Discussion Paper 4

Institute for Empirical Macroeconomics
Federal Reserve Bank of Minneapolis
250 Marquette Avenue
Minneapolis, Minnesota 55480

June 1988

UNDERSTANDING UNIT ROOTERS: A HELICOPTER TOUR

Christopher A. Sims* Harald Uhlig*

Institute for Empirical Macroeconomics
and University of Minnesota

ABSTRACT

For the first-order univariate autoregression without constant term, the joint p.d.f (corresponding to a flat prior) for the true coefficient ρ and the least squares estimate $\hat{\rho}$ is estimated by Monte Carlo and graphically displayed. The graphs show how the symmetric distribution of $\rho|\hat{\rho}$ coexists with the assymmetric distribution of $\hat{\rho}|\rho$. Treating tail areas of the $\hat{\rho}|\rho$ distribution as if they summarized evidence in the data about the location of ρ amounts to ignoring the shrinkage in the variance of $\hat{\rho}|\rho$ as ρ approaches one. Prior p.d.f.'s implicit in treating classical significance levels as if they were Bayesian conditional probabilities are calculated. They are shown to depend sensitively on $\hat{\rho}$ and to put substantial probability on ρ's above one.

JEL number 211

Keywords: unit roots, Bayesian econometrics, autoregression

*Research supported in part by NSF Grant SES 8608078.

This material is based on work supported by the National Science Foundation under Grant No. SES-8722451. The Government has certain rights to this material.

Understanding Unit Rooters: A Helicopter Tour

In an earlier paper (Sims [1988]) one of the authors of this paper pointed out that -- in contrast to the usual situation in econometric inference -- in time series models with unit roots Bayesian probability statements about the unknown parameters conditional on the data are different even asymptotically from classical confidence statements about the probability of random intervals covering the true value of the parameter. That paper argued that the Bayesian probability statements are more useful as well as much easier to handle analytically than the classical confidence statements.

Many economists are not used to having to make careful distinctions between probability statements about the location of unknown parameters (Bayesian inference) and probability statements about the behavior of statistics computed from the data in repeated samples (classical confidence statements). The earlier paper included an example which aimed at guiding intuition about these distinctions, but the example used discrete data and had no evident connection to the unit root time series context. This paper explores in more detail the distinction between confidence statements and probability statements about parameters, in a simple time series model which may show a unit root.

We first graphically summarize the results of a Monte Carlo study of the joint p.d.f. of an unknown autoregressive coefficient ρ and its least squares estimate $\hat{\rho}$, when ρ is treated as uniformly distributed. Bayesian conditional p.d.f.'s for ρ are cross sections of this joint p.d.f. along a fixed-$\hat{\rho}$ line, while classical distributions for $\hat{\rho}$ are sections of the joint p.d.f. along a fixed-ρ line. We display several views of the joint p.d.f., sliced in various ways (the helicopter tour of the title).

Careful econometrics based on classical procedures often considers both the distribution of the data conditional on a parameter value near the point estimate and the distribution conditional on some special point or subspace in which there is a unit root. The relative plausibility of the special subspace and the region around the point estimate is then judged based on p-values for the two null hypotheses given the observed data.[1]

This sort of procedure, while better than asymmetric treatment of the unit root hypothesis and its alternatives, is flawed because it relies on classical p values as measures of the relative plausibilities of the competing hypotheses. It can happen, and does happen in models with unit roots, that p values are quite misleading as measures of relative plausibility. We consider the following way of proceeding: After the sample has been observed, for each possible true value of ρ a test statistic for the null hypothesis that this ρ is the true one is formed, and its p value tabulated. Assuming that these p values have a single peak at $\rho=\hat{\rho}$, they are treated as if they trace out a c.d.f. for ρ, with $P[\rho_0<\rho]$ given by the p value for the test of the ρ null hypothesis when $\rho<\hat{\rho}$ and by one minus the p value when $\rho>\hat{\rho}$. This is of course not formally justified by either a Bayesian or a classical argument; it represents an attempt to capture formally what one is doing when one uses p values as if they imply probabilities for sets of ρ's given the observed data.

Of course by varying prior p.d.f.'s, a Bayesian can emerge with varying posterior p.d.f.'s. For a given sample and observed $\hat{\rho}$, the pseudo-c.d.f. we are constructing here can be rationalized as consistent with Bayesian inference based on some prior. The reason the procedure nonetheless has no Bayesian rationale is that its Bayesian interpretation is different for different $\hat{\rho}$'s. That is, there is no way to specify <u>one</u> Bayesian prior which leads to this p-value c.d.f. as a posterior for every $\hat{\rho}$. Nonetheless it is interesting to compute the implied prior p.d.f. for various $\hat{\rho}$'s, and we do so. The results show that when $\hat{\rho}$ is near one, the p-value c.d.f. implicitly

---

[1]Christiano and Ljungqvist [1988] provide an example of careful work along this line.

puts increasingly high prior weight on larger ρ's, even for ρ's above one.

1. The Model

We consider the simple univariate autoregressive model

$$y(t) = \rho y(t-1) + \varepsilon(t) \ , \tag{1}$$

with i.i.d. $\varepsilon(t) \sim N(0,\sigma^2)$. If we observe $y(t)$, $t=0,\ldots,T$, we can form the least squares estimate $\hat{\rho}$ of ρ. In this model $\sqrt{T}(\hat{\rho}-\rho)$ is asymptotically normal if, say ε is i.i.d. with bounded variance and $|\rho|<1$. When $\rho=1$, $\hat{\rho}$ is not asymptotically normal. The likelihood, conditional on the initial observation $y(0)$, is Gaussian in shape as a function of ρ, however, and this result of course does not depend on whether the data is actually generated by a process with a unit root or not.

Because the likelihood depends on both $\hat{\rho}$ and

$$\hat{\sigma}^2_\rho = \frac{\sum_{t=1}^{T} (y(t)-\hat{\rho}y(t-1))^2}{T \sum_{t=1}^{T} y(t-1)^2} \ , \tag{2}$$

there is no one-dimensional way to summarize the sample evidence. In order to develop insight into the relation between Bayesian and classical inference, however, it is helpful to artificially simplify the situation further. We will consider the situation where one cannot observe the full sample -- only $\hat{\rho}$. We will also assume that $\sigma^2=1$ and is known.

These simplifying assumptions make the shape of the likelihood nonnormal and difficult to derive. Their appeal is only that they make the Bayesian analytical framework consist of a two-dimensional joint p.d.f., that of ρ

and $\hat{\rho}$. A function of two arguments is easily visualized as a surface in three dimensions, while a function of three arguments is much harder to visualize.

We can be sure in advance that the likelihood will remain symmetric in ρ around a peak at $\hat{\rho}$, because conditional on $\hat{\sigma}$ it has these properties and it therefore will not lose them when $\hat{\sigma}$ is integrated out.

In the next section we will proceed to construct, by Monte Carlo, an estimated joint p.d.f. for ρ and $\hat{\rho}$ under a uniform prior p.d.f. on ρ. We choose 59 values of ρ, from .815 to 1.105 at intervals of .005. We draw 10000 100×1 i.i.d. N(0,1) vectors of random variables to use as realizations of ε. For each of the 10000 ε vectors and for each of the 59 ρ values, we construct a y vector with y(0)=0, y(t) generated by equation (1). For each of these y vectors, we construct $\hat{\rho}$. Using as bins the intervals $[-\infty,.815)$, $[.815,.820)$, $[.820,.825)$, etc. we construct a histogram which estimates the p.d.f. of $\hat{\rho}$ for each fixed ρ value. When these histograms are lined up side by side, they form a surface which is the joint p.d.f. for ρ and $\hat{\rho}$ under a flat prior on $\hat{\rho}$.

2. The Helicopter Tour

Figure 1-5 display different views of the same surface, the estimated joint p.d.f. for ρ and $\hat{\rho}$. Figure 1 shows the surface sliced along the $\hat{\rho}=1$ and $\rho=1$ planes. This angle gives a good view of the surface shape, but the view from lower down, centered on the corner of the viewing box, shown in Figure 2, makes it easier to be convinced that the distribution of $\hat{\rho}|\rho=1$, one side of which is the section generated by the left-hand panel in Figure 2, really is more skewed toward lower values than the conditional distribution of $\rho|\hat{\rho}=1$, one side of which is the section generated by the right-hand panel in Figure 2.

Figure 3 is sliced only along the $\rho=1$ plane, so the section is just the p.d.f. of $\hat{\rho}|\rho=1$. Here the well known result that $\hat{\rho}$ is asymmetric, with a peak at 1 but much more probability below than above one, is easily visible. The section along the $\hat{\rho}=1$ plane shown in figure 4 confirms the theoretical result that this p.d.f. is symmetric about $\rho=1$. Figure 5 shows that the distribution remains symmetric along the $\hat{\rho}=.95$ plane, though it is more dispersed. This result, that the ρ distributions spread out as $\hat{\rho}$ get smaller, is what generates the skewness when the joint p.d.f. is sliced in the other direction. The two sections shown in Figures 3 and 4 are displayed on top of each other in a two-dimensional graph in Figure 6, with both normalized to have the same integral. In Figure 7 they are displayed normalized to have the same height at their peaks, so the contrast between the symmetry of the conditional distribution of ρ and the asymmetry of the conditional distribution of $\hat{\rho}$ is sharper.

If we were studying many instances of the model (1), with true values of ρ drawn at random from a distribution which was uniform over (.84,1.06) (and possibly nonuniform, but not too wildly behaved outside that interval), then any reasonable person would have to agree that what the data implies about the likely location of ρ given that we observe $\hat{\rho}=1$ is given by taking the dotted line in Figure 6 as a p.d.f. for the unknown ρ. The difference between Bayesian and classical statistics is not over the logic of Bayes' rule, but over whether it can legitimately be applied when there is no "objective" source of randomness on which to base the notion of a probability distribution for ρ.

So let us suppose that we really have an application where, say, someone is generating ρ's uniformly by flipping coins or drawing numbers out of a hat. Everyone should agree that, on observing $\hat{\rho}=1$, our uncertainty about ρ is symmetric about $\rho=1$. What if we nonetheless try comparing the p-values of the null hypotheses $\rho=.98$ $\rho=1.02$ by classical procedures? The natural classical test of $\rho=.98$, assuming we can see only $\hat{\rho}$ and not the whole sample, is obtained by normalizing the $\rho=.98$ section of our $\rho,\hat{\rho}$ p.d.f. to integrate to one, then computing the area under the curve to the right of the observed $\hat{\rho}$. This area is the p-value, and one would reject $\rho=.98$ if it fell below some critical level, say $\alpha=.05$. Our Monte Carlo joint p.d.f. implies that the p-value for $\rho=.98$ given an observed $\hat{\rho}=1$ is .033, while the p-value for $\rho=1.02$ given $\hat{\rho}=1$ is .245.[2] We can reject .98 at the .05 level, in other words, while easily accepting 1.02. The actual conditional probability of $\rho>1.02$ given observed $\hat{\rho}=1$ is .097, which is the same as the conditional probability of $\rho<.98$ given $\hat{\rho}=1$.[3]

How can this be, given that we are already sure that any reasonable person must agree that our beliefs about ρ are symmetrically distributed about $\rho=1$?

---

[2] Our bins for $\hat{\rho}$ are centered in between the grid points of our true ρ values, with no bin centered at 1.0. Therefore the significance levels presented here were computed by interpolation.

[3] Again, this computed probability involves some interpolation, and it averages the two separate probabilities from the Monte Carlo study for $\rho<.98$ and $\rho>1.02$. The two separate computations are .106 for $\rho<.98$ and .125 for $\rho>1.02$; the difference is in line with the theoretical standard error of the Monte Carlo calculations.

The answer is that the p-values are distorted by some irrelevant information. It is indeed about equally likely that an observed $\hat{\rho}=1$ is generated by a true $\rho=1.02$ or a true $\rho=.98$. However $\hat{\rho}$'s much below 1 are much more likely given $\rho=1.02$ than are $\hat{\rho}$'s much above 1 given $\rho=.98$. In this particular sample we have observed $\hat{\rho}=1$, not $\hat{\rho}$ much above one or $\hat{\rho}$ much below one; for deciding what this sample tells us about ρ the implications of the competing hypotheses about $\hat{\rho}$'s we have not observed are irrelevant.

A similar result holds for an observed $\hat{\rho}=.95$. Given this observation the probability that $\rho \geq 1$ is .07, which is the same as the probability that $\rho \leq .9$. A test of the null hypothesis $\rho=1$ constructed as described in the preceding paragraph yields a p-value of .15 for $\hat{\rho}=.95$, while a test of the null hypothesis $\rho=.90$ yields a p-value of .031. While the asymmetry is not as dramatic here, clearly classical hypothesis testing, naively applied, would still lead to the mistaken view that when $\hat{\rho}=.95$ is observed, it is much more likely that the true ρ is 1 or higher than that the true ρ is .9 or lower. Much recent econometric work on models with possible unit roots could be interpreted as taking just this naive approach to interpreting classical hypothesis tests.

3. Implicit Priors

In the standard normal linear regression model, and asymptotically in most econometric applications, Bayesian probability statements about the location of ρ approximately coincide with corresponding p-values. Econometricians could therefore easily form the habit of treating an observation that $\hat{\rho}=.95$, which has a p-value of .15 on the null hypothesis $\rho=1$, as suggesting that the probability of $\rho \geq 1$ is about .15 given an observation of $\hat{\rho}=.95$. This amounts to summing the joint $\rho, \hat{\rho}$ p.d.f. along each constant-ρ line to form a family of c.d.f.'s, but then treating the values of these c.d.f.'s along a constant-$\hat{\rho}$ line as if they formed a c.d.f. for ρ conditional on the observed $\hat{\rho}$. As we have already seen, for $\hat{\rho}$ near one the resulting c.d.f. puts much more weight on $\rho>1$ than $\rho<1$, even though a flat prior would imply a conditional c.d.f. symmetric about $\rho=1$. It may be of interest to see what prior is implicit in inference based on treating p-values as generating a c.d.f. and how the implied prior shifts as the observed $\hat{\rho}$ changes.

Letting $f(\rho,\hat{\rho})$ be the joint p.d.f. of ρ and $\hat{\rho}$, the pseudo-p.d.f. for ρ we are considering is

$$g(\rho|\hat{\rho}) = \frac{\partial}{\partial\rho}\int_{-\infty}^{\hat{\rho}} f(\rho,s)\ ds \ . \qquad (3)$$

The actual conditional p.d.f. for $\rho|\hat{\rho}$ based on a flat prior over ρ is proportional to $f(\rho,\hat{\rho})$. For g to emerge as the conditional p.d.f. for $\rho|\hat{\rho}$, therefore, requires that the prior p.d.f. on ρ be proportional to $g(\rho|\hat{\rho})/f(\rho,\hat{\rho})$. We make an approximate calculation of this implied prior p.d.f. by cumulating our Monte Carlo estimate of f along constant-ρ planes, then differencing the result along constant-$\hat{\rho}$ planes, finally dividing by the original estimated f. For unlikely values of $(\rho,\hat{\rho})$, these estimates are ratios of small numbers with high proportional standard errors. Thus in the tails the estimates are quite erratic.

The results are displayed in Figures 8 and 9. Figure 8 shows the full range .8 to 1.1, while Figure 9 cuts off the section above 1.0, where very high p.d.f. values occur. One can see that the p.d.f. shifts increasing weight toward the region above $\rho=1$ as $\hat{\rho}$ gets closer to 1. For $\hat{\rho}=.95$ the implicit prior makes ρ's around 1 2 to 3 times more likely than ρ's araound .9. Furthermore, the prior p.d.f.'s for all $\hat{\rho}$ values keep increasing in the region above $\rho=1$ for as far as the estimates retain any reliability. Thus naive use of classical tests' p-values not only gives special prior weight to $\rho=1$, it implies a priori belief that a ρ of 1.05 is more likely than a ρ of .95.

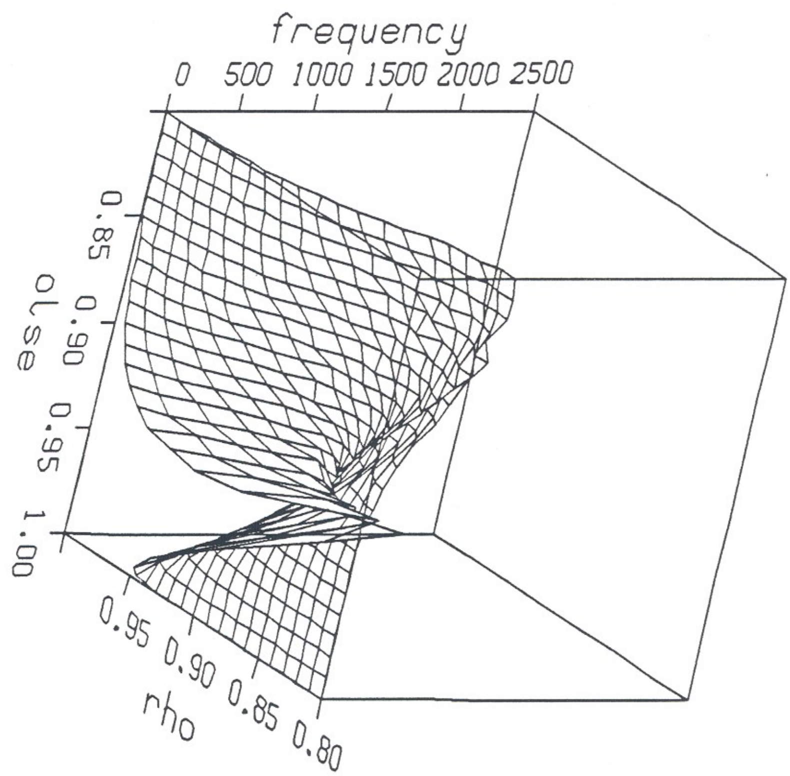
4. Conclusions

Use of classical statistical tests as measures of the plausibility of hypotheses is logically unsound. We have shown that in the case of a simple time series model with a unit root, it amounts to acting as if one had a stronger prior belief in a root at or above one, the closer to one is the estimated value $\hat{\rho}$ of the root. What the data tell us about the parameter is summarized in the likelihood, which can be conveniently described by normalizing it to integrate to one and treating the result as a p.d.f., i.e. by summarizing the implications of a flat-prior Bayesian analysis.

REFERENCES

Sims, Christopher A. [1988]. "Bayesian Skepticism on Unit Root Econometrics," Institute for Empirical Macroeconomics, Minneapolis, Discussion Paper No. 3.

Figure 1



frequency
0
500
1000
1500
2000
2500
olse
0.85
0.90
0.95
1.00
rho
0.95
0.90
0.85
0.80

Figure 2

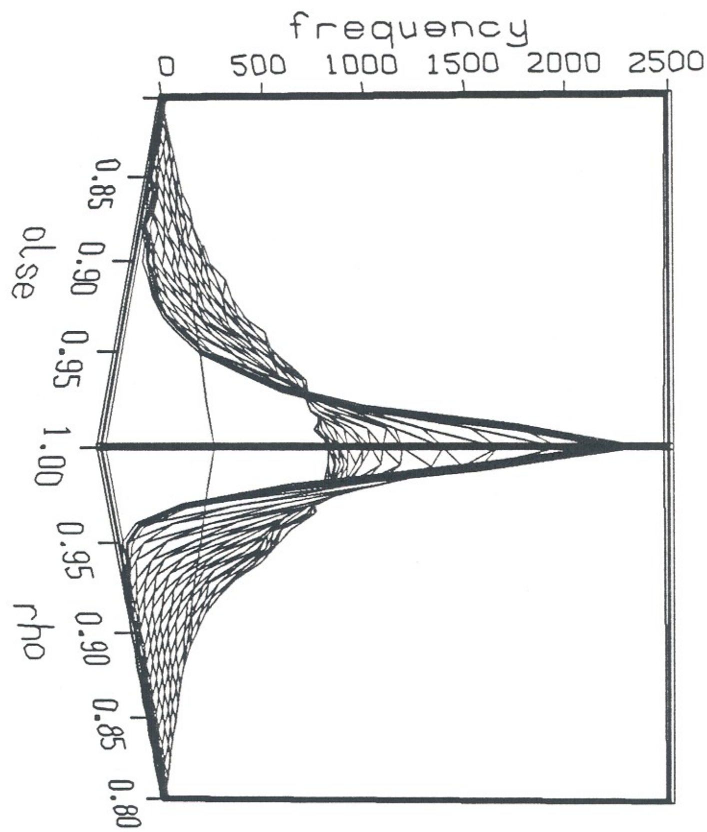

frequency
0
500
1000
1500
2000
2500
olse
0.85
0.90
0.95
1.00
0.95
0.90
0.85
0.80
rho

Figure 3

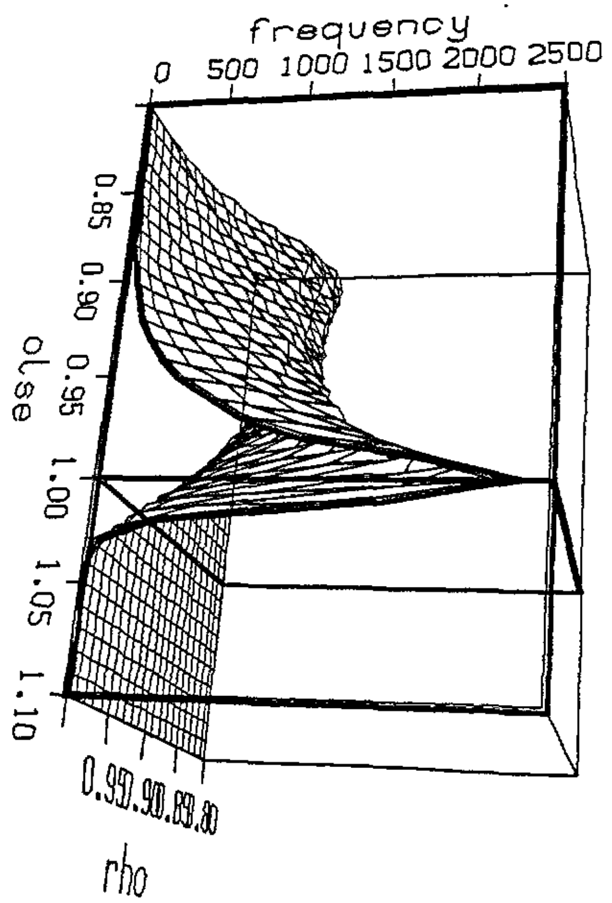
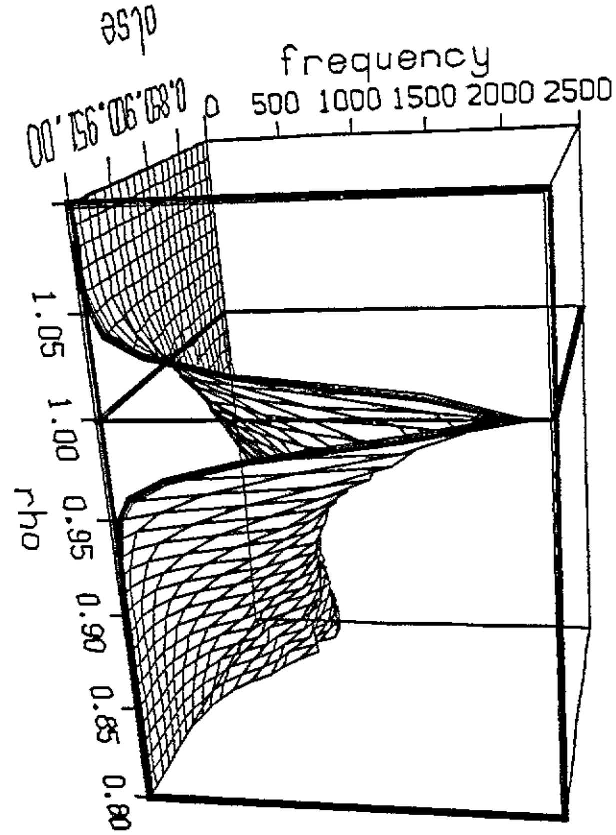

frequency
0
500
1000
1500
2000
2500
olse
0.85
0.90
0.95
1.00
1.05
1.10
rho

Figure 4



frequency
0 500 1000 1500 2000 2500
rho
0.80 0.85 0.90 0.95 1.00 1.05
olse

Figure 5

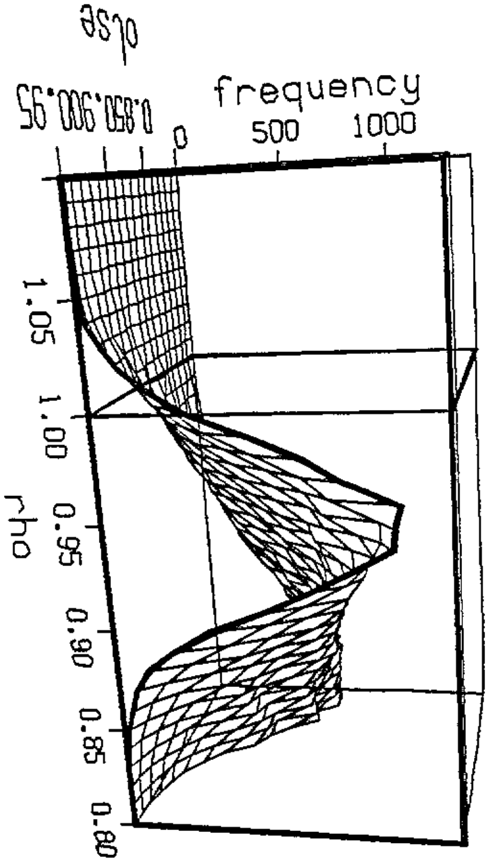


Figure 6

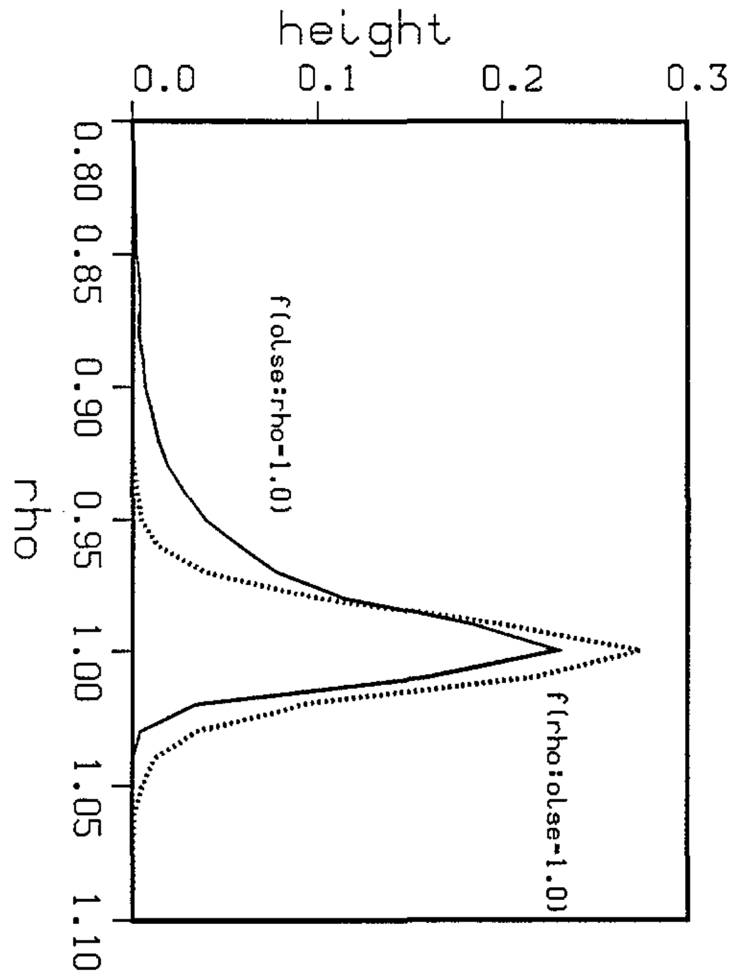

height
0.0
0.1
0.2
0.3
rho
0.80
0.85
0.90
0.95
1.00
1.05
1.10
f(olse:rho=1.0)
f(rho:olse=1.0)

Figure 7

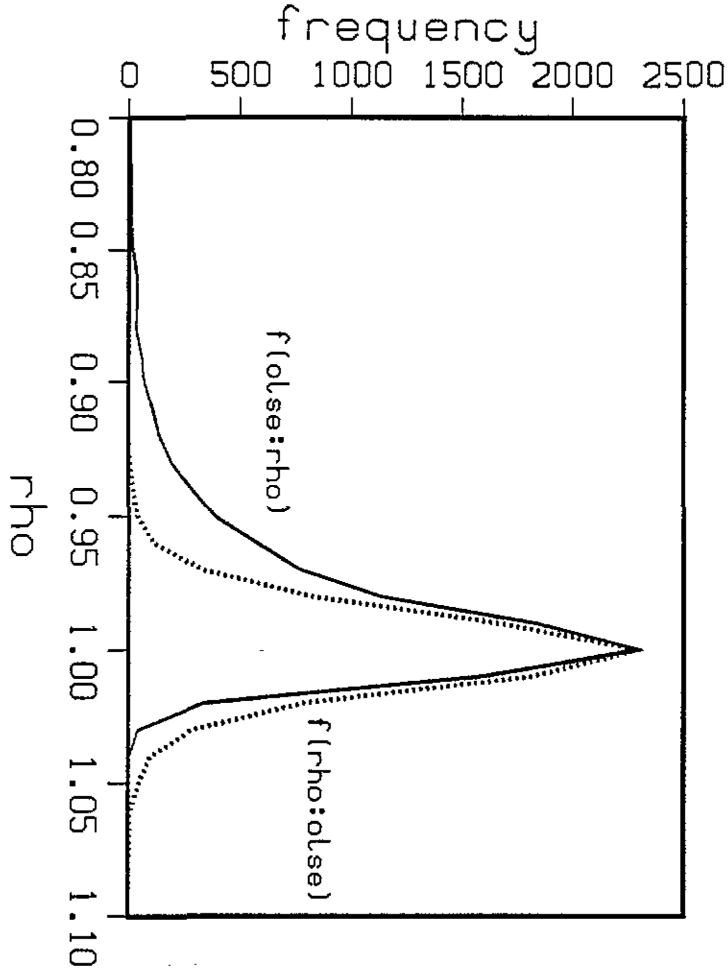

frequency
0
500
1000
1500
2000
2500
rho
0.80
0.85
0.90
0.95
1.00
1.05
1.10
f(olse:rho)
f(rho:olse)

Figure 8

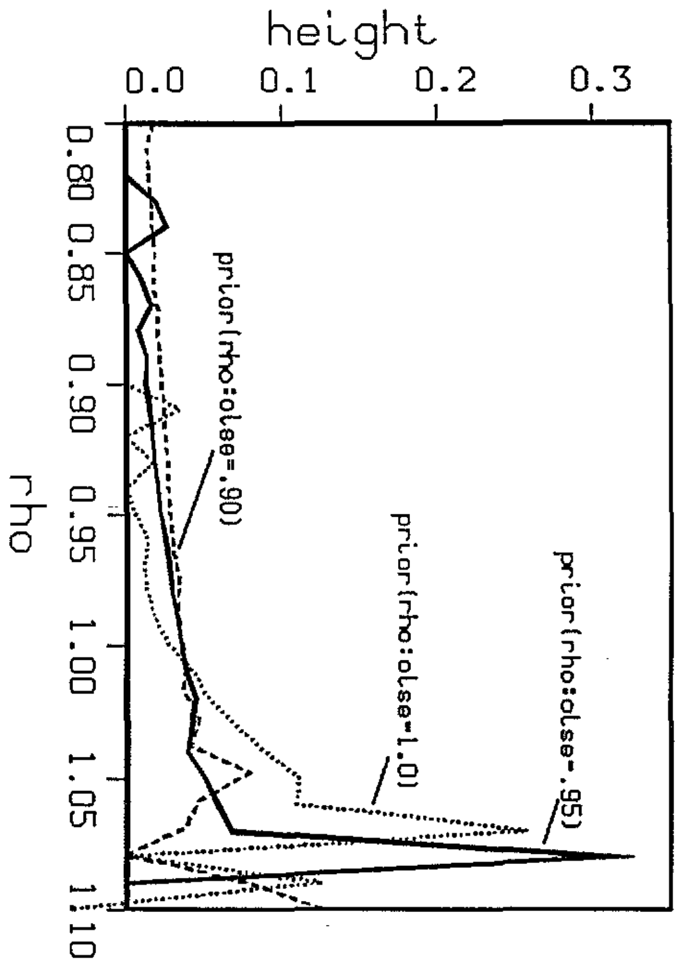
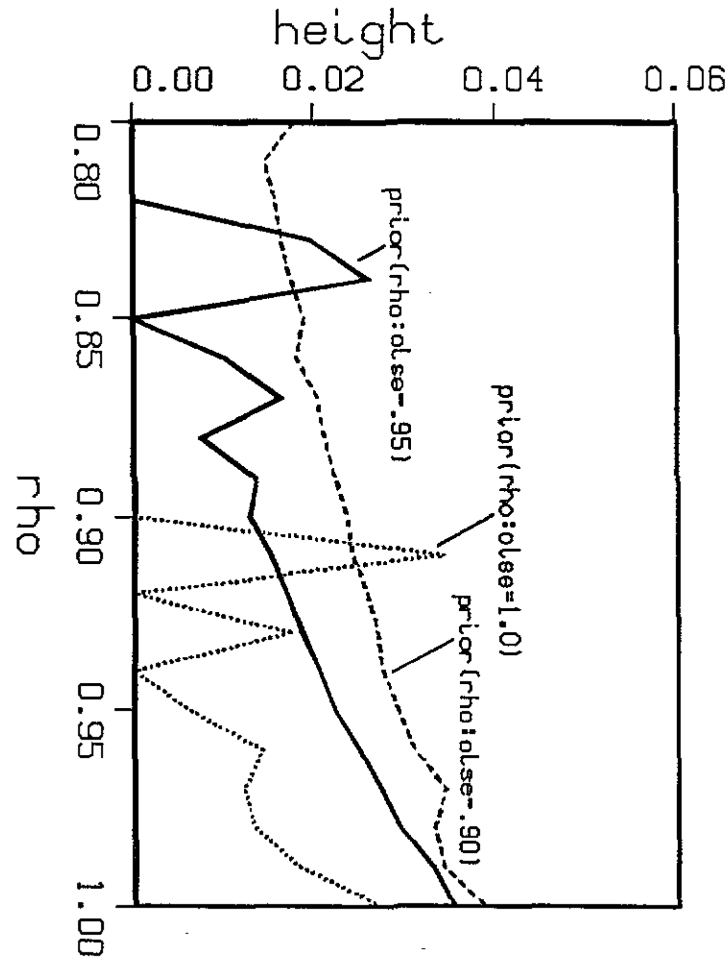

height
0.0
0.1
0.2
0.3
rho
0.80
0.85
0.90
0.95
1.00
1.05
1.10
prior(rho:olse=.90)
prior(rho:olse=1.0)
prior(rho:olse=.95)

Figure 9


height
0.00
0.02
0.04
0.06
rho
0.80
0.85
0.90
0.95
1.00
prior(rho:olse=.95)
prior(rho:olse=1.0)
prior(rho:olse=.90)