Testing Calibrated General Equilibrium Models

Fabio Canova*
Department of Economics, Universitat Pompeu Fabra, Balmes 132,
E-08008 Barcelona, Spain;
Department of Economics, University of Modena, I-41100 Modena, Italy;
and CEPR

and

Eva Ortega
Department of Economics, European University Institute, Via dei Roccettini, 9
I-50016 San Domenico di Fiesole (FI), Italy

First draft: October 1995
Preliminary and Incomplete

Key words: Calibration, Simulation, Evaluation, Saving and Investment Correlations.

JEL Classification nos: C15, C52, D58.

1 Introduction

Simulation techniques are now used in many fields of applied research. As shown elsewhere in this book, they have been employed to compute estimators in situations where standard methods are impractical or fail, to evaluate the properties of parametric and nonparametric econometric estimators, to provide cheap estimates of posterior integral in Bayesian analysis and computationally simple ways to undertake linear and nonlinear filtering.

The task of this chapter is to describe how simulation based methods can be used to evaluate the fit of dynamic general equilibrium models whose parameters have been calibrated to the data, to compare and contrast their usefulness relative to more standard econometric approaches and to provide an explicit example where the various features of the approach can be highlighted and discussed.

The structure of this chapter is as follows. First, we provide a definition of what we mean by calibrating a model and discuss the philosophy underlying the approach and how it differs from standard dynamic time series modelling. Second, we discuss various approaches to the selection of model parameters, how to choose the vector of statistics used to compare actual with simulated data and how simulations are performed. Third, we describe how to evaluate the quality of the model approximation to the data and discuss alternative approaches to account for the uncertainty faced by a simulator in generating time paths for the relevant variables. Although we present a general overview of alternative evaluation techniques, the focus is on simulation based approaches. Finally, we present an example, borrowed from Baxter and Crucini (1993), where the features of the various approaches to evaluation can be stressed. We conclude by indicating how simulation techniques can be used in constructing additional tests measuring the fit of the model to the data.

2 What is Calibration?

2.1 A Definition

Although it is more than a decade since calibration techniques emerged in the main stream of dynamic macroeconomics (see Kydland and Prescott (1982)), a precise statement of what it means to calibrate a model has yet to appear in the literature. In general, it is common to think of calibration as an (unorthodox) procedure to select the parameters of a model. One may want to calibrate a model because there is no available data to estimate its parameters, for example when we want to study the effect of certain taxes in a newly born country. Some, on the other hand, interpret calibration as a particular econometric technique where the parameters of the model are estimated using an "economic" instead of a "statistical" criteria (see e.g. Canova (1994)). Alternatively, it is possible to view calibration as a cheap way to evaluate models. For example, calibration is considered by some a more formal version of the standard back-of-the-envelope calculations that theorists perform to judge the validity of their models (see e.g. Pesaran and Smith (1992)). According to others, calibration is a way to conduct quantitative experiments using models which are known to be "false", i.e. improper or simplified approximation of the true data generating process of the actual data (see e.g. Kydland and Prescott (1991)).

Pagan (1994) stressed that the unique feature of calibration does not lie so much in the way parameters are estimated, as the literature has provided alternative ways of doing so, but in the particular collection of procedures used to test tightly specified (and false) theoretical models against particular empirical facts. Here we take a more general point of view and identify 6 steps which, we believe, capture the essence of the methodology. We call calibration a procedure which involves:

(i) Formulating an economic question to be addressed.

(ii) Selecting a model design which bears some relevance to the question asked.

(iii) Choosing functional forms for the primitives of the model and finding a solution for the endogenous variables in terms of the exogenous ones and the parameters.

(iv) Choosing parameters and stochastic processes for the exogenous variables and simulating paths for the endogenous variables of the model.

(v) Selecting a metric and comparing the outcomes of the model relative to a set of "stylized facts".

(vi) Doing policy analyses if required.

By "stylized facts" the literature typically means a collection of sample statistics of the actual data which (a) do not involve estimation of parameters and (b) are self-evident (such as means, variances, correlations, etc.). More recently, however, the first requirement has been waived and the parameters of a VAR (or the impulse responses) have also been taken as the relevant stylized facts to be matched by the model (see e.g. Smith (1993), Cogley and Nason (1994)).

The next two subsections describe in details both the philosophy behind the first four steps and the practicalities connected with their implementation.

2.2 Formulating a question and choosing a model

The first two steps of a calibration procedure, to formulate a question of interest and a model which bears relevance to the question, are self evident and require very little discussion. In general, the questions posed display four types of structures (see e.g. Kollintzas (1992) and Kydland (1992)):

- How much of the fact X can be explained with impulses of type Y?
- Is it possible to generate Z using theory W?
- Is it possible to reduce the discrepancy D of the theory from the data by introducing feature F in the model?
- What happens to the endogenous variables of the model if the stochastic process for the control variable V is modified ?

Two economic questions which have received considerable attention in the literature in the last 10 years are the so-called equity premium puzzle, i.e. the inability of a general equilibrium model with complete financial markets to quantitatively replicate the excess returns of equities over bonds over the last hundred years (see e.g. Mehra and Prescott (1985)) and how much of the variability of GNP can be explained by a model whose only source of dynamics are technology disturbances (see e.g. Kydland and Prescott (1982)). As is clear from these two examples, the type of questions posed are very specific and the emphasis is on the numerical implications of the exercise. Generic questions with no numerical quantification are not usually studied in this literature.

For the second step, the choice of an economic model, there are essentially no rules except that it has to have some relevance with the question asked. For example, if one is interested in the equity premium puzzle, one can choose a model which is very simply specified on the international and the government side, but very well specified on the financial side so that it is possible to calculate the returns on various assets. Typically, one chooses dynamic general equilibrium models. However, several authors have used model designs coming from different paradigms (see e.g. the neo-keynesian model of Gali (1992), the non-walrasian model of Danthine and Donaldson (1992) and the model with union bargaining of Eberwin and Kollintzas (1995)). There is nothing in the procedure that restricts the class of model design to be used. The only requirement is that the question that the researcher formulates is quantifiable within the context of the model and that the theory, in the form of a model design, is fully specified.

It is important to stress that a model is chosen on the basis of the question asked and not on its being realistic or being able to best replicate the data (see Kydland and Prescott (1991) or Kydland (1992)). In other words, how well it captures reality is not a criteria to select models. What is important is not whether a model is realistic or not but whether it is able to provide a quantitative answer to the specific question the researcher poses.

This brings us to discuss an important philosophical aspect of the methodology. From the point of view of a calibrator all models are approximations to the DGP of the data and, as such, false. This aspect of the problem has been appreciated by several authors even before the appearance of the seminal article of Kydland and Prescott. For example, Hansen and Sargent (1979) also concede that an economic model is a false DGP for the data. Because of this and in order to test the validity of the model using standard statistical tools, they complete the probabilistic structure of the model by adding additional sources of variability, in the form of measurement errors or unobservable variables, to the fundamental forces of the economy.

For calibrators, the model is not a null hypothesis to be tested but an approximation of a few dimensions of the data. A calibrator is not interested in veryfying whether the model is true (the answer is already known from the outstart), but in identifying which aspects of the data a false model can replicate and whether different models give different answers because they are false in different dimensions. A calibrator is satisfied with his effort if, through a process of theoretical respecification, a simple and highly stylized model captures an increasing number of features of the data (confront this activity with the so-called normal science of Kuhn (1970)).

Being more explicit, consider the realization of a vector of stochastic processes y_t (our data) and some well specified theoretical model $x_t = f(z_t, \gamma)$ which has something to say about y_t, where z_t are exogenous and predetermined variables and γ parameters. Because the model

does not provide a complete description of the phenomenon under investigation we write

$$y_t = x_t + u_t \tag{1}$$

where u_t is an error representing what is missing from $f(z_t,\gamma)$ to reproduce the stochastic process generating y_t and whose properties are, in general, unknown (it need not be necessarily mean zero, serially uncorrelated, uncorrelated with the x's and so on). Let B_y and B_x be continuous and differentiable functions of actual and simulated data, respectively. Then, standard econometric procedures judge the coherence of the model to the data by testing whether or not $B_x = B_y$, given that the difference between B_x and B_y and their estimated counterpart $\widehat{B}_x$ and $\widehat{B}_y$ arises entirely from sampling error. While this is a sensible procedure when the null hypothesis is expected to represent the data, it is less sensible when it is known that the model does not completely capture all aspects of the data.

The third step of a calibration exercise concerns the solution of the model. To be able to obtain quantitative answers from a model it is necessary to find an explicit solution for the endogenous variables of the model in terms of the exogenous and predetermined variables and the parameters. For this reason it is typical to parameterize the objective function of the agents so that manipulation of the first order conditions is analytically tractable. For example, in general equilibrium models, it is typical to choose Cobb-Douglas production functions and constant relative risk aversion utility functions. However, although the main objective is to select simple enough functional forms, it is well known that almost all general equilibrium models and many partial equilibrium models have exact analytical solutions only in very special situations.

For general equilibrium models, a solution exists if the objective function is quadratic and the constraints linear (see e.g. Hansen and Sargent (1979)) or when the objective function is log-linear and the constraints linear (see e.g. Sargent (1987, ch.2)). In the other cases, analytical expressions relating the endogenous variables of the model to the "states" of the problem does not exist and it is necessary to resort to numerical techniques to find solutions which approximate equilibrium functionals either locally or globally. There has been substantial theoretical development in this area in the last few years and several solution algorithms have appeared in the literature (see e.g. the special January 1990 issue of the JBES or Marcet (1995)).

The essence of the approximation process is very simple. The exact solution of a model is a relationship between the endogenous variables x_t, the exogenous and predetermined variables z_t and a set of "deep" parameters γ of the type $x_t = f(z_t,\gamma)$ where f is generally unknown. The approximation procedures generate a relationship of the type $x_t^* = g(z_t,\theta)$ where $\theta = h(\gamma)$ and where $||f - g|| < \epsilon$ is minimal for some local or global metric. Examples of these types of procedures appear in Kydland and Prescott (1982), Coleman (1989), Tauchen and Hussey (1991), Novales (1990), Baxter (1992) and Marcet (1992), among others. The choice of a particular approximation procedure depends on the question asked. If one is concerned with the dynamics of the model around the steady state, local approximations are sufficient. On the other hand, if one is interested in comparing economic policies requiring drastic changes in the parameters of the control variables, global approximation methods must be used.

2.3 Selecting Parameters and Exogenous Processes

Once an approximate solution has been obtained, a calibrator needs to select the parameters γ and the exogenous stochastic process z_t to be fed into the model in order to generate time series for x_t^*. There are several approaches to the choice of these two features of the model. Consider first the question of selecting z_t. This choice is relatively uncontroversial. One either chooses it on the basis of tractability or to give the model some realistic connotation. For example, one can assume that z_t is an AR process with innovations which are transformations of a $N(0,1)$ process and draw one or more realizations for z_t using standard random number generators, or select the Solow residuals of the actual economy, or the actual path of government expenditure or of the money supply. Obviously, the second alternative is typically preferred if policy analyses are undertaken. Note that while in both cases z_t is the realization of a stochastic process, in the first case the DGP is known while in the second it is not and this has implications for the way one measures the uncertainty in the outcomes of the model.

Next, consider the selection of the vector of parameters γ. Typically, they are chosen so that the model reproduces certain observations. Taking an example from physics, if one is interested in measuring water temperature in various situations it will be necessary to calibrate a thermometer for the experiments. For this purpose a researcher arbitrarily assigns the value 0 C to freezing water and the value 100 C to boiling water and interpolates values in the middle with, say, a linear scale. Given this calibration of the thermometer, one can then proceed to measure the results of the experiments: a value close to 100 C indicates "hot" water, a value close to 30 C indicates "tepid" water, and so on. To try to give answers to the economic question he poses, a calibrator must similarly select observations to be used to calibrate the model-thermometer. There are at least three approaches in the literature. One can follow the deterministic computable general equilibrium (CGE) tradition, summarized, e.g. in Showen and Walley (1984), the dynamic general equilibrium tradition pioneered by Kydland and Prescott (1982) or employ more standard econometric techniques. There are differences between the first two approaches. The first one was developed for deterministic models which do not necessarily possess a steady state while the second one has been applied to dynamic stochastic models whose steady state is unique. Kim and Pagan (1994) provide a detailed analysis of the differences between these two approaches. Gregory and Smith (1993) supplement the discussion by adding interesting insights in the comparison of the first two approaches with the third.

In CGE models, a researcher solves the model linearizing the system of equations by determining the endogenous variables around a hypothetical equilibrium where prices and quantities are such that there is no excess demand or excess supply. It is not necessary that this equilibrium exists. However, because the coefficients of the linear equations determining endogenous variables are functions of these equilibrium values, it is necessary to measure this hypothetical equilibrium. The main problem for this literature is therefore to find a set of "benchmark data" and to calibrate the model so that it can reproduce this data. Finding this data set is the most complicated part of the approach since it requires a lot of judgement and ingenuity. The process of specification of this data set leaves some of the parameters of the model typically undetermined, for example, those that describe the utility function of agents. In this situation a researcher either assigns arbitrary values or fixes them to values estimated in

other studies in the literature. Although these choices are arbitrary, the procedure is coherent with the philosophy of the models: a researcher is interested in examining deviations of the model from a hypothetical equilibrium, not from an economy in real time.

In stochastic general equilibrium models, the equilibrium used to calibrate the model is, typically, the steady state: parameters are chosen so that the model, in the steady state, produces values for the endogenous variables which exactly match corresponding averages of the actual data. In both this approach and the CGE approach, point estimates of the parameters used to calibrate the model to the equilibrium are taken to be exact (no standard deviations are typically attached to these estimates). As in the previous setup, the steady state does not necessarily pin down all the parameters of the model. Canova (1994) and Gregory and Smith (1993) discuss various methods to select the remaining parameters. Briefly, a researcher can choose parameters a-priori, pin them down using values previously estimated in the literature, can informally estimate them using simple method of moment conditions where such a parameter may appear or formally estimate them using procedures like GMM (see e.g. Christiano and Eichenbaum (1992)), SMM (see e.g. Duffie and Singleton (1993)) or maximum likelihood (see e.g. McGratten, Rogerson and Wright (1993)). As pointed out by Kydland and Prescott (1991), choosing parameters using the information contained in other studies imposes a coherence criteria among various branches of the profession. For example, in the business cycle literature one uses growth models to examine business cycle fluctuations and checks the implications of the model using parameters obtained typically, in micro studies, which do not employ data having to do with aggregate business cycle fluctuations (e.g. micro studies of labor markets).

If one follows a standard econometric approach, all the parameters are chosen by minimizing the MSE of the error u_t in (1), arbitrarily assuming that the error and the model designs are orthogonal, or by minimizing the distance between moments of the actual data and the model or maximizing the likelihood function of the data given the model design. As we already pointed out, this last approach is the least appealing one from the point of view of a calibrator since it makes assumptions on the time series properties of u_t which are hard to justify from an economic theory point of view.

To clearly understand the merits of each of these procedures it is useful to discuss their advantages and their disadvantages. Both the CGE and the Kydland and Prescott approach where some of the parameters are chosen a-priori or selected using values obtained from a very selected group of studies are problematic in several respects. First, there is a selectivity bias problem (see Canova (1995)). There exists a great variety of estimates of the parameters in the literature and different researchers may refer to different studies even when they are examining the same problem. Second, there is a statistical inconsistency problem which may generate very spurious and distorted inference. As Gregory and Smith (1989) have shown, if some parameters are set a-priori and others estimated by simulation, estimates of the latter may be biased and inconsistent unless the parameters of the former group are the true parameters of the DGP of the data or consistent estimates of them. Third, since any particular choice is arbitrary, extensive sensitivity analysis is necessary to evaluate the quality of the results. To solve these problems Canova (1994)-(1995) suggests an approach for choosing parameters which allows, at a second stage, to draw inferences about the quality of the approximation of the model to the data. The idea is very simple. Instead of choosing one set of parameters over another he suggests to

calibrate each parameter of the model to an interval, use the empirical information to construct a distribution over this interval (the likelihood of a parameter given existing estimates) and conduct simulation by drawing parameter vectors from the corresponding joint "empirical" distribution. An example may clarify the approach. If one of the parameters of interest is the coefficient of constant relative risk aversion of the representative agent, one typically chooses a value of 2 and tries a few values above and below this one to see if results change. Canova suggests to take a range of possible values, possibly dictated by economic theory, say [0,20], and then over this range construct an histogram using existing estimates of this parameter. Most of the estimates are in the range [1,2] and in some asset pricing models researchers have tried values up to 10. Given this information, the resulting empirical distribution for this parameter can be very closely approximated by a $\chi^2(4)$, which has the mode at 2 and about 5% probability in the region above 6.

The selection of the parameters of theoretical models through statistical estimation has advantages and disadvantages. The main advantage is that these procedures avoid arbitrary choices and explicitly provide a measure of dispersion for the estimates which can be used at a second stage to evaluate the quality of the approximation of the model to the data. The disadvantages are of various kinds. First of all, to undertake a formal or informal estimation it is typically necessary to select the moments one wants to fit, and this choice is arbitrary. The standard approach suggested by Kydland and Prescott can indeed be thought as a method of moment estimation where one chooses parameters so as to set only the discrepancy between the first moment of the model and the data (i.e. the long run averages) to zero. The formal approach suggested by Christiano and Eichenbaum (1992) or Langot and Fève (1994), on the other hand, can be thought of as a method of moment estimation where a researcher fits to zero the discrepancies between model and data first and second moments. The approach of choosing parameters setting to zero certain moments has the disadvantage of reducing the number of moments over which it will be possible to evaluate the quality of the model. Moreover, it is known that estimates obtained with the method of moments or GMM may be biased. Therefore, simulations and inference conducted with these estimates may lead to spurious inference (see e.g. Canova, Finn and Pagan (1994)). In addition, informal SMM may lead one to select parameters even though they are not identifiable (see Gregory and Smith (1989)). Finally, one should note that the type of uncertainty which is imposed on the model via an estimation process does not necessarily reflect the uncertainty a calibrator faces when choosing the parameter vector. As is clear from a decade of GMM estimation, once the moments are selected and the data given, sample uncertainty is pretty small. The true uncertainty is in the choice of moments and in the data set to be used to select parameters. This uncertainty is disregarded when parameters are chosen using extremum estimators like GMM.

Finally, it is useful to compare the parameter selection process used by a calibrator à-la Kydland and Prescott and the one used by a traditional econometric approach. As we have mentioned, in a traditional econometric approach parameters are chosen so as to minimize some **statistical** criteria, for example, the MSE. Such criteria do not have any economic content, impose stringent requirements on the structure of u_t and are used, primarily, because there exists a well established statistical and mathematical literature on the subject. In other words, the parameter selection criteria used by traditional econometricians does not have economic justification. On the other hand, the parameter selection criteria used by followers of the

Kydland and Prescott's methodology can be thought of as being based on **economic** criteria. For example, if the model is calibrated so that, in the steady state, it matches the long run features of the actual economy, parameters are implicitly selected using the condition that the sum (over time) of the discrepancies between the model and the data is zero. In this sense there is an important difference between the two approaches which has to do with the assumptions that one is willing to make on the errors u_t. By calibrating the model to long run observations a researcher selects parameters assuming $E(u) = 0$, i.e. using a restriction which is identical to the one imposed by a GMM econometrician who chooses parameters using only first moment conditions. On the other hand, to conduct classical inference a researcher imposes restrictions on the first and second moments of u_t.

The comparison we have made so far concerns, obviously, only those parameters which enter the steady state conditions of the model. For the other parameters a direct comparison with standard econometric practice is not possible. However, if all parameters are calibrated to intervals with distributions which are empirically determined, the calibration procedure we have described shares a tight connection with Bayesian Inferential methods such as Consensus Analysis or Meta-Analysis (see e.g. Genest and Zidak (1986) or Wolf (1986)).

Once the parameters and the stochastic processes for the exogenous variables are selected and an (approximate) solution to the model has been found, simulated paths for x_t^* can be generated using standard Monte Carlo simulation techniques.

3 Evaluating Calibrated Models

The questions of how well the model matches the data and how much confidence a researcher should give to the results constitute the most crucial step in the calibration procedure. In fact, the most active methodological branch of this literature concerns methods to evaluate the fit of a model to the data. The evaluation of a model requires three steps: first, the selection of stylized facts; second, the choice of a metric to compare functions of actual and simulated data and third, a (statistical) evaluation of the magnitude of the distance. Formally, let S_y be a set of statistics (stylized facts) of the actual data and let $S_{x^*}(z_t, \gamma)$ be a set of statistics of simulated data, given a vector of parameters γ and a vector of stochastic processes z_t. Then model evaluation consists in selecting a function $\psi(S_y, S_{x^*}(z_t, \gamma))$ that measures the distance between S_y and S_{x^*} and in assessing its magnitude.

The choice of which stylized facts one wants to match obviously depends on the question asked and on the type of models used. For example, if the question is what is the proportion of actual cyclical fluctuations in GNP and consumption explained by the model, one would choose stylized facts based on variances and covariances of the data. As an alternative to the examination of second moments of the data, one could summarize the properties of actual data via a VAR and study the properties of simulated data, for example, comparing the number of unit roots in the two sets of data (as in Canova, Finn and Pagan (1994)), the size of VAR coefficients (as in Smith (1993) or Ingram, DeJong and Whiteman (1995)) or the magnitude of certain impulse responses (as in Cogley and Nason (1994)). Also, it is possible to evaluate the discrepancy of a model to the data by choosing specific events that one wants the model to replicate e.g. business cycle turning points, (as in King and Plosser (1994) or Simkins (1994)) or variance bounds (as in Hansen and Jagannathan (1991)).

Classical pieces in the calibration literature (see e.g. Kydland and Prescott (1982) or (1991)) are typically silent on the metric one should use to evaluate the quality of the approximation of the model to the data. The approach favored by most calibrators is to glare over the exact definition of the metric used and informally assess the properties of simulated data by comparing them to the set of stylized facts. In this way a researcher treats the experiment conducted as a measurement exercise where the task is to gauge the proportion of some observed statistics reproduced by the theoretical model. This informal approach is also shared by cliometricians (see e.g. Summers (1991)) who believe that rough reproduction of simple sample statistics are everything that is needed to evaluate the implications of the model ("either you see it with naked eyes or no fancy econometric procedure will find it").

There are, however, alternatives to this informal approach. To gain some understanding of the differences among approaches, but at the cost of oversimplifying the matter, we divide evaluation approaches into five classes:

- Informal approaches.
- Approaches which do not consider sampling variability of the actual or simulated data, but instead use the statistical properties of u_t in (1) to impose restrictions on the time series properties of ψ. This allows them to provide an R^2 type measure of fit between the model and the data (see Watson (1993)).
- Approaches which use the sampling variability of the **actual** data (affecting S_y and, in some cases, estimated γ) to provide a measure of the distance between the model and the data. Among these we list the GMM based approach of Christiano and Eichenbaum (1992), Cecchetti, Lam and Mark (1994) or Fève and Langot (1994), and the frequency domain approaches of Diebold, Ohanian and Berkowitz (1994) and Ortega (1995).
- Approaches which use sampling variability of the **simulated** data to provide a measure of distance between the model and the data. Among these procedures we can distinguish those who take z_t as stochastic and γ as given, such as Gregory and Smith (1991), Söderlind (1994) or Cogley and Nason (1994) and those who take both z_t and γ as stochastic, such as Canova (1994) and (1995).
- Finally, approaches which consider the sampling variability of both the actual and simulated data to evaluate the fit of the model. Once again we can distinguish approaches which, in addition to taking S_y as random, allow for variability in the parameters of the model (keeping z_t fixed) such as DeJong, Ingram and Whiteman (1995) from those which allow for both z_t and γ to vary such as Canova and De Nicoló (1995).

Because the emphasis of this book is on simulation techniques, we will briefly examine the first three approaches and discuss more in detail the last two, which make extensive use of simulation techniques to conduct inference. Kim and Pagan (1994) provide a thorough critical review of several of these evaluation techniques and additional insights on the relationship among them.

The evaluation criteria that each of these approaches proposes is tightly linked to the parameter selection procedure we discussed in the previous section.

A calibrator typically chooses parameters using steady state conditions, and those which do not appear in the steady state equations are chosen a-priori or referring to a particular study. In this case, because S_y is a vector of numbers, there are no free parameters. No uncertainty is allowed in the selected parameter vector; one is forced to use an informal metric to compare the model to the data. This is because, apart from the uncertainty present in the exogenous variables, the model links the endogenous variables to the parameters in a deterministic fashion. Therefore, once we have selected the parameters and we have a realization of S_y it is not possible to measure the dispersion of the distance $\psi(S_y, S_{x^*}(z_t, \gamma))$. From the point of view of the majority of calibrators this is not a problem. As emphasized by Kydland and Prescott (1991) or Kydland (1992), the trust a researcher has in an answer given by the model does not depend on a statistical measure of discrepancy, but on how much he believes in the economic theory used and in the measurement undertaken.

Taking this as the starting point of the analysis Watson (1993) suggests an ingenious way to evaluate models which are known to be an incorrect DGP for the actual data. Watson asks how much error should be added to x_t^* so that its autocovariance function equals the autocovariance function of y_t. Writing $y_t = x_t^* + u_t^*$ where u_t^* includes the approximation error due to the use x_t^* in place of x_t, the autocovariance function of this error is given by

$$A_{u^*}(z) = A_y(z) + A_{x^*}(z) - A_{x^*y}(z) - A_{yx^*}(z) \tag{2}$$

To evaluate the last two terms in (10) we need a sample from the joint distribution of (x_t^*, y_t) which is not available. In this circumstances it is typical to assume that either u_t^* is a measurement error or a signal extraction noise (see e.g. Sargent (1989)), but in the present context none of the two assumptions is very appealing. Watson suggests to choose $A_{x^*y}(z)$ so as to minimize the variance of u_t^* subject to the constraint that $A_{x^*}(z)$ and $A_y(z)$ are positive semidefinite. Intuitively, the idea is to select $A_{x^*y}(z)$ to give the best possible fit between the model and the data (i.e. the smallest possible variance of u_t^*). The exact choice of $A_{x^*y}(z)$ depends on the properties of x_t^* and y_t, i.e. whether they are serially correlated or not, scalar or vectors, full rank processes or not. In all cases, the selection criteria chosen imply that x_t^* and y_t are perfectly linearly correlated where the matrix linking the two vectors depends on their time series properties and on the number of shocks buffeting the model. Given this framework of analysis, Watson suggests measures of fit statistics, similar to a $1 - R^2$ from a regression, of the form

$$r_j(\omega) = \frac{A_{u^*}(\omega)_{jj}}{A_y(\omega)_{jj}} \tag{3}$$

$$R_j(\omega) = \frac{\int_{\omega \in Z} A_{u^*}(\omega)_{jj} d\omega}{\int_{\omega \in Z} A_y(\omega)_{jj} d\omega} \tag{4}$$

where the first statistic measures the variance of the j-th component of the error relative to the variance of the $j - th$ component of the data for each frequency and the second statistic is the sum of the first over a set of frequencies and may be useful to evaluate the model, say, at business cycle frequencies. It should be stressed that (11) and (12) are lower bounds. That is, when $r_j(\omega)$ or $R_j(\omega)$ are large, the model poorly fits the data. However, when they are small they do not necessarily indicate good fit since the model may still fit poorly the data, if we change the assumptions about $A_{x^*y}(z)$.

To summarize, Watson chooses the autocovariance function of y as the set of stylized facts of the data to be matched by the model, the ψ function as the ratio of A_{u^*} to A_y and evaluates the size of ψ informally. Note that in this approach, γ and z_t are fixed, and A_{x^*} and A_y are assumed to be measured without error.

When a calibrator is willing to assume that parameters are measured with error because, given an econometric technique and a sample, parameters are imprecisely estimated, then model evaluation can be conducted using measures of dispersion for simulated statistics which reflect parameter uncertainty. There are various versions of this approach. One is the criteria of Christiano and Eichenbaum (1992), Cecchetti, Lam and Mark (1994) and Fève and Langot (1994) which use a version of a J-test to evaluate the fit of a model. In this case S_y are moments of the data while ψ is a quadratic function of the type

$$\psi(S_y, S_{x^*}(z_t, \gamma)) = [S_y - S_{x^*}(\gamma)]V^{-1}[S_y - S_{x^*}(\gamma)]' \tag{5}$$

where V is a matrix which linearly weights the covariance matrix of S_{x^*} and S_y, and S_{x^*} is random because γ is random. Formal evaluation of this distance can be undertaken following Hansen (1982): under the null that $S_y = S_{x^*}(z_t, \gamma)$ the statistic defined in (13) is asymptotically distributed as a χ^2 with the number of degrees of freedom equal to the number of overidentifying restrictions, i.e. the dimension of S_y minus the dimension of the vector γ. Note that this procedure is correct asymptotically and that it implicitly assumes that $x_t = f(z_t, \gamma)$ (or its approximation) is the correct DGP for the data and that the relevant loss function measuring the distance between actual and simulated data is quadratic.

Although the setup is slightly different, also the methods proposed by Diebold, Ohanian and Berkowitz (DOB) (1994) and Ortega (1995) can be broadly included into this class of approaches.

For DOB the statistic of interest is the spectral density matrix of y_t and, given a sample, this is assumed to be measured with error. They construct the uncertainty around point estimates of the spectral density matrix by using parametric and nonparametric bootstrap approaches and Bonferroni tunnels to obtain (small sample) 90% confidence bands around these point estimates. On the other hand, they take calibrated parameters and the realization of z_t as given so that the spectral density matrix of simulated data can be estimated without error simply by simulating very long time series for x_t^*. Ortega (1995) also takes the spectral density matrix as the set of stylized facts of the data to be matched by the model. Differently to DOB, she estimates jointly the spectral density matrix of actual and simulated data and constructs the uncertainty around point estimates using asymptotic distribution theory.

In both cases, the measure of fit of the model to the data is generically given by:

$$C(\gamma, z_t) = \int_0^{\pi} \psi(F_y(\omega), F_{x^*}(\omega, \gamma, z_t))W(\omega)d\omega \tag{6}$$

where $W(\omega)$ is a set of weights applied to different frequencies and F are the spectral density matrices of actual and simulated data.

DOB suggest various options for ψ (quadratic, ratio, likelihood type) but do not construct a test statistic to examine the magnitude of ψ. Instead, they compute a small sample distribution of the event that $C(\gamma, z_t)$ is close to zero. Ortega, on the other hand, explicitly uses

a quadratic distance function and constructs an asymptotic χ^2 test to measure the magnitude of the discrepancy between the model and the data. The set of tools she develops can also be used to compare the fit of two alternative models to the data and decide which one is the best using asymptotic criteria.

If a calibrator is willing to accept the idea that the stochastic process for the exogenous variables is not fixed, he can then compute measures of dispersion for simulated statistics by changing the realization of z_t while maintaining the parameters fixed. Such a methodology has its cornerstone in the fact that it is the uncertainty in the realization of the exogenous stochastic process (e.g. the technology shock), an uncertainty which one can call extrinsic, and not the uncertainty in the parameters, which one can call intrinsic, which determines possible variations in the sample statistics of simulated data. Once the dispersion of simulated statistics is obtained, the sampling variability of simulated data can be used to evaluate the distance between moments of simulated and of actual data (as e.g. Gregory and Smith (1991) and (1993)).

If one uses such an approach, the evaluation of a model can be conducted with a probabilistic metric using well known techniques in the Monte Carlo literature. For example, one may be interested in finding out in what decile of the simulated distribution the actual value of a particular statistics lies calculating, in practice, the "size" of calibration tests. This approach requires that the evaluator takes the model economy as the true DGP for the data and that differences between S_y and S_{x^*} may occur only because of sampling variability. To be specific, Gregory and Smith take S_y be a set of moments of the data and they assume that they can be measured without error. Then they construct a distribution of $S_{x^*}(z_t, \gamma)$ by drawing realizations of the z_t process from a given distribution for fixed γ. The metric ψ used is probabilistic, i.e. they calculate the probability $Q = P(S_{x^*} \leq S_y)$, and the judgement of the fit of the model is informal, e.g. measuring how close Q is to 0.5.

An interesting variation on this setup is provided by the work of Söderlind (1994) and Cogley and Nason (1994). Söderlind employs the spectral density matrix of the actual data as the relevant set of statistics to be matched while Cogley and Nason look at a "structural" impulse response function. Söderlind maintains a probabilistic metric and constructs empirical rejection rates for an event (that the actual spectral density matrix of y_t lies inside the asymptotic 90% confidence band for the spectral density matrix of the simulated data), and the event is replicated by drawing vectors z_t for a given distribution. Cogley and Nason choose a quadratic measure of distance which has an asymptotic χ^2 distribution under the null that the DGP for the data is the model. Then they tabulate the empirical rejection rates of the test, by repeatedly constructing the statistic drawing realizations of the z_t vector. To be specific the ψ function is given by

$$\psi_{k,j}(\gamma) = [IRF_{x^*}(z_t^j, \gamma) - IRF_y]V^{-1}[IRF_{x^*}(z_t^j, \gamma) - IRF_y]' \tag{7}$$

where j refers to replication and k to the $k-th$ step, IRF is the impulse response function and V is the asymptotic covariance matrix of the impulse response function of x^*. Because for every k and for fixed j $\psi_{k,j}(\gamma)$ is asymptotically χ^2, they can construct both a simulated distribution for $\psi_{k,j}$ and compare it with a χ^2 and the rejection frequency for each model specification they examine.

In practice, all three approaches are computer intensive and rely on Monte Carlo methods to conduct inference. Also, all three methods verify the validity of the model by assuming that the model is the correct DGP for y_t and computing the "size" of the calibration tests.

The approach of Canova (1994)-(1995) also belongs to this category of methods, but in addition to allowing the realization of the stochastic process for the exogenous variables to vary he also allows parameter variability in measuring the dispersion of simulated statistic. The starting point, as discussed earlier, is that parameters are uncertain not so much because of sample variability, but because there are many estimates of the same parameter obtained in the literature, since estimation techniques, samples and frequency of the data tend to differ. If one calibrates the parameter vector to an interval, instead of to a particular value, and draws values for the parameters from the empirical distributions of their estimates, it is then possible to use intrinsic uncertainty in addition or in alternative to the extrinsic one, to evaluate the fit of the model. The evaluation approach is then very similar to the one of Gregory and Smith: one simulates the model by drawing parameter vectors from the empirical "prior" distribution and realizations of the exogenous stochastic process z_t from some given distribution. Once the empirical distribution of the statistics of the model is constructed is then possible to compute either the size of calibration tests or the percentiles where the statistics found in the actual data lie.

The last set of approaches recently developed in the literature considers the uncertainty present in the statistics of both actual and simulated data to measure the fit of the model to the data. In essence what these approaches attempt is to formally measure the degree of overlap between the (possibly) multivariate distributions of S_y and S_x using Monte Carlo techniques. There are differences on the way these distributions are constructed. Canova and De Nicoló (1995) use a parametric bootstrap algorithm to construct distributions for the multivariate statistic of the actual data they consider (first and second moments of the equity premium and of the risk free rate). DeJong, Ingram and Whiteman (DIW) (1995), on the other hand, suggest to represent the actual data with a VAR. Then they compute posterior distribution estimates for the moments of interest by drawing VAR parameters from their posterior distribution and constructing, at each replication, the moments of interest from the AR(1) companion matrix representation of the VAR. In constructing distributions of simulated statistics, Canova and De Nicoló take into account both the uncertainty in exogenous processes and parameters while DIW only consider parameter uncertainty. In addition they differ in the way the "prior" uncertainty in the parameters is introduced in the model. The former paper follows Canova (1995) and chooses empirical based distributions for the parameter vector. DIW use subjectively specified prior distributions (typically normal) whose location parameter is set at the value typically calibrated in the literature while the dispersion parameter is free. The authors use this parameter in order to (informally) minimize the distance between the actual and the simulated distributions of the statistics of interest. By enabling the specification of a sequence of increasingly diffuse priors over the parameter vector, the procedure can therefore illustrate whether uncertainty in the choice of the model's parameters can mitigate differences between the model and the data.

There are few differences in the two approaches also in deciding the degree of overlap of

the two distributions. Canova and De Nicoló choose a particular contour probability for one of the two distributions and ask how much of the other distribution is inside this contour. In other words, the fit of the model is examined very much in the style of the Monte Carlo literature, where a good fit is indicated by a high probability covering of the two regions. In addition, to study the features of the two distributions, they repeat the exercise varying the chosen contour probability, say, from 50% to 75%, 90%, 95% and 99%. In this way it is possible to detect anomalies in the shape of the two distributions due to clustering of observations in one area, skewness or leptokurtic behavior. In this approach actual data and simulated data are used symmetrically in the sense that one can either ask whether the actual data could be generated by the model, or, viceversa, whether simulated data are consistent with the distribution of the empirical sample. This symmetry allows to a much better understanding of the properties of error u_t in (1) and resembles very much to switching the null and the alternative in standard classical hypothesis testing.

DeJong, Ingram and Whiteman take the point of view that there are no well established criteria to judge the adequacy of a model's "approximation" to reality. For this reason they present two statistic aimed at synthetically measuring the degree of overlap among distributions. One they call Confidence Interval Criterion (CIC) is the univariate version of the contour probability criteria used by Canova and De Nicoló and is defined as

$$CIC_{ij} = \frac{1}{1-\alpha}\int_a^b P_j(s_i)ds_i \tag{8}$$

where s_i, $i = 1, \ldots, n$ is a set of functions of interests, $a = \frac{\alpha}{2}$ and $b = 1 - a$ are the quantiles of $D(s_i)$, the distribution of the statistic in the actual data, $P_j(s_i)$ is the distribution of simulated statistic where j is the diffusion index of the prior on the parameter vector and $1 - \alpha = \int_a^b D(s_i)ds_i$. Note that with this definition, CIC_{ij} ranges between 0 and $\frac{1}{1-\alpha}$. For CIC close to zero, the fit of the model is poor, i.e. either the overlap is small or P_j is very diffuse. For CIC close to $\frac{1}{1-\alpha}$ the two distributions overlap substantially. Note also that if $CIC > 1$, $D(s_i)$ is diffuse relative to $P_j(s_i)$, i.e. the data is found relatively uninformative regarding s_i.

To distinguish among the two possible interpretations when CIC is close to zero, DeJong, Ingram and Whiteman suggest a second summary measure analogous to a t-statistic for the mean of $P_j(s_i)$ in the $D(s_i)$ distribution, i.e.,

$$d_{ji} = \frac{EP_j(s_i) - ED(s_i)}{\sqrt{var\,D(s_i)}} \tag{9}$$

Large values of (17) would indicate that the location of $P_j(s_i)$ is quite different from the location of $D(s_i)$.

The final problem is to choose α. DeJong, Ingram and Whiteman fix a particular value ($\alpha = 0.01$) but, as in Canova and De Nicoló, varying α for a given j is probably a good thing to do in order to study the shape of the distributions. This is particularly useful when we are interested in partitions of the joint distributions of s_i and/or graphical methods are not particularly informative about distributions in high dimensional spaces.

4 Evaluating Calibrated Models

The questions of how well the model matches the data and how much confidence a researcher should give to the results constitute the most crucial step in the calibration procedure. In fact, the most active methodological branch of this literature concerns methods to evaluate the fit of a model to the data. The evaluation of a model requires three steps: first, the selection of stylized facts; second, the choice of a metric to compare functions of actual and simulated data and third, a (statistical) evaluation of the magnitude of the distance. Formally, let S_y be a set of statistics (stylized facts) of the actual data and let $S_{x^*}(z_t, \gamma)$ be a set of statistics of simulated data, given a vector of parameters γ and a vector of stochastic processes z_t. Then model evaluation consists in selecting a function $\psi(S_y, S_{x^*}(z_t, \gamma))$ that measures the distance between S_y and S_{x^*} and in assessing its magnitude.

The choice of which stylized facts one wants to match obviously depends on the question asked and on the type of models used. For example, if the question is what is the proportion of actual cyclical fluctuations in GNP and consumption explained by the model, one would choose stylized facts based on variances and covariances of the data. As an alternative to the examination of second moments of the data, one could summarize the properties of actual data via a VAR and study the properties of simulated data, for example, comparing the number of unit roots in the two sets of data (as in Canova, Finn and Pagan (1994)), the size of VAR coefficients (as in Smith (1993) or Ingram, DeJong and Whiteman (1995)) or the magnitude of certain impulse responses (as in Cogley and Nason (1994)). Also, it is possible to evaluate the discrepancy of a model to the data by choosing specific events that one wants the model to replicate e.g. business cycle turning points, (as in King and Plosser (1994) or Simkins (1994)) or variance bounds (as in Hansen and Jagannathan (1991)).

Classical pieces in the calibration literature (see e.g. Kydland and Prescott (1982) or (1991)) are typically silent on the metric one should use to evaluate the quality of the approximation of the model to the data. The approach favored by most calibrators is to glare over the exact definition of the metric used and informally assess the properties of simulated data by comparing them to the set of stylized facts. In this way a researcher treats the experiment conducted as a measurement exercise where the task is to gauge the proportion of some observed statistics reproduced by the theoretical model. This informal approach is also shared by cliometricians (see e.g. Summers (1991)) who believe that rough reproduction of simple sample statistics are everything that is needed to evaluate the implications of the model ("either you see it with naked eyes or no fancy econometric procedure will find it").

There are, however, alternatives to this informal approach. To gain some understanding of the differences among approaches, but at the cost of oversimplifying the matter, we divide evaluation approaches into five classes:

- Informal approaches.
- Approaches which do not consider sampling variability of the actual or simulated data, but instead use the statistical properties of u_t in (1) to impose restrictions on the time series properties of ψ. This allows them to provide an R^2 type measure of fit between the model and the data (see Watson (1993)).
- Approaches which use the sampling variability of the **actual** data (affecting S_y and, in

some cases, estimated γ) to provide a measure of the distance between the model and the data. Among these we list the GMM based approach of Christiano and Eichenbaum (1992), Cecchetti, Lam and Mark (1994) or Fève and Langot (1994), and the frequency domain approaches of Diebold, Ohanian and Berkowitz (1994) and Ortega (1995).

- Approaches which use sampling variability of the **simulated** data to provide a measure of distance between the model and the data. Among these procedures we can distinguish those who take z_t as stochastic and γ as given, such as Gregory and Smith (1991), Söderlind (1994) or Cogley and Nason (1994) and those who take both z_t and γ as stochastic, such as Canova (1994) and (1995).

- Finally, approaches which consider the sampling variability of both the actual and simulated data to evaluate the fit of the model. Once again we can distinguish approaches which, in addition to taking S_y as random, allow for variability in the parameters of the model (keeping z_t fixed) such as DeJong, Ingram and Whiteman (1995) from those which allow for both z_t and γ to vary such as Canova and De Nicoló (1995).

Because the emphasis of this book is on simulation techniques, we will briefly examine the first three approaches and discuss more in detail the last two, which make extensive use of simulation techniques to conduct inference. Kim and Pagan (1994) provide a thorough critical review of several of these evaluation techniques and additional insights on the relationship among them.

The evaluation criteria that each of these approaches proposes is tightly linked to the parameter selection procedure we discussed in the previous section.

A calibrator typically chooses parameters using steady state conditions, and those which do not appear in the steady state equations are chosen a-priori or referring to a particular study. In this case, because S_y is a vector of numbers, there are no free parameters. No uncertainty is allowed in the selected parameter vector; one is forced to use an informal metric to compare the model to the data. This is because, apart from the uncertainty present in the exogenous variables, the model links the endogenous variables to the parameters in a deterministic fashion. Therefore, once we have selected the parameters and we have a realization of S_y it is not possible to measure the dispersion of the distance $\psi(S_y, S_{x^*}(z_t, \gamma))$. From the point of view of the majority of calibrators this is not a problem. As emphasized by Kydland and Prescott (1991) or Kydland (1992), the trust a researcher has in an answer given by the model does not depend on a statistical measure of discrepancy, but on how much he believes in the economic theory used and in the measurement undertaken.

Taking this as the starting point of the analysis Watson (1993) suggests an ingenious way to evaluate models which are known to be an incorrect DGP for the actual data. Watson asks how much error should be added to x_t^* so that its autocovariance function equals the autocovariance function of y_t. Writing $y_t = x_t^* + u_t^*$ where u_t^* includes the approximation error due to the use x_t^* in place of x_t, the autocovariance function of this error is given by

$$A_{u^*}(z) = A_y(z) + A_{x^*}(z) - A_{x^*y}(z) - A_{yx^*}(z) \tag{10}$$

To evaluate the last two terms in (10) we need a sample from the joint distribution of (x_t^*, y_t) which is not available. In this circumstances it is typical to assume that either u_t^* is a measurement error or a signal extraction noise (see e.g. Sargent (1989)), but in the present context none

of the two assumptions is very appealing. Watson suggests to choose $A_{x^*y}(z)$ so as to minimize the variance of u_t^* subject to the constraint that $A_{x^*}(z)$ and $A_y(z)$ are positive semidefinite. Intuitively, the idea is to select $A_{x^*y}(z)$ to give the best possible fit between the model and the data (i.e. the smallest possible variance of u_t^*). The exact choice of $A_{x^*y}(z)$ depends on the properties of x_t^* and y_t, i.e. whether they are serially correlated or not, scalar or vectors, full rank processes or not. In all cases, the selection criteria chosen imply that x_t^* and y_t are perfectly linearly correlated where the matrix linking the two vectors depends on their time series properties and on the number of shocks buffeting the model. Given this framework of analysis, Watson suggests measures of fit statistics, similar to a $1 - R^2$ from a regression, of the form

$$r_j(\omega) = \frac{A_{u^*}(\omega)_{jj}}{A_y(\omega)_{jj}} \tag{11}$$

$$R_j(\omega) = \frac{\int_{\omega \in Z} A_{u^*}(\omega)_{jj} d\omega}{\int_{\omega \in Z} A_y(\omega)_{jj} d\omega} \tag{12}$$

where the first statistic measures the variance of the j-th component of the error relative to the variance of the $j - th$ component of the data for each frequency and the second statistic is the sum of the first over a set of frequencies and may be useful to evaluate the model, say, at business cycle frequencies. It should be stressed that (11) and (12) are lower bounds. That is, when $r_j(\omega)$ or $R_j(\omega)$ are large, the model poorly fits the data. However, when they are small they do not necessarily indicate good fit since the model may still fit poorly the data, if we change the assumptions about $A_{x^*y}(z)$.

To summarize, Watson chooses the autocovariance function of y as the set of stylized facts of the data to be matched by the model, the ψ function as the ratio of A_{u^*} to A_y and evaluates the size of ψ informally. Note that in this approach, γ and z_t are fixed, and A_{x^*} and A_y are assumed to be measured without error.

When a calibrator is willing to assume that parameters are measured with error because, given an econometric technique and a sample, parameters are imprecisely estimated, then model evaluation can be conducted using measures of dispersion for simulated statistics which reflect parameter uncertainty. There are various versions of this approach. One is the criteria of Christiano and Eichenbaum (1992), Cecchetti, Lam and Mark (1994) and Fève and Langot (1994) which use a version of a J-test to evaluate the fit of a model. In this case S_y are moments of the data while ψ is a quadratic function of the type

$$\psi(S_y, S_{x^*}(z_t, \gamma)) = [S_y - S_{x^*}(\gamma)]V^{-1}[S_y - S_{x^*}(\gamma)]' \tag{13}$$

where V is a matrix which linearly weights the covariance matrix of S_{x^*} and S_y, and S_{x^*} is random because γ is random. Formal evaluation of this distance can be undertaken following Hansen (1982): under the null that $S_y = S_{x^*}(z_t, \gamma)$ the statistic defined in (13) is asymptotically distributed as a χ^2 with the number of degrees of freedom equal to the number of overidentifying restrictions, i.e. the dimension of S_y minus the dimension of the vector γ. Note that this procedure is correct asymptotically and that it implicitly assumes that $x_t = f(z_t, \gamma)$ (or its approximation) is the correct DGP for the data and that the relevant loss function measuring the distance between actual and simulated data is quadratic.

Although the setup is slightly different, also the methods proposed by Diebold, Ohanian and Berkowitz (DOB) (1994) and Ortega (1995) can be broadly included into this class of approaches.

For DOB the statistic of interest is the spectral density matrix of y_t and, given a sample, this is assumed to be measured with error. They construct the uncertainty around point estimates of the spectral density matrix by using parametric and nonparametric bootstrap approaches and Bonferroni tunnels to obtain (small sample) 90% confidence bands around these point estimates. On the other hand, they take calibrated parameters and the realization of z_t as given so that the spectral density matrix of simulated data can be estimated without error simply by simulating very long time series for x_t^*. Ortega (1995) also takes the spectral density matrix as the set of stylized facts of the data to be matched by the model. Differently to DOB, she estimates jointly the spectral density matrix of actual and simulated data and constructs the uncertainty around point estimates using asymptotic distribution theory.

In both cases, the measure of fit of the model to the data is generically given by:

$$C(\gamma, z_t) = \int_0^{\pi} \psi(F_y(\omega), F_{x^*}(\omega, \gamma, z_t))W(\omega)d\omega \tag{14}$$

where $W(\omega)$ is a set of weights applied to different frequencies and F are the spectral density matrices of actual and simulated data.

DOB suggest various options for ψ (quadratic, ratio, likelihood type) but do not construct a test statistic to examine the magnitude of ψ. Instead, they compute a small sample distribution of the event that $C(\gamma, z_t)$ is close to zero. Ortega, on the other hand, explicitly uses a quadratic distance function and constructs an asymptotic χ^2 test to measure the magnitude of the discrepancy between the model and the data. The set of tools she develops can also be used to compare the fit of two alternative models to the data and decide which one is the best using asymptotic criteria.

If a calibrator is willing to accept the idea that the stochastic process for the exogenous variables is not fixed, he can then compute measures of dispersion for simulated statistics by changing the realization of z_t while maintaining the parameters fixed. Such a methodology has its cornerstone in the fact that it is the uncertainty in the realization of the exogenous stochastic process (e.g. the technology shock), an uncertainty which one can call extrinsic, and not the uncertainty in the parameters, which one can call intrinsic, which determines possible variations in the sample statistics of simulated data. Once the dispersion of simulated statistics is obtained, the sampling variability of simulated data can be used to evaluate the distance between moments of simulated and of actual data (as e.g. Gregory and Smith (1991) and (1993)).

If one uses such an approach, the evaluation of a model can be conducted with a probabilistic metric using well known techniques in the Monte Carlo literature. For example, one may be interested in finding out in what decile of the simulated distribution the actual value of a particular statistics lies calculating, in practice, the "size" of calibration tests. This approach requires that the evaluator takes the model economy as the true DGP for the data and that differences between S_y and S_{x^*} may occur only because of sampling variability. To be specific, Gregory and Smith take S_y be a set of moments of the data and they assume that they can be

measured without error. Then they construct a distribution of $S_{x^*}(z_t, \gamma)$ by drawing realizations of the z_t process from a given distribution for fixed γ. The metric ψ used is probabilistic, i.e. they calculate the probability $Q = P(S_{x^*} \leq S_y)$, and the judgement of the fit of the model is informal, e.g. measuring how close Q is to 0.5.

An interesting variation on this setup is provided by the work of Söderlind (1994) and Cogley and Nason (1994). Söderlind employs the spectral density matrix of the actual data as the relevant set of statistics to be matched while Cogley and Nason look at a "structural" impulse response function. Söderlind maintains a probabilistic metric and constructs empirical rejection rates for an event (that the actual spectral density matrix of y_t lies inside the asymptotic 90% confidence band for the spectral density matrix of the simulated data), and the event is replicated by drawing vectors z_t for a given distribution. Cogley and Nason choose a quadratic measure of distance which has an asymptotic χ^2 distribution under the null that the DGP for the data is the model. Then they tabulate the empirical rejection rates of the test, by repeatedly constructing the statistic drawing realizations of the z_t vector. To be specific the ψ function is given by

$$\psi_{k,j}(\gamma) = [IRF_{x^*}(z_t^j, \gamma) - IRF_y]V^{-1}[IRF_{x^*}(z_t^j, \gamma) - IRF_y]' \tag{15}$$

where j refers to replication and k to the $k-th$ step, IRF is the impulse response function and V is the asymptotic covariance matrix of the impulse response function of x^*. Because for every k and for fixed j $\psi_{k,j}(\gamma)$ is asymptotically χ^2, they can construct both a simulated distribution for $\psi_{k,j}$ and compare it with a χ^2 and the rejection frequency for each model specification they examine.

In practice, all three approaches are computer intensive and rely on Monte Carlo methods to conduct inference. Also, all three methods verify the validity of the model by assuming that the model is the correct DGP for y_t and computing the "size" of the calibration tests.

The approach of Canova (1994)-(1995) also belongs to this category of methods, but in addition to allowing the realization of the stochastic process for the exogenous variables to vary he also allows parameter variability in measuring the dispersion of simulated statistic. The starting point, as discussed earlier, is that parameters are uncertain not so much because of sample variability, but because there are many estimates of the same parameter obtained in the literature, since estimation techniques, samples and frequency of the data tend to differ. If one calibrates the parameter vector to an interval, instead of to a particular value, and draws values for the parameters from the empirical distributions of their estimates, it is then possible to use intrinsic uncertainty in addition or in alternative to the extrinsic one, to evaluate the fit of the model. The evaluation approach is then very similar to the one of Gregory and Smith: one simulates the model by drawing parameter vectors from the empirical "prior" distribution and realizations of the exogenous stochastic process z_t from some given distribution. Once the empirical distribution of the statistics of the model is constructed is then possible to compute either the size of calibration tests or the percentiles where the statistics found in the actual data lie.

The last set of approaches recently developed in the literature considers the uncertainty present in the statistics of both actual and simulated data to measure the fit of the model

to the data. In essence what these approaches attempt is to formally measure the degree of overlap between the (possibly) multivariate distributions of S_y and S_x using Monte Carlo techniques. There are differences on the way these distributions are constructed. Canova and De Nicoló (1995) use a parametric bootstrap algorithm to construct distributions for the multivariate statistic of the actual data they consider (first and second moments of the equity premium and of the risk free rate). DeJong, Ingram and Whiteman (DIW) (1995), on the other hand, suggest to represent the actual data with a VAR. Then they compute posterior distribution estimates for the moments of interest by drawing VAR parameters from their posterior distribution and constructing, at each replication, the moments of interest from the AR(1) companion matrix representation of the VAR. In constructing distributions of simulated statistics, Canova and De Nicoló take into account both the uncertainty in exogenous processes and parameters while DIW only consider parameter uncertainty. In addition they differ in the way the "prior" uncertainty in the parameters is introduced in the model. The former paper follows Canova (1995) and chooses empirical based distributions for the parameter vector. DIW use subjectively specified prior distributions (typically normal) whose location parameter is set at the value typically calibrated in the literature while the dispersion parameter is free. The authors use this parameter in order to (informally) minimize the distance between the actual and the simulated distributions of the statistics of interest. By enabling the specification of a sequence of increasingly diffuse priors over the parameter vector, the procedure can therefore illustrate whether uncertainty in the choice of the model's parameters can mitigate differences between the model and the data.

There are few differences in the two approaches also in deciding the degree of overlap of the two distributions. Canova and De Nicoló choose a particular contour probability for one of the two distributions and ask how much of the other distribution is inside this contour. In other words, the fit of the model is examined very much in the style of the Monte Carlo literature, where a good fit is indicated by a high probability covering of the two regions. In addition, to study the features of the two distributions, they repeat the exercise varying the chosen contour probability, say, from 50% to 75%, 90%, 95% and 99%. In this way it is possible to detect anomalies in the shape of the two distributions due to clustering of observations in one area, skewness or leptokurtic behavior. In this approach actual data and simulated data are used symmetrically in the sense that one can either ask whether the actual data could be generated by the model, or, viceversa, whether simulated data are consistent with the distribution of the empirical sample. This symmetry allows to a much better understanding of the properties of error u_t in (1) and resembles very much to switching the null and the alternative in standard classical hypothesis testing.

DeJong, Ingram and Whiteman take the point of view that there are no well established criteria to judge the adequacy of a model's "approximation" to reality. For this reason they present two statistic aimed at synthetically measuring the degree of overlap among distributions. One they call Confidence Interval Criterion (CIC) is the univariate version of the contour probability criteria used by Canova and De Nicoló and is defined as

$$CIC_{ij} = \frac{1}{1-\alpha}\int_a^b P_j(s_i)ds_i \tag{16}$$

where s_i, $i = 1, \ldots, n$ is a set of functions of interests, $a = \frac{\alpha}{2}$ and $b = 1 - a$ are the quantiles of

$D(s_i)$, the distribution of the statistic in the actual data, $P_j(s_i)$ is the distribution of simulated statistic where j is the diffusion index of the prior on the parameter vector and $1-\alpha = \int_a^b D(s_i)ds_i$. Note that with this definition, CIC_{ij} ranges between 0 and $\frac{1}{1-\alpha}$. For CIC close to zero, the fit of the model is poor, i.e. either the overlap is small or P_j is very diffuse. For CIC close to $\frac{1}{1-\alpha}$ the two distributions overlap substantially. Note also that if $CIC > 1$, $D(s_i)$ is diffuse relative to $P_j(s_i)$, i.e. the data is found relatively uninformative regarding s_i.

To distinguish among the two possible interpretations when CIC is close to zero, DeJong, Ingram and Whiteman suggest a second summary measure analogous to a t-statistic for the mean of $P_j(s_i)$ in the $D(s_i)$ distribution, i.e.,

$$d_{ji} = \frac{EP_j(s_i) - ED(s_i)}{\sqrt{varD(s_i)}} \tag{17}$$

Large values of (17) would indicate that the location of $P_j(s_i)$ is quite different from the location of $D(s_i)$.

The final problem is to choose α. DeJong, Ingram and Whiteman fix a particular value ($\alpha = 0.01$) but, as in Canova and De Nicoló, varying α for a given j is probably a good thing to do in order to study the shape of the distributions. This is particularly useful when we are interested in partitions of the joint distributions of s_i and/or graphical methods are not particularly informative about distributions in high dimensional spaces.

5 Policy Analyses

Although it is not the purpose of this chapter to discuss in detail the use of calibrated models for policy analyses, it is useful to briefly describe how they can be undertaken and what they involve. As we have already mentioned, a model is typically calibrated to provide quantitative answer to very precise questions. Some of these questions have potential policy implications: If one wants to be confident in the answer given by the model, it is necessary to undertake sensitivity analysis to check how results change when certain assumptions are modified.

As we have seen, the answers of the model come in the form of continuous functions $h(x_t^*) = h(g(z_t, \gamma))$ of simulated data. In theory, once g has been selected, the uncertainty in h is due to the uncertainty in γ and in z_t. In standard calibration exercises the γ vector is fixed; it is therefore typical to examine the sensitivity of the results in the neighborhood of the calibrated values for γ. Such experiments may be local, if the neighborhood is small, or global, in which case one measures the sensitivity of the results to perturbations of the parameters over the entire range. This type of exercises may provide two types of information. First, if results are robust to variations of a parameter in a particular range, then its exact measurement is not crucial. In other words, the uncertainty present in the choice of such a parameter does not make the answers of the model tenuous and economic inference groundless. On the other hand, if results crucially depend on some parameters, it is clearly necessary to improve upon existing measurement of the calibrated parameters.

A local sensitivity analysis can be undertaken informally, replicating the experiments for different values of the parameters (as in Kydland and Prescott (1982)) or more formally, calculating the elasticity of h with respect to γ (as in Pagan and Shannon (1985)). A global

sensitivity analysis can be efficiently undertaken with Monte Carlo methods or numerical semi-deterministic techniques (see e.g. Niederreiter (1988)) if the function g is known, once the distribution of the γ vector is specified. If g is only an approximation to the functional linking x to z and γ, one can use techniques like Importance Sampling (see Geweke (1989)) to take into account this additional source of uncertainty. Clearly the two types of sensitivity analysis are not incompatible and should both be undertaken to assess the degree of trust a researcher can attach to the answer given by the model. Finally, one should note that the type of sensitivity analysis one may want to undertake depends also on the way parameters are selected and models evaluated. For example, if one uses the approach of Canova (1994)-(1995) or DeJong, Ingram and Whiteman (1995), the evaluation procedure automatically and efficiently provides sensitivity analysis to global perturbations of the parameters within an economic reasonable range.

Once the answer of the model has been made robust to variations of the parameters, a researcher undertakes policy analyses by changing the realization of the stochastic process for z_t or varying a subset of the γ vector, which may be under the control of, say, the government. Analyses involving changes in the distribution of z_t of in the g function are also possible, but care should be exercised in order to compare results across specifications.

6 An example

In the field of international economics, robust stylized facts are usually hard to obtain. One of the most stable regularities observed in the data is the fact that national saving rates are highly correlated with national investment rates, both in time series analysis of individual countries and in cross sections in which the average over time of these variables is treated as a single data point for each country. Moreover, high saving and investment correlations arise in small economies as well as in large ones, although the correlation tends to be lower for smaller countries. These findings have been interpreted as indicating that the world is characterized by capital immobility. Yet most economists believe that the world is evolving toward an increasingly higher degree of international capital mobility. Baxter and Crucini (1993) have provided a model in which there is perfect mobility of financial and physical capital (so that the degree of international capital mobility is high) which generates high time series correlations of saving and investment, therefore solving the apparent puzzle present in the data. Their analysis is entirely within the standard Kydland and Prescott approach, i.e. the parameters are fixed at some reasonably chosen values, no uncertainty is allowed in actual and simulated statistics and the metric used to compare actual and simulated data is informal.

The task of this section is to study whether their model is indeed able to solve the puzzle when its performance is formally examined with the tools described in the previous section. We will provide a generic measure of fit of the model to the data using several variants of the simulation-based procedures we described in previous sections and compare and contrast the outcomes of the evaluation procedure with the ones proposed by more standard measures of fit.

6.1 The model

It is a two country model with a single consumption good. Each country is populated by a large number of identical agents and labor is assumed to be immobile across countries. Preferences of the representative agent of country $h = 1, 2$ are given by:

$$U \equiv E_0 \sum_{t=0}^{\infty} \frac{\beta^t}{1-\sigma} [c_{ht}^{\mu} l_{ht}^{(1-\mu)}]^{1-\sigma} \tag{18}$$

where c_{ht} is private consumption of the single composite good by the representative agent of country h and l_{ht} is leisure, β is the discount factor, σ the coefficient of relative risk aversion and μ the share of consumption in utility. Leisure choices are constrained by:

$$0 \leq l_{ht} + N_{ht} \leq 1 \qquad \forall\ h \tag{19}$$

where the total endowment of time in each country is normalized to 1 and N_t represents the number of hours worked. The goods are produced with a Cobb-Douglas technology:

$$Y_{ht} = A_{ht}(K_{ht}^{1-\alpha})(X_{ht}N_{ht})^{\alpha} \qquad h = 1, 2 \tag{20}$$

where K_t is the capital input, α is the share of labor in GDP, and where $X_{ht} = \lambda X_{ht-1}\ \forall h$ with $\lambda \geq 1$. X_{ht} represents labor-augmenting Harrod-neutral technological progress with deterministic growth rate equal to λ. Production requires domestic labor and capital inputs and is subject to a technological disturbance A_{ht} with the following properties:

$$\begin{bmatrix} A_{1t} \\ A_{2t} \end{bmatrix} = \begin{bmatrix} \rho & \nu \\ \nu & \rho \end{bmatrix} * \begin{bmatrix} A_{1t-1} \\ A_{2t-1} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}$$

where $e_t = [e_{1t} e_{2t}]' \sim N(0, \begin{bmatrix} \sigma^2 & \phi \\ \phi & \sigma^2 \end{bmatrix})$ and where the parameter ϕ controls for the contemporaneous spillover of the shocks.

Capital goods are accumulated according to:

$$K_{ht+1} = (1 - \delta_h)K_{ht} + v(I_{ht}/K_{ht})K_{ht} \qquad h = 1, 2 \tag{21}$$

where $v(\frac{I_{ht}}{K_{ht}}) > 0$ is concave and represents the cost of installing new capital or moving new capital from the location where it is produced to the other country (costs of adjustment in capital). As explained in Baxter and Crucini (1993), there is no need to specify a functional form for v, it is sufficient to describe its behavior near the steady state by specifying some parameters. One is $\frac{1}{v'}$ which corresponds to Tobin's Q, i.e. the price of existing capital in one location relative to the price of new capital. Another one is the elasticity of the marginal adjustment cost function, $\xi_{v'}$.

Governments finance their consumption purchases, g_{ht}, by taxing national outputs with a distorting tax and then transferring what remains back to domestic residents. For simplicity we assume that $g_{ht} = g_h$, $\forall t$. The government budget constraint is given by:

$$g_h = TR_{ht} + \tau_h Y_{ht} \qquad \forall\ h \tag{22}$$

where τ_h are tax rates and TR_h are lump sum transfers in country h.

The economy wide resource constraint is given by:

$$\pi*(Y_{1t}-g_{1t}-c_{1t}-k_{1t+1})+(1-\pi)*(Y_{2t}-g_{2t}-c_{2t}-k_{2t+1})\geq -(1-\delta_1)k_{1t}-(1-\delta_2)k_{2t} \quad (23)$$

where π is the fraction of world population living in country 1 and where we have implicitly accounted for the fact that new investment is costly.

Finally, following Baxter and Crucini (1993) we assume complete financial markets and free mobility of financial capital across countries so that agents can write and trade in every kind of contingent securities they wish. To find a solution to the model we first detrend those variables which drift over time by taking ratios of the original variables with respect to the labor augmenting technological progress, e.g. $y_{ht}=\frac{Y_{ht}}{X_{ht}}$, etc. Second, since there are distortionary taxes in the model, the competitive equilibrium is not Pareto optimal and the competitive solution differs from the social planner's solution. As in Baxter and Crucini (1993) we solve the problem faced by a pseudo social planner, modifying the optimality conditions to take care of the distortions. The weights in the social planner problem are chosen to be proportional to the number of individuals living in each of the countries. The modified optimality conditions are then approximated with a log-linear expansion around the steady state as in King, Plosser and Rebelo (1988). Time series for saving and investment in each of the two countries are computed analytically from the approximate optimality conditions. The variances and the correlation (or the spectra and the coherence) of saving and investment of actual and simulated data are computed after eliminating from the raw time series a linear trend.

The parameters of the model are $\gamma=[\beta,\ \mu,\ \sigma,\ \alpha,\ \rho,\ \nu,\ \sigma^2,\ \xi_{v'},\ \pi,\ \tau,\ \lambda]$ plus the steady state Tobin's Q which we set equal to 1. The exogenous processes of the model are the two productivity disturbances so that $z_t=[A_{1t}A_{2t}]'$.

The actual data we use are aggregate basic saving (i.e. computed as S=Y-C-G) and investment for the period 1971:1-1993:4 for the US and for Europe, measured in constant prices, seasonally adjusted, in billions of US dollars using 1985 average exchange rates and are from OECD Main Economic Indicators.

The statistics we care about are the diagonal elements of the 4×4 spectral density matrix of the data and the coherences between saving and investment of the two "countries". Although interesting, cross country coherences are not considered here for the sake of presentation. To maintain a close relationship between the model and the actual data we linearly detrend the actual data before the spectral density matrix is computed. Plots of the detrended series appear in figure 1.

In the basic experiment the parameters of the model are the same as in Baxter and Crucini and are reported in the first column of table 1. When we allow for parameters to be random we take two approaches: the one of Canova and the one of DeJong, Ingram and Whiteman. In the first case empirical based distributions are constructed using existing estimates of these parameters or, when there are none, choosing a-priori an interval on the basis of theoretical considerations and imposing uniform distributions on it. The distributions from which the parameters are drawn and their features appear in the second column of table 1. In the second case distributions for the parameters are assumed to be normal, with means which are the same as the basic calibrated parameters while dispersions are a-priori chosen. The third column of table 1 report these distributions for the parameters of interest.

In each case we generate samples of 96 observations to match the sample size of actual data. Because the initial conditions for the capital stock are set arbitrarily, the initial 50 observations for each replication of the model are discarded to eliminate initial condition problems. The number of replications for this preliminary version of the paper is set to 100 because of time constraints, but in later versions we hope to be able to present results obtained with a much larger number of replications.

6.2 The Results

To evaluate the fit of the model to the data we use four different approaches. The first one, which we use as a benchmark, is the one of Watson. Given the spectral density matrix of the actual saving and investment for the two economic blocks, we calculate the spectral density matrix of the approximation error and compute the measure of fit (11). Since in the model there are two technology disturbances, the spectral density matrix of simulated saving and investment for the two countries is singular and of rank equal to two. Therefore, to minimize the variance of the approximation error we consider two different identification schemes: one where we minimize jointly the error term of the saving and investment of the first country and one where we minimize jointly the saving and investment errors of the second country. Note that to generate the measure (11) we make two important assumptions: (i) that the spectral density matrix of the actual and simulated data can be measured without error, (ii) that the parameters of the model can be selected without error. We report the results of our test graphically in figure 2. Column 1 refers to the case where the variance of saving and investment of the first country (the US) is minimized, Column 2 to the case where the variance of saving and investment of the second country (Europe) is minimized. Vertical bars comprise business cycle frequencies (cycles of 3-8 years).

We see that the model does sufficiently well in replicating the features of the spectral density matrix of the data when we minimize the errors of US saving and investment. In particular, at business cycle frequencies, model's generated data are more volatile than actual US data, they are slightly less volatile than European saving, and substantially less volatile than European investment. This features shows up also in terms of coherences: the coherences of US saving and investment is well matched at business cycle frequencies, while the coherences of European saving and investment is not. When we minimize the errors of European saving and investment a similar picture emerges. Two features however need to be mentioned. First, under this identifying restriction the model's performance is much worse at low frequencies, for all four variables. Second, the variance of European investment at business cycle frequencies is definitely worsely matched relative to the first identification scheme but the coherence of European saving and investment is better matched. Overall, these statistics suggest that the model accounts for the second order properties of US data better than the second order properties of European data. There are variations across frequencies but the model seems to be reasonable, especially at business cycle frequencies. Also, the two identification schemes produce the same qualitative features even though, quantitatively, the first one appears to do better.

To show how Monte Carlo techniques we have discussed in this paper can be used to evaluate the quality of the model's approximation to the data we compute three types of statistics. First we report how many times, for each frequency, the diagonal elements of the

spectral density matrix and the coherences of the model generated data lie within a 95% confidence band for the diagonal elements of the spectral density matrix and the coherences of actual data. If the spectral density matrix of the actual data is taken to be the object of interest to be replicated, these numbers report the the "power" of a test which assumes that the model is the correct DGP for the actual. If we are not willing to assume that the model is the correct DGP for the actual data, these numbers judge the quality of the approximation by informally examining the magnitude of the probability coverings. No matter which interpretation we take, a number as close as possible to 0.95 at a particular frequency would indicate a "good" model performance.

We compute 95% confidence bands for the actual data in two ways: using asymptotic distribution theory (as in Ortega (1995)) and using a version of the parameteric bootstrap procedure of Diebold, Ohanian and Berkowitz (1995). In this latter case, we run a four variables VAR with 6 lags and a constant, construct replications for saving and investment for the two countries by bootstrapping the residuals of the VAR model, estimate the spectral density matrix of the data for each replication and extract 95% confidence bands after ordering the replications, frequency by frequency.

Replications for the time series generated by the model are constructed using Monte Carlo techniques in three different ways. In the first case we simply randomize on the innovations of the technology shocks, keeping their distribution fixed (as in Gregory and Smith (1991)). In the second and third cases parameters are random and drawn from the distributions listed in the second and third columns of table 1. Once again we present results of our simulation graphically. To economize on space and because simulated results are similar when the parameters are randomly drawn from the two sets of distributions and when the 95% confidence bands for actual data are computed asymptotically or by bootstrap, in figure 3 we present the percentage of times the model spectra is inside the asymptotic 95% band, frequency by frequency, when only the stochastic processes of the model are randomized, while in figure 4 we present the percentage of times the model spectra is inside the small sample 95% band, again frequency by frequency, when we randomize on the stochastic processes of the model and the parameters of the model are drawn from empirically based distributions.

The message contained in these statistics is somewhat different from those contained in figure 2. First of all, and considering figure 3, the percentage of times the model spectra is inside the 95% band for the actual spectra is very similar for US saving and investment, it is high at business cycle frequencies, but is very different across the frequencies of the spectrum. Despite this favorable partially outcome, the coherence between simulated saving and investment is rarely inside the 95% band for the coherence of actual saving and investment. In general, simulated correlation are too high with the given parameter values, and this is clearly the case at business cycle frequencies where the largest percentage at any frequency in this range is only 6%. For European variables the match is much worse than for US variables, in particular for investments, where at most 4% of the times the diagonal elements simulated spectra are inside the band for the diagonal elements of the actual spectra at each frequency. On the contrary, the performance of the model in terms of coherence is, relatively speaking, more acceptable even though the model appears to be failing at business cycle frequencies.

A different picture emerges when we randomize on the parameters and use small sample 95% confidence bands (figure 4). The model is doing very well in matching variance of US saving

and investment over all frequencies and does a reasonable job for the variance of European savings. However, the performance for European investment is definitely unacceptable while for coherences there is no substantial changes relative to the situation described in figure 3. Of the two changes we have introduced, parameter uncertainty is responsible for the improved performances for US variables, while substitution of small sample confidence measures for asymptotic ones alter the quantitative but not the qualitative features of the results.

In sum, figures 3 and 4 suggest that the model is doing better in matching the variance of US saving and investment than the variance of European saving and investment, but also that the coherences of saving and investment across the two continental blocks are equally mismatched. Therefore, contrary to Baxter and Crucini's claim, the model seems to fail to generate the same type of saving and investment correlation we see in the data (in general, they are too high).

The second statistic we compute is the percentile of the simulated distribution of the spectral density matrix of saving and investment for the two countries where the value of the spectral density matrix of actual data, taken here to be estimated without an error, lies, frequency by frequency. Implicitly, this p-value reports, for each frequency, the proportion of replications for which the simulated data is less than the historical value. In other words, if $\bar{S}_y(\omega)$ is the spectral density matrix of the actual data we report $P[-\infty < S_x(\omega) \leq \bar{S}_y(\omega)] = \int_{-\infty}^{\bar{S}_y}(\omega)p(x)dx$ where $p(x)$ is the empirical distribution of the simulated spectral density matrix for the four series. Seen through these lenses the sample spectral density matrix is treated as a "critical value" in examining the validity of the theoretical model. Values close to zero (one) indicate that the actual spectral density matrix is in left (right) tail of the simulated distribution of the spectral density matrix of simulated data at that particular frequency, in which case the model performs poorly in reproducing second moments of saving and investment in the two countries. Values close to 0.5 indicate that the actual spectral density matrix at those frequencies is close to the median of the simulated distribution of the spectral density matrix of simulated data. In this latter case, the model does a good job in reproducing the data at those frequencies. Note also that values in the range $[\alpha, 1-\alpha]$, where α is a chosen confidence size, indicate that the model is not significantly at odds with the data.

Also in this case we report results graphically. In figure 5, we present the percentile for the four diagonal elements of the spectral density matrix and the two coherences, when only the innovations of the technology disturbances are randomized. In figure 6 we present results when also the parameters of the model are randomized. Once again results do not depend very much on the exact prior distribution used for the parameters, and in this figure we report results using normally distributed priors.

Figure 5 indicates that, in general, the model with fixed parameters tends to generate too much variability both for US and European variables at the most interesting set of frequencies. In particular, at business cycle frequencies, the diagonal elements of the spectral density matrix of the actual data lie between the 25 and the 40 percentile of the simulated distribution. Note also that at high frequencies the model does not generate enough variability. Similarly, the model generates a correlation between saving and investment which is too high for both continental blocks: at business cycle frequencies, the actual coherences are in the first decile of the simulated coherences. Adding parameter variability improves the picture in one direction but worsens it in another (see figure 6). The diagonal elements of the spectral density matrix

of the actual data are at most in fourth percentiles of the simulated distribution, regardless of the frequency, suggesting that the model variability is definitely too high. However, such a high level of variability generates a distribution of coherence by frequencies which is much wider than the one obtained in figure 5. Consequently, at least at business cycle frequencies, the model appears to do much better and, on average, the actual coherence is around the 50% of simulate coherence for both US and Europe saving and investment. Note, however, that there are substantial differences within business cycle frequencies.

Finally, we compute by Monte Carlo methods the distributional properties of the log of the approximation error, i.e. the error needed to match the spectral density matrix of the actual data, given the model's simulated spectral density matrix. To compute the distributional properties of the log of the error, we draw, at each replication, parameters and innovations from the posterior distribution of the VAR representation of the actual data, construct time series of interest following the procedure of DeJong, Ingram and Whiteman (1995) and estimate the spectral density matrix of the four series. At each replication, we also draw parameters and innovations from the distributions presented in table 1 and construct the spectral density matrix of simulated data. Let $\log(S_u^i(\omega)) = \log(\frac{S_y^i(\omega)}{S_x^i(\omega)})$ be the log error in matching the spectral density matrix of the data, $S_y^i(\omega)$ at replication i. By drawing a large number of replications we can construct a nonparametric estimate of this distribution (using e.g. kernels) and compute moments and fractiles at each frequency. If the model was the correct DGP for the data, the distribution for this error would be degenerated at each frequency. If the model captures the serial correlation properties of the data well, then the distribution of this error should be clustered around the same values for each frequency of the spectrum. Finally, if the bands are tight, then the error has a small variance and the uncertainty in estimated actual and simulated spectra is not important in evaluating the quality of the approximation of the model to the data. Figure 7 presents the 90% confidence range at each frequency for the six elements of interest of the spectral density matrix. Once again, we have performed the calculations randomizing both on the stochastic processes of the model and on the stochastic process and the parameters of the model. The qualitative features of the results are pretty similar across the two sets of replications. Therefore, we report only results where simulated time series incorporates the two types of uncertainty.

The results suggest that the model fails to generate enough variability for all variables at business cycle frequencies. Moreover, the magnitude of the failure is different across frequencies: it is much more evident at low frequencies and is still relevant, but smaller at higher frequencies. Overall, the model does not generate enough variability at low and business cycle frequencies for the US and at business cycle frequencies for Europe. In terms of coherences, the model appears to be again inappropriate at low frequencies while at business cycle frequencies the 90% range for the error is within a reasonable range. Finally, uncertainty in estimating spectra appears to be more of a problem for US variables: the band for European variables is relatively tighter and the distribution closer to normal.

7 Conclusions

The task of this chapter was to illustrate how simulation techniques can be used to evaluate the quality of a model approximation to the data, where the basic theoretical model design is one which fits into what we call calibration exercise. In section 2 we provide first a definition of what calibration is and describe in details the steps needed to generate time series from the model and to select relevant statistics of actual and simulated data. In section 3 we overview four different types of approaches which have been suggested in the literature, comparing and contrasting them on the basis of what type of variability they use to judge the closeness of the model's approximation to the data. In section 4 we describe how we can undertake policy analysis with models which have been calibrated and evaluated along the lines discussed in the previous two sections. Section 5 presents a concrete example borrowed from Baxter and Crucini (1993) where we evaluate whether or not the model is able to reproduce the spectral density matrix of saving and investment for the US and Europe. Here we present four different simulation based statistics which allow us to get different perspectives on the quality of the model approximation to the data. We show that, contrary to Baxter and Crucini's claims, the model fails according to some of our measures to account for the coherence of saving and investment in the two continental blocks and we stress that the model is more at odds with the actual data at low and business cycle frequencies than at high frequencies. Overall, the example clearly demonstrates that simulation based evaluation techniques are a very useful tool to judge the quality of the approximation of fully specified general equilibrium models to the data and may uncover features of the model which are left hidden by more simple but more standard informal evaluation techniques.

References

[1] Baxter, M. (1991) "Approximating Suboptimal Dynamic Equilibria: An Euler Equation Approach", *Journal of Monetary Economics*, 27, 173-200.

[2] Baxter, M. and Crucini, M. (1993) "Explaining Saving-Investment Correlations", *American Economic Review*, 83, 416-436.

[3] Canova, F. (1994) " Statistical Inference in Calibrated Models", *Journal of Applied Econometrics*, 9, S123-S144.

[4] Canova, F. (1995) " Sensitivity Analysis and Model Evaluation in Simulated Dynamic General Equilibrium Economies", *International Economic Review*, 36, 477-501.

[5] Canova, F. and De Nicoló, G. (1995), " The Equity Premium and the Risk Free Rate: A Cross Country, Cross Maturity Examination", CEPR working paper 1119.

[6] Canova, F., Finn, M. and Pagan, A. (1994), "Evaluating a Real Business Cycle Model", in C. Hargreaves, *Nonstationary Time Series Analyses and Cointegration*, Oxford, UK: Oxford University Press.

[7] Cecchetti, S.G., Lam, P. and N. Mark (1994), "Testing Volatility Restrictions on Intertemporal Marginal Rates of Substitution Implied by Euler Equations and Asset Returns", *Journal of Finance*, 49, 123-152.

[8] Christiano, L. and M. Eichenbaum (1992), " Current Business Cycle Theories and Aggregate Labor Market Fluctuations", *American Economic Review*, 82, 430-450.

[9] Cogley, T. and Nason, J.M. (1994), "Testing the Implications of Long-run Neutrality for Monetary Business Cycle Models", *Journal of Applied Econometrics*, 9, S37-S70.

[10] Coleman, W. (1989) "An Algorithm to Solve Dynamic Models", Board of Governors of the Federal Reserve System, International Finance Division, Discussion Paper no. 351.

[11] Danthine, J.P. and Donaldson, (1992), "Non-Walrasian Economies", Cahiers de Recherche Economique, Université de Lausanne, No.9301.

[12] DeJong, D., Ingram, B. and C. Whiteman, (1995), "A Bayesian Approach to Calibration", forthcoming, *Journal of Business and Economic Statistics.*

[13] Diebold, F., Ohanian, L. and J. Berkowitz, (1994), "Dynamic Equilibrium Economies: A Framework for Comparing Models and Data", University of Pennsylvania, manuscript.

[14] Duffie, D. and Singleton, K. (1993, "Simulated Moments Estimation of Markov Models of Asset Prices", *Econometrica*, 61, 929-950.

[15] Eberwin, and Kollintzas, T. (1995), *Review de Economic et Statistics*,

[16] Fève, P. and Langot, F. (1994), "The RBC Models through Statistical Inference: An Application with French Data", *Journal of Applied Econometrics*, 9, S11-S37.

[17] Gali, J. (1992), "How Well Does the IS-LM Model Fit Postwar U.S. Data?", *Quarterly Journal of Economics*, 107, 709-738.

[18] Geweke, J. (1989), " Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317-1339.

[19] Genest, C. and Zidak, M. (1986) "Combining Probability Distributions: A Critique and an Annotated Bibliography", *Statistical Science*, 1, 114-148.

[20] Gregory, A. and Smith, G. (1989), " Calibration as Estimation", *Econometric Reviews*, 9(1), 57-89.

[21] Gregory, A. and Smith, G. (1991), " Calibration as Testing: Inference in Simulated Macro Models", *Journal of Business and Economic Statistics*, 9(3), 293-303.

[22] Gregory, A. and Smith, G. (1993), " Calibration in Macroeconomics", in Maddala, G.S. (ed.), *Handbook of Statistics*, vol. 11, Amsterdam, North Holland.

[23] Hansen, L. (1982), " Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, 50, 1029-1054.

[24] Hansen, L. and Sargent, T. (1979), "Formulating and Estimating Dynamic Linear Rational Expectations Models", *Journal of Economic Dynamic and Control*, 2, 7-46.

[25] Hansen, L. and Jagannathan, R. (1991), "Implications of Security Market Data for Models of Dynamic Economies", *Journal of Political Economy*, 99, 225-262.

[26] Journal of Business and Economic Statistics, January 1990.

[27] King, R., and Plosser, C. (1994), " Real Business Cycles and the test of the Adelmans" , *Journal of Monetary Economics*, 33, 405-438.

[28] King, R., Plosser, C. and Rebelo S. (1988), " Production, Growth and Business Cycles: I and II", *Journal of Monetary Economics*, 21, 195-232 and 309-342.

[29] Kim, K. and Pagan, A. (1994) "The Econometric Analysis of Calibrated Macroeconomic Models", forthcoming, in Pesaran, H. and M. Wickens, eds., *Handbook of Applied Econometrics*, vol.I, London: Blackwell Press.

[30] Kollintzas, T. (1992)," Comment to J.P. Danthine: Calibrated Macroeconomic Models:What and What for", manuscript, Athens University.

[31] Kuhn, T. (1970), *The Structure of Scientific Revolution*, Chicago, Il.: Chicago University Press.

[32] Kydland, F. (1992) " On the Econometrics of World Business Cycles", *European Economic Review*, 36, 476-482.

[33] Kydland, F. and Prescott, E. (1982), "Time To Build and Aggregate Fluctuations", *Econometrica*, 50, 1345-1370 .

[34] Kydland, F. and Prescott, E. (1991), "The Econometrics of the General Equilibrium Approach to Business Cycles", *The Scandinavian Journal of Economics*, 93(2), 161-178.

[35] Marcet, A. (1992) "Solving Nonlinear Stochastic Models by Parametrizing Expectations: An Application to Asset Pricing with Production", Universitat Pompeu Fabra, working paper 5.

[36] Marcet, A. (1995), "" in Sims, C. (ed.)

[37] Mehra, R. and Prescott, E. (1985), "The Equity Premium: A Puzzle", *Journal of Monetary Economics*, 15, 145-162.

[38] McGratten, E. , Rogerson, B. and Wright, R. (1993), "Estimating the Stochastic Growth Model with Household Production", Federal Reserve Bank of Minneapolis, manuscript.

[39] Niederreiter, H. (1988), " Quasi Monte Carlo Methods for Multidimensional Numerical Integration", *International Series of Numerical Mathematics*, 85, 157-171.

[40] Novales, A. (1990), "Solving Nonlinear Rational Expectations Models: A Stochastic Equilibrium Model of Interest Rates", *Econometrica*, 58, 93-111.

[41] Ortega, E. (1995), "Assessing and Comparing Computable Dynamic General Equilibrium Models", manuscript, European University Institute.

[42] Pagan, A. (1994), "Calibration and Econometric Research: An Overview", *Journal of Applied Econometrics*, 9, S1-S10.

[43] Pagan, A. and Shannon, (1985), "Sensitivity Analysis for Linearized Computable General Equilibrium Models", in J.Piggott and J. Whalley (eds.) *New Developments in Applied General Equilibrium Analysis*, Cambridge: Cambridge University Press.

[44] Pesaran, H. and Smith, R. (1992), "The Interaction between Theory and Observation in Economics", University of Cambridge, manuscript.

[45] Sargent, T. (1987), *Dynamic Macroeconomic Theory*, Cambridge, Ma: Harvard University Press.

[46] Showen, J. and Whalley, J. (1984), "Applied General Equilibrium Models of Taxation and International Trade: An Introduction and Survey", *Journal of Economic Literature*, 22, 1007-1051.

[47] Simkins, S.P. (1994), "Do Real Business Cycle Models Really Exhibit Business Cycle Behavior?", *Journal of Monetary Economics*, 33, 381-404.

[48] Smith, T. (1993) "Estimating Nonlinear Time Series Models Using Simulated VAR", *Journal of Applied Econometrics*, 8, S63-S84.

[49] Söderlind, P. (1994), "Cyclical Properties of a Real Business Cycle Model", *Journal of Applied Econometrics*, 9, S113-S122.

[50] Summers, L. (1991), "Scientific Illusion in Empirical Macroeconomics", *Scandinavian Journal of Economics*, 93(2), 129-148.

[51] Tauchen, G. and Hussey, R. (1991) "Quadrature Based Methods for obtaining Approximate Solutions to Integral Equations of Nonlinear Asset Pricing Models", *Econometrica*, 59, 371-397.

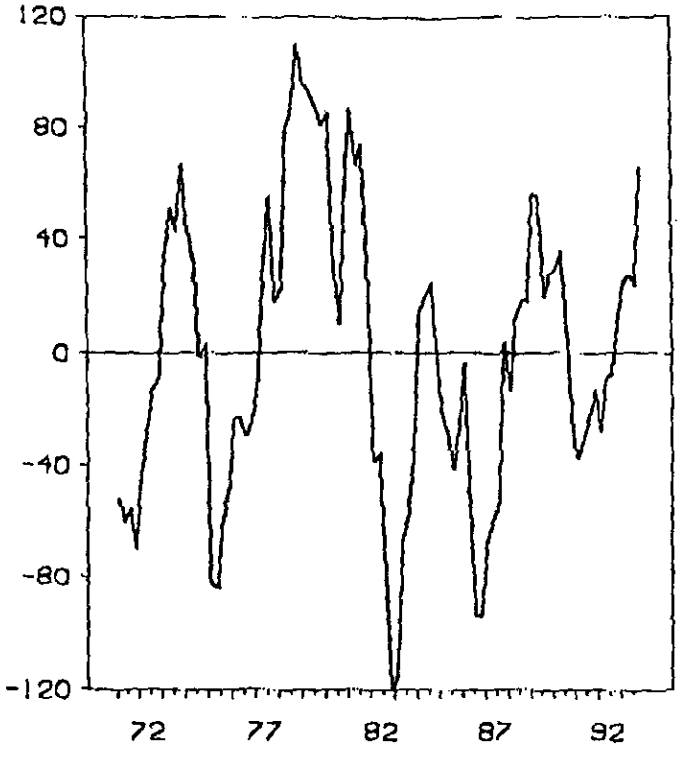
[52] Watson, M. (1993) "Measures of Fit for Calibrated Models", *Journal of Political Economy*, 101, 1011-1041.

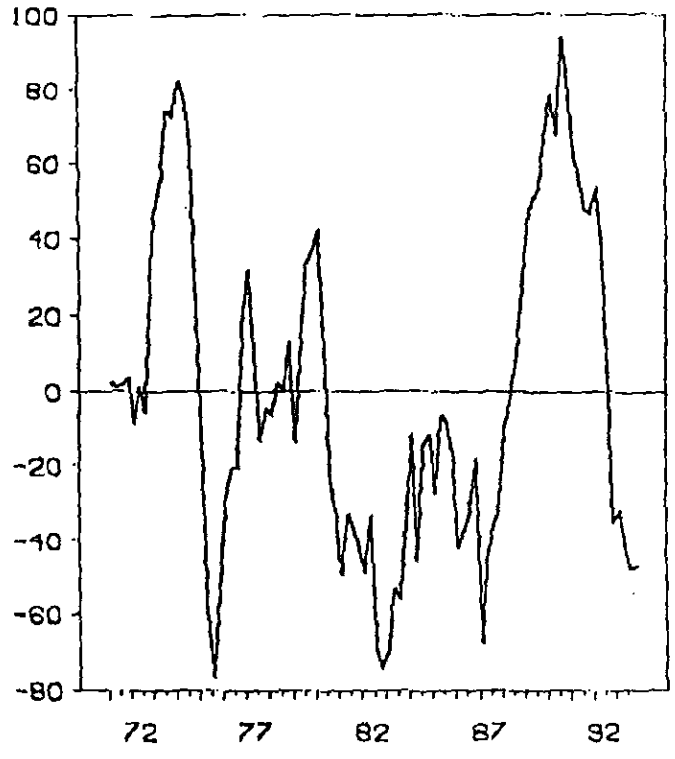
[53] Wolf, F. (1986) *Meta-Analysis: Quantitative Methods for Research Synthesis*, Beverly Hill, Ca.: Sage Publishers.

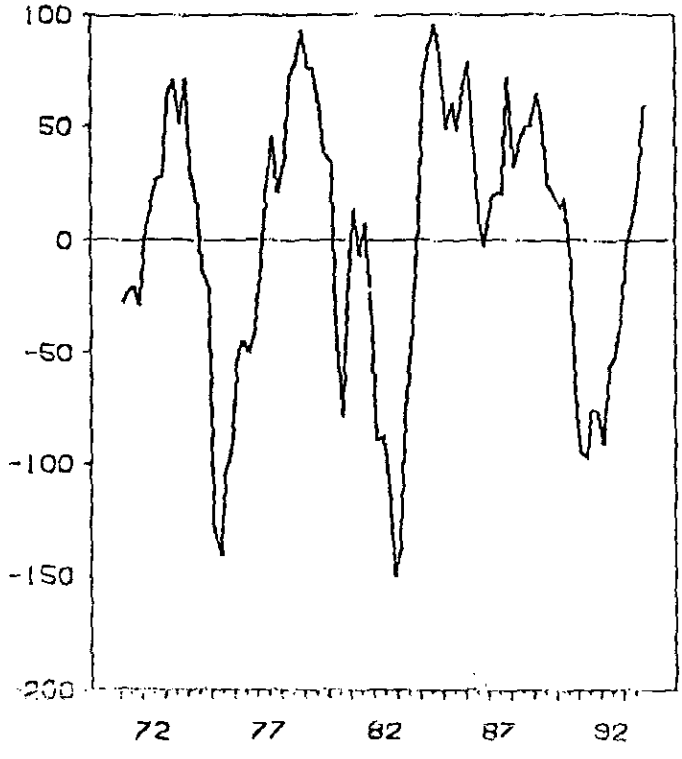
Table 1: Parameter values used in the simulation

Parameter	Basic	Empirical Density	Prior Density
Steady State hours ($\bar{H}$)	0.2	Uniform[0.2, 0.35]	Normal(0.2, 0.02)
Discount Factor (β)	0.9873	Truncated Normal[0.9855, 1.002]	Normal(0.9873, 0.01)
Risk Aversion (σ)	2	Truncated $\chi^2(2)[0, 10]$	Normal(2, 1)
Share of Labor in Output (α)	0.58	Uniform[0.50, 0.75]	Normal(0.58, 0.05)
Growth rate (λ)	1.004	Normal(1.004, 0.001)	1.004
Depreciation Rate of Capital (δ)	0.025	Uniform[0.02, 0.03]	Normal(0.025, 0.01)
Persistence of Disturbances (ρ)	0.93	Normal(0.93, 0.02)	Normal(0.93, 0.025)
Lagged Spillover Disturbances (ν)	0.05	Normal(0.05, 0.03)	Normal(0.05, 0.02)
Standard Deviation of Technology Innovations (σ_e)	1	Truncated $\chi^2(1)$ [0, 0.0091]	Normal(1, 0.004)
Contemporaneous Spillover (ϕ)	0.40	Normal(0.35, 0.03)	Normal(0.4, 0.02)
Country Size (π)	0.50	Uniform[0.10, 0.50]	0.5
Elasticity of marginal adjustment cost function ($\xi_{\nu'}$)	-0.075	-0.075	-0.075
Steady State Tobin's Q (ψ')	1.0	1.0	1.0
Tax Rate (τ)	0.0	0.0	0.0

Notes: "Empirical density" refers to distributions for the parameters constructed using either existing estimates or a-priori intervals as in Canova (1994). "Prior density" refers to distributions for the parameters which are a-priori chosen as in DeJong, Ingram and Whiteman (1995). Inside brackets there are the upper and the lower the range for the parameters. Inside parentheses there are the mean and the standard deviation for the distribution.







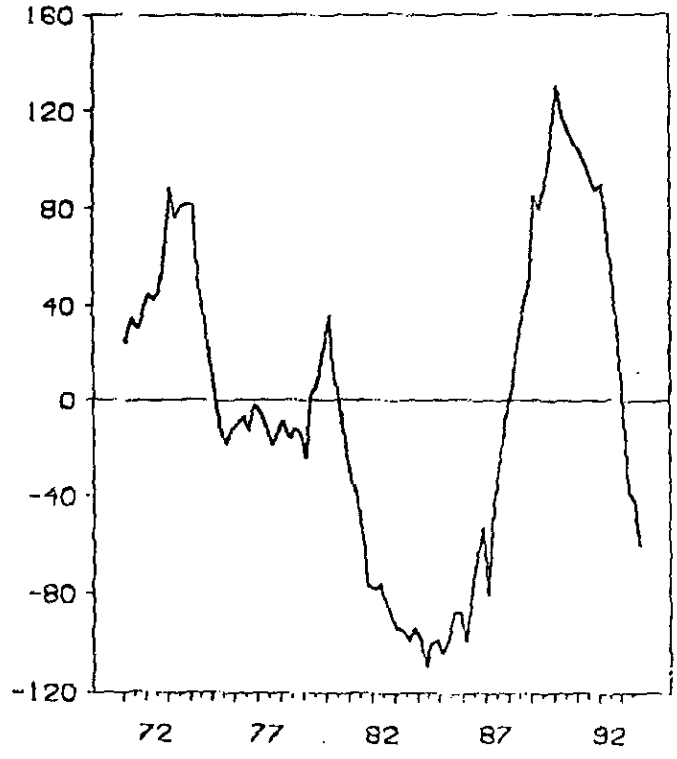


FIGURE 1: Time series plots, detrended data.

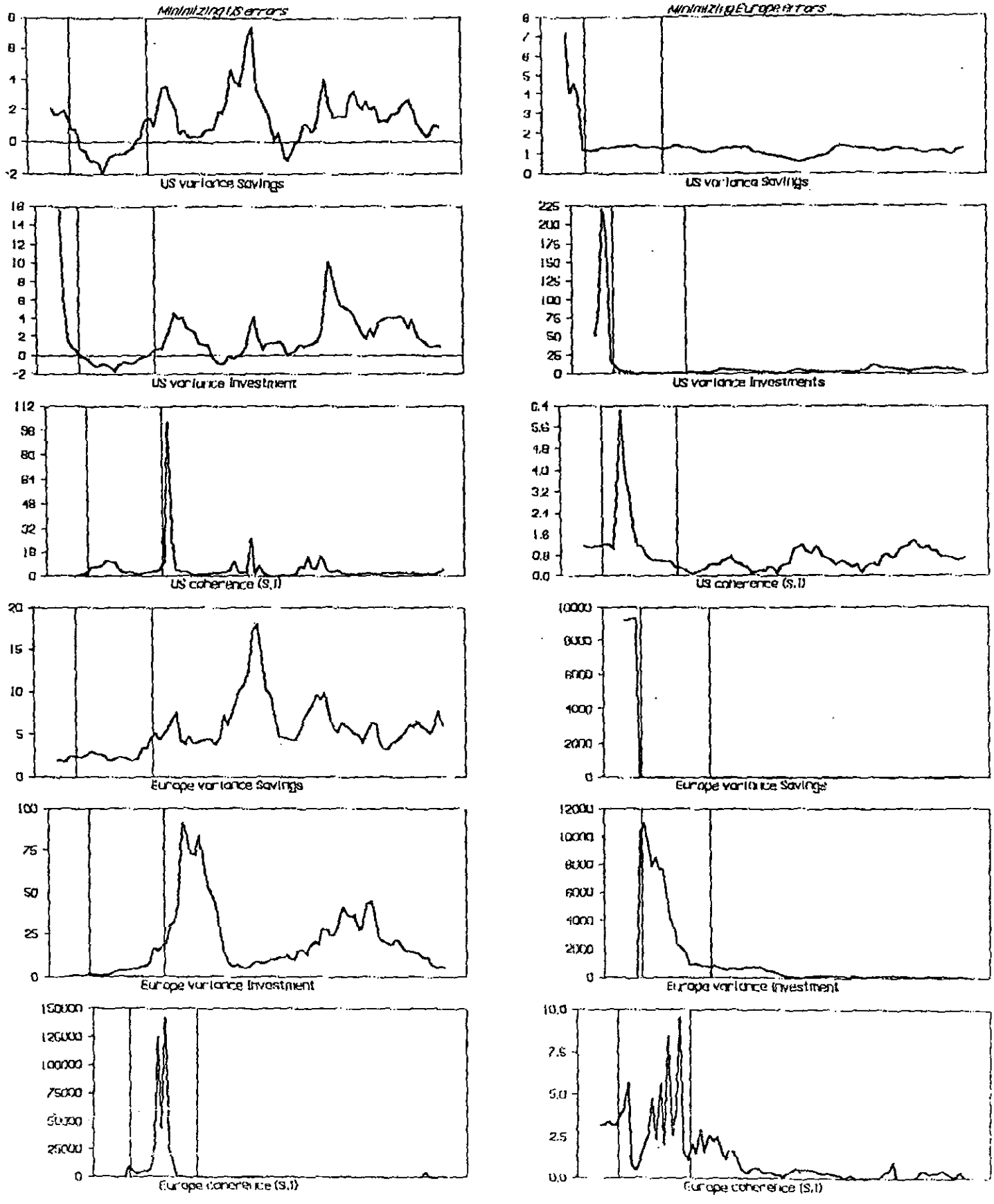


FIGURE 2: Watson's statistics.

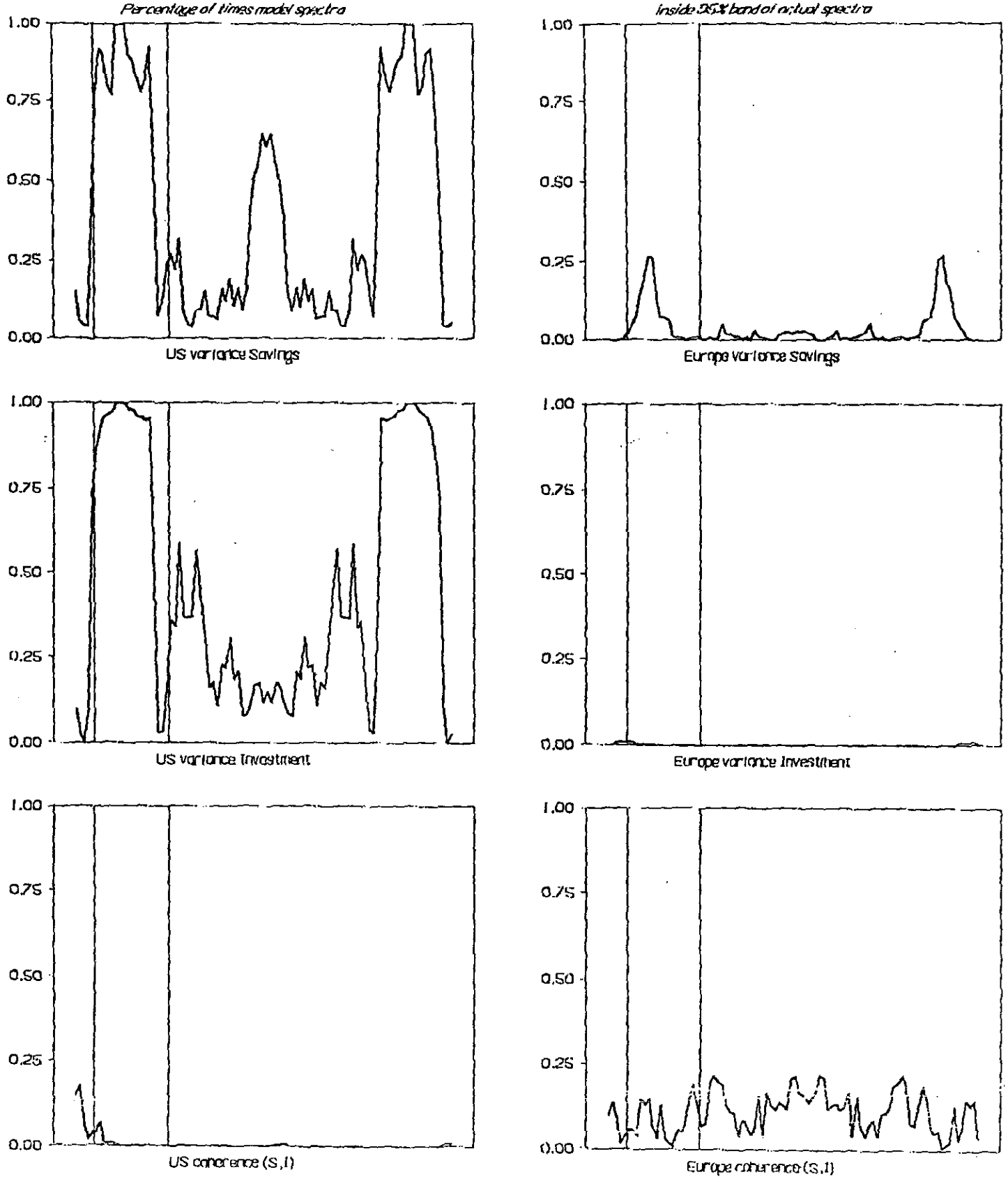


FIGURE 3: Fixed parameters.

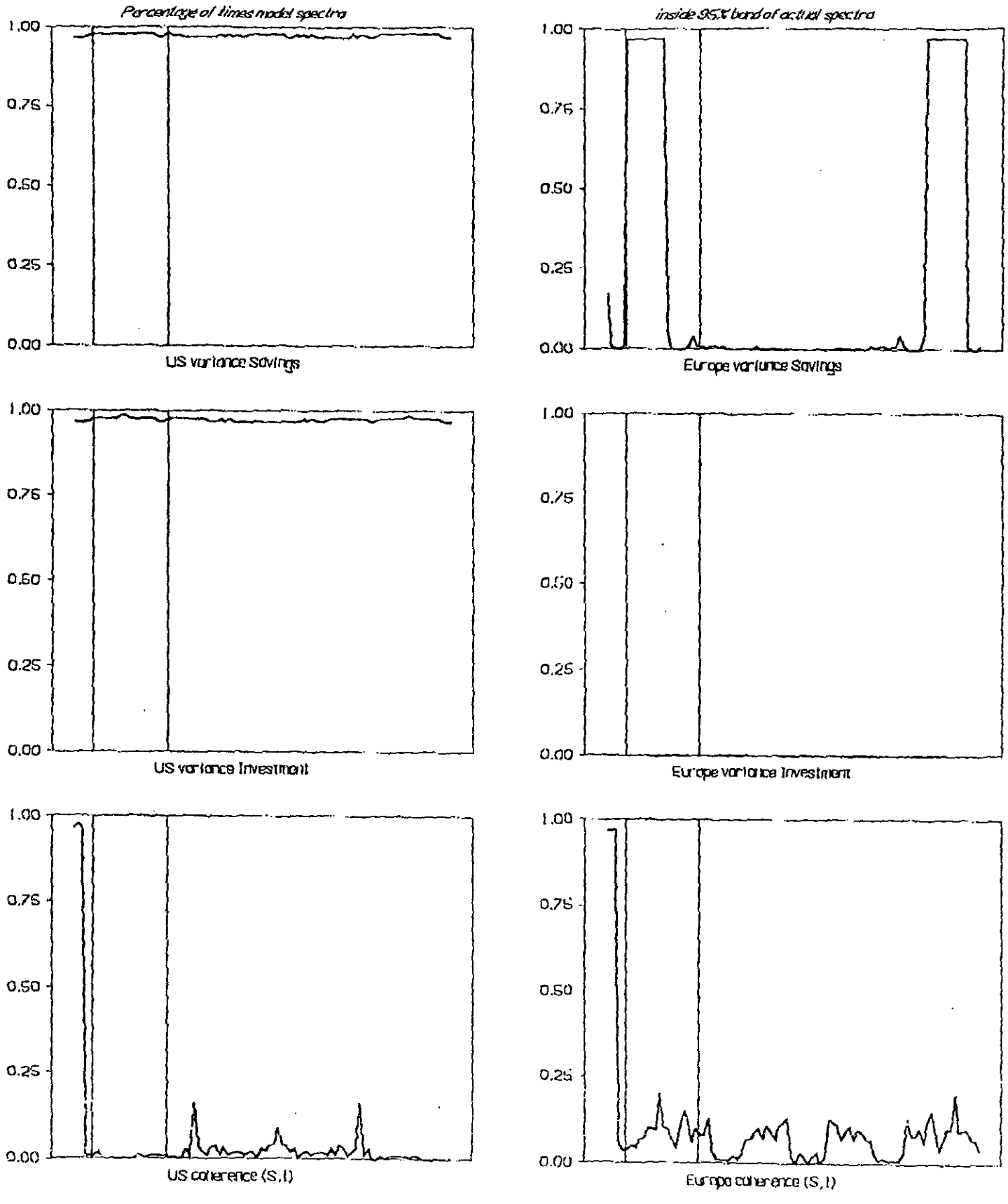


FIGURE 4: Random parameters.

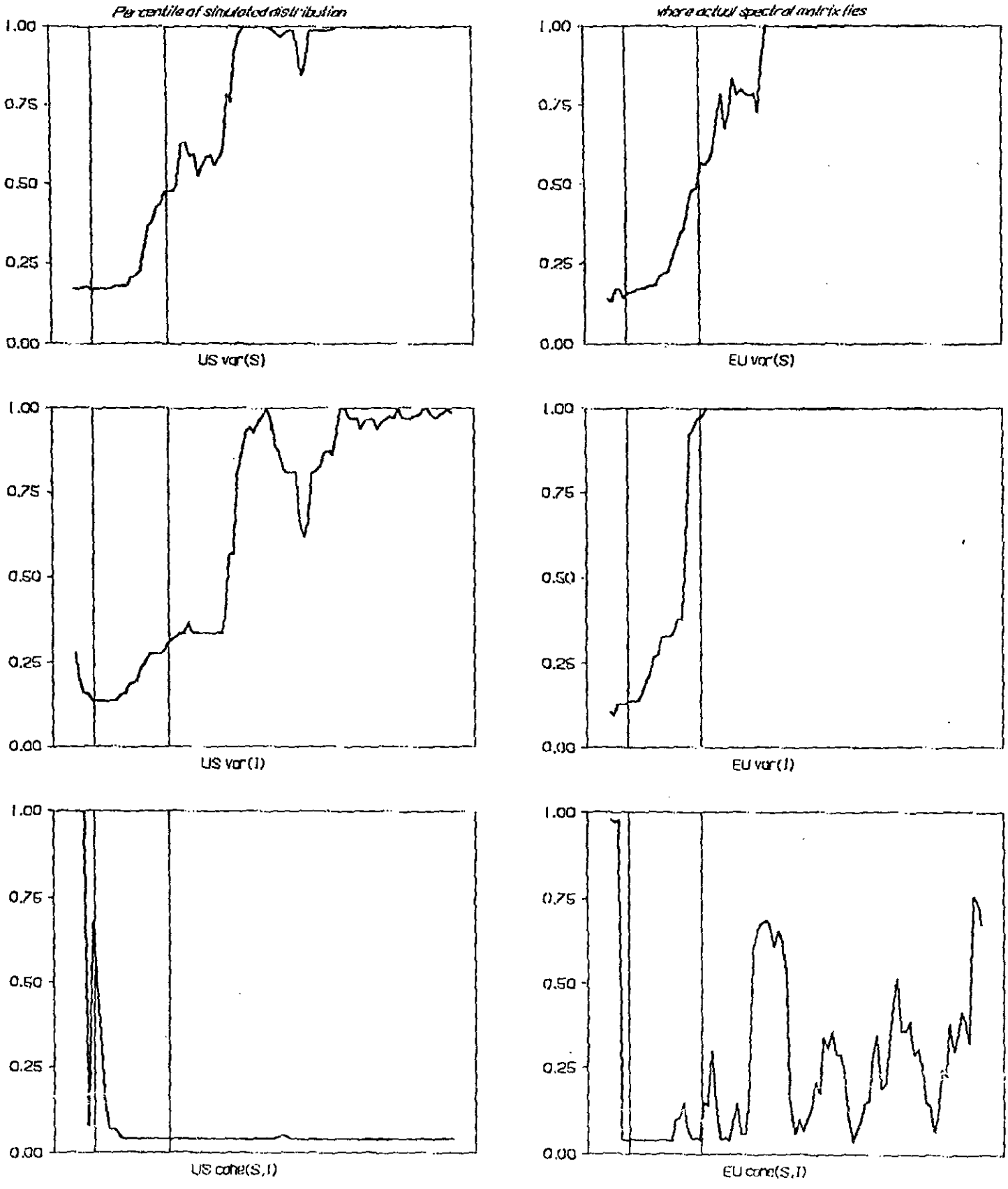


FIGURE 5: Fixed parameters.

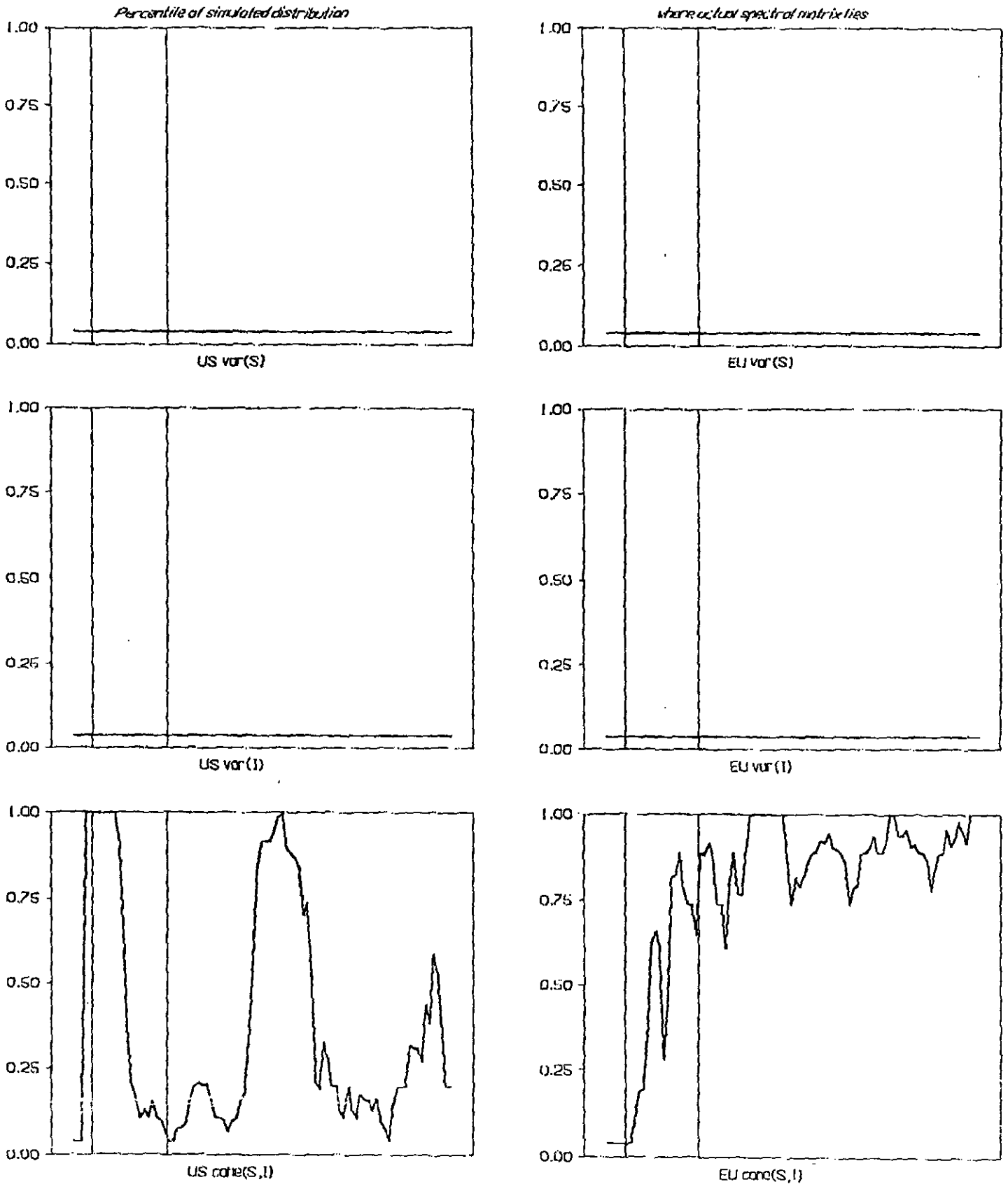


FIGURE 6: Random parameters.

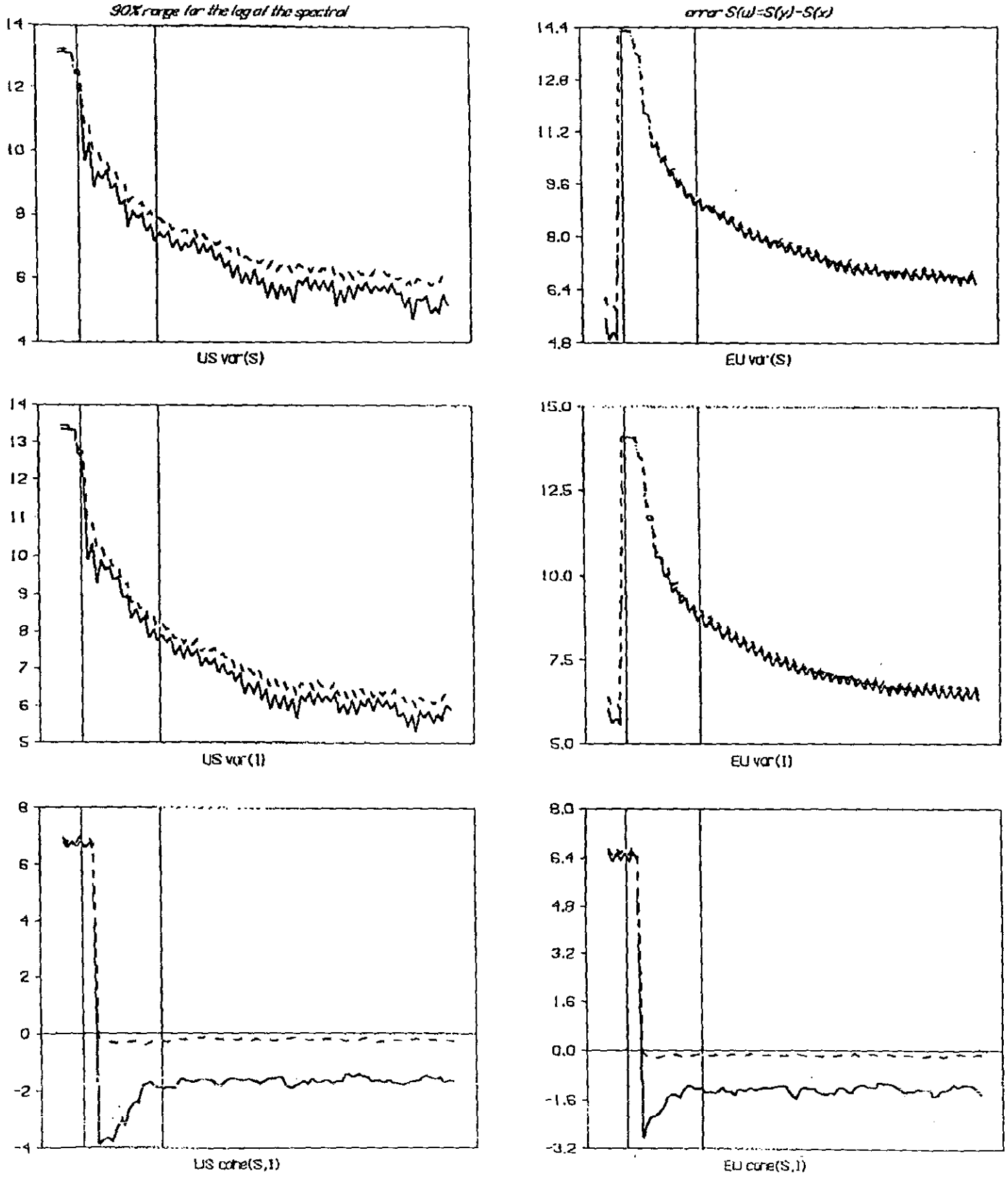


FIGURE 7: Random parameters.