Decision maps for bivariate time series with potential threshold cointegration

Robert M. Kunst
University of Vienna,
Department of Economics,
BWZ, Bruenner Strasse 72,
A–1210 Wien (Vienna), Austria
Abstract

Bivariate time series data often show strong relationships between the two components, while both individual variables can be approximated by random walks in the short run and are obviously bounded in the long run. Three model classes are considered for a time-series model selection problem: stable vector autoregressions, cointegrated models, and globally stable threshold models. It is demonstrated how simulated decision maps help in classifying observed time series. The maps process the joint evidence of two test statistics: a canonical root and an LR-type specification statistic for threshold effects.

Keywords: Bayes test, model selection, nonlinear time series, cointegration, simulation.
1 Introduction

Bivariate time series data often exhibit features such as the pair of retail interest rates in Figure 1. The two component series appear to be closely linked such that deviations from each other are relatively small. For long time spans, the variables move up or down the positive diagonal, with very little memory with respect to this motion. Individually, the series appear to be well described as random walks or at least as first-order integrated processes with low autocorrelation in the differences. Eventually, however, the upward or downward motion appears to hit upon some outer boundary and is reversed, such that all observations are contained in a bounded interval, with unknown bounds. A good example for such pairs of time series data are interest rates, such as saving and loan rates or bill and bond rates, though a deep economic analysis of interest rates is outside the scope of this paper. It can be observed that they attain their maximum in phases of high inflation and that their minimum is usually strictly positive. Similar pairs of time series may appear in non-economic data, such as water flow data from two rivers (see Tsay, 1998). Such series are characterized by three features: individual behavior close to but not exactly equal to random walks for considerable time spans; parallel movement in both variables, though no exact relation and possibly slight deviations of the ‘equilibrium’ from the positive diagonal; in the long run, ‘mean reversion’ determined by unknown upper and lower bounds.

This paper is concerned with selecting an appropriate time series model for such data, if the set of available classes is given by the three following ideas. Firstly, the observed reversion to some distributional center may suggest a linear stable vector autoregression. This model class has the drawback that the autoregression is unable to reflect the random movement in the short run. Secondly, this movement and the obvious link between the components may suggest a cointegrated vector autoregression. Particularly for interest rates of different maturity, this idea has been popular in the econometric literature, see for example Campbell and Shiller (1987), Hall et al. (1992), or Johansen and
Figure 1: Time-series scatter plot of monthly data on U.S. interest rates on loans and 7-days deposits, 1963–2002.
Juselius (1992). Enders and Siklos (2001) state that “it is generally agreed that interest-rate series are I(1) variables that should be cointegrated”. The drawback of this model class is that it does not match the long-run bounded-ness condition. The model contains a unit root, is non-ergodic and inappropriate from a longer-run perspective (see Weidmann, 1999, for a similar argument in economics). Thirdly, one may consider a mixture of the two models, with cointegration prevailing in a ‘normal’ regime and global mean reversion in an ‘outer’ regime. This idea yields a threshold cointegration specification, as used by Jumah and Kunst (2002), for interest rates on loans and deposits. The drawback of the model is that it is nonlinear and contains some poorly identified parameters.

The concept of threshold cointegration was developed by Balke and Fomby (1997, henceforth BF) who assumed a version with cointegration in the outer regime and an integrated process without cointegration in an inner regime (a ‘band’). Jumah and Kunst (2002) suggested a modification of BF’s model that is in focus here. Threshold cointegration models were also considered by Enders and Granger (1998), Enders and Falk (1998), and Enders and Siklos (2001). These contributions focus on asymmetric adjustment to disequilibrium and hence they mainly use two-regime or single-threshold models. Hansen and Seo (2002) analyze hypothesis testing for single-threshold cointegration models and assume cointegration in both regimes, though with different cointegrating vectors. Like BF, Lo and Zivot (2001) consider the case of three regimes, with cointegration in the lower and upper regimes and no cointegration in the central regime. While these authors allow for asymmetric reaction or different cointegrating structures across regimes, only symmetric reaction outside the band will be considered here, due to the limited information that is provided by the few observations in the outer region of our models. Weidmann (1999) analyzed bivariate time series of an inflation index and an interest rate and suggested a three-regimes model that is close to the one used here. While his model assumes different cointegration structures across regimes and achieves global stability by the choice of cointegrating vectors in the outer regimes, we

Threshold cointegration models are particular threshold vector autoregressions of the SETAR (self-exciting threshold autoregression) type that was considered by Chan et al. (1985), Tong (1990), Chan (1993), and Chan and Tsay (1998). Particularly Chan et al. (1985) found that stability in the outer regimes is sufficient, though not necessary, for global stability and geometric ergodicity. It follows that the models suggested here are globally stable and geometrically ergodic. A proof of this important property has been added to this paper as an appendix.

The decision set consists of three elements: the stable vector autoregression, the linear cointegration model, and the threshold cointegration model. Each model class deserves attention as a possible data-generating mechanism. It is therefore interesting to study methods that allow a selection among the classes on the basis of observed data. Candidates for discriminatory statistics are likelihood ratios for any two model classes or approximations thereof. The theory of likelihood ratios for the first and the second class has been developed by Johansen (1988), hence it is convenient to include this ratio in the vector statistic. As another statistic, we add an approximation to the likelihood ratio of the first and the third class.

Model selection is a finite-action problem. It is obvious that a mere application of the standard binary Neyman-Pearson framework of null and alternative hypotheses severely restricts the set of decision procedures. A joint evaluation of test statistics allows a much richer set than consideration of test sequences only. Here, the three competing model classes are modeled as three alternative Bayes data measures. Each measure can be conditioned on the values of a vector statistic, such that the probability for each hypothesis can be evaluated conditional on the given or observed test statistic. The model or hypothesis with maximum probability is then the suggested choice for the observed value of the statistic. In
the space of null fractiles of the test statistics, one obtains ‘decision maps’ that show distinct regions of preference for each model class. This approach builds on a suggestion by Hatanaka (1996) and was used by Kunst and Reutter (2002), among others. For the present problem, non-informative prior distributions are elicited on the basis of Jordan distributions (see also Kunst, 2001). Discrete uniform priors over the three hypotheses are allotted implicitly. All calculations of decision maps have been conducted by means of simulation. Decision maps are particularly well suited for model selection based on the joint evaluation of two test statistics. Conditional on the priors for the rival hypotheses, the decision according to the map is optimal in the sense that any deviation from the suggested selection increases the probability of incorrect classification.

The remainder of this paper is structured as follows. Section 2 explains the decision maps approach and details the properties of the entertained models, including the elicitation of non-informative prior distributions. Section 3 reports the simulation results, including the maps and their tentative interpretation. Section 4 concludes.

2 Designing the simulation

2.1 Decision maps

The decision maps approach can be applied to all parameterized problems \((f_\theta, \theta \in \Theta)\) for finite-dimensional \(\Theta\), where a decision is searched among a finite number of rival hypotheses (a partition of \(\Theta\)), preferably on the basis of a bivariate vector statistic. For a univariate statistic, the maps collapse to intervals on the real line. For vector statistics with a higher dimension, the visual representation of the maps meets technical difficulties.

Assume a decision concerns the indexed set of hypotheses \(\{H_j = \{\theta \in \Theta_j\}, j = 1, \ldots, h\}\). Usually, model selection utilizes \(h - 1\) likelihood-ratio statistics \(S_{1(j), 2(j)}\) or approximations thereof, for ‘null’ hypothesis \(H_{1(j)}\) and alternative \(H_{2(j)}\), with \(1 \leq j < h, 1(j) \neq 2(j)\), and \(1(j), 2(j) \in \{1, \ldots, h\}\). We focus on
this typical case, though there may be problems where less than \( h - 1 \) statistics suffice for a discrimination or others where the evaluation of \( h \) statistics is convenient. For ease of notation, let the statistics be collected in an \( h - 1 \)-dimensional vector statistic \( \mathbf{S} = (S_1, \ldots, S_{h-1})' \), such that \( S_j \) and \( S_{1(j)2(j)} \) can be used equivalently. The classical approach requires nested hypotheses, such that \( \Theta_{i(j)} \subset \Theta_{z(j)} \), where bars denote topological closure. If parameter sets can be completely ordered, one may write \( \Theta_j \subset \Theta_{j+1} \) for \( 1 \leq j < h \). Then, two typical choices of vector likelihood-ratio statistics are \( \mathbf{S} = (S_1, S_2, \ldots, S_{h-1}, h)' \) and \( \mathbf{S} = (S_{1,h}, S_{2,h}, \ldots, S_{h-1,h})' \).

To an investigator who calculates \( \mathbf{S} \) from data, a natural question is which of the \( h \) classes should be selected for a given value of \( \mathbf{S} = s \). An autoral answer is to use the class that is most probable. The determination of such probabilities requires a Bayesian setup.

Let weighting prior distributions be defined on each \( \Theta_j \), \( 1 \leq j \leq h \) by their densities \( \pi_j \). For any pair \((1(j), 2(j))\), \( 1 \leq j < h \), the collection of distributions \( \{f_\theta, \theta \in \Theta_{1(j)}\} \) defines a null distribution \( f_{1(j)} \) of the statistic \( S_j \) via the implied p.d.f. of this statistic under \( \theta \), denoted as \( f_{j; \theta} \) by

\[
\int_{\Theta_j} f_{j; \theta}(x) \pi_j(\theta) d\theta = f_{1(j)}(x) .
\]

Note that it is equally possible to define a null distribution \( f_{2(j)} \) of the same statistic. Let \( F_{1(j)} \) denote the c.d.f. that corresponds to \( f_{1(j)} \). Then, the preference area for \( \Theta_j \) is defined as

\[
PA_j = \{ \mathbf{z} = (z_1, \ldots, z_{h-1})' \in (0,1)^{h-1} \mid j = \arg \max_j P(H_j|\mathbf{S}) , \quad z_k = F_{1(k)}(S_k) \}.
\]

The transformation \( F_{1(k)} \) conveniently represents the preference areas in a simplex instead of some possibly unbounded subspace of \( \mathbb{R}^{h-1} \). A graphical representation of preference areas is called a decision map.

The meaning of (2) can be highlighted by considering its counterpart in classical statistics. Suppose \( h = 3 \), and two statistics are evaluated. In a \((0,1)^2\)-diagram for the fractiles of the null distributions for \( S_{12} \) and \( S_{23} \), a
given sample of observations defines a point of realized statistics. A classical statistician would base a decision on rejections in a test sequence, for example starting from the more general decision on $S_{23}$. If one-sided tests are used that reject for the upper tails of their null distributions, the *classical preference area* for hypothesis $\Theta_3$ or $H_3$ is the rectangle $A_3 = (0, 1) \times (0.95, 1)$. If $\Theta_2$ is not rejected, test $S_{12}$ will be applied and separates the preference areas for $\Theta_1$, $A_1 = (0, 0.95) \times (0, 0.95)$, and for $\Theta_2$, $A_2 = (0.95, 1) \times (0, 0.95)$ . Classical statistics may face difficulties in uniquely determining null distributions, due to non-similarity of the tests. The $(0, 1)^2$-chart split into the three rectangles constitutes a *classical decision map*.

Bayesian decision maps are more complex than classical decision maps, as the boundaries among the preference areas may be general curves. Except for few simple decision problems, it is difficult to calculate the value of a conditional probability for a given point of the $(0, 1)^2$-square in the fractile space. It is more manageable to generate, by numerical simulation, a large number of statistics from the priors with uniform weights across the hypotheses and to collect the observed statistics in a bivariate grid over the fractile space. Within each ‘bin’, the maximum of the observed values can be evaluated easily. In computer time, the simulation can be time-consuming but requires little more storage than $h^2 g^2$, where $h$ is the number of considered hypotheses and $g$ is the inverse resolution of the grid, i.e., there are $g^2$ bins.

The numerical calculation of a decision map consists of sequential steps. At first, a conveniently large number of replications for each of the two statistics under their respective null distributions are generated. From the sorted simulated data, a grid of fractiles is calculated. Note that ‘coding’ by the null fractiles merely serves to obtain a convenient representation, as it implies a uniform distribution under the ‘null’ and some deviation from uniformity under the ‘alternative’, such that all bins in the $(0, 1)$-square will be sufficiently populated. The exact null distribution is not required. Different distributions may even be preferred, particularly if these have a closed analytical form.

In a second step, both statistics are generated from any of the competing
hypotheses, i.e., from the prior distributions $\pi_j$. These statistics are allotted into the bins as defined by the fractiles grid. Finally, each bin is marked as ‘belonging’ to that hypothesis from which it has collected the maximum number of entries.

The simulation of the null fractiles requires a number of replications that is large enough to ensure a useful precision of the fractiles. For the purpose of the graphical map, a high precision may not be required. A large number of replications in this preparatory step slows down the simulation considerably due to the sorting algorithm that is applied in order to determine empirical fractiles.

By contrast, a fairly large number of replications can be attained for the simulation of statistics conditional on the respective hypotheses. Computer time is limited by the calculation of the statistics only, which may take time if some iteration or non-linear estimation is required, while only a $g \times g$ matrix of bins for each hypothesis is stored during the simulation. For $g = 100$, $10^6$ replications yield acceptable maps. For $h$ hypotheses, this gives $10^6 \cdot h$ replications. It was found that kernel smoothing of the bin entries improves the visual impression of the map more than considerably increasing the number of replications (see Section 2.4).

### 2.2 The model hypotheses

For model selection, prior distributions with point mass on lower-dimensional parameter sets are required. Then, flat or Gaussian distributions are used on the higher-dimensional sets, with respect to a convenient parameterization. The specific requirements of decision map simulations rule out improper or Jeffreys priors, hence the priors do not coincide with the suggestions for unit-root test priors in the literature (see Bauwens et al., 1999). They are perhaps closest to the reference priors of Berger and Yang (1994), that peak close to the lower-dimensional sets and are relatively flat elsewhere.
The first model $H_1$ is the stable vector autoregression

$$
\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \mu + \sum_{j=1}^{p} \Phi_j \begin{pmatrix} x_{t-j} \\ y_{t-j} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}
$$

(3)

with all zeros of the polynomial $Q(z) = \det(I_2 - \sum_{j=1}^{p} \Phi_j z^j)$ outside the unit circle. For $p = 1$, this condition is equivalent to the condition that all latent values of $\Phi = \Phi_1$ have modulus less than one. This again is equivalent to the property that $\Phi$ has a Jordan decomposition

$$
\Phi = T J T^{-1}
$$

(4)

with non-singular transformation matrix $T$ and ‘small’ Jordan matrix $J$. If one restricts attention to real roots and to non-derogatory Jordan forms, the matrix $J$ is diagonal with both diagonal elements in the interval $(-1, 1)$. Therefore, the prior distribution for this model $\pi_1$ can be simply taken from the family of Jordan distributions and is defined by

$$
t_{12}, t_{21} \sim N(0, 1)\\
t_{11} = t_{22} = 1\\
j_{11}, j_{22} \sim U(-1, 1)\\
\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})' \sim N(0, I_2),
$$

(5)

where the notation $J = (j_{kt})$ etc. is used. The concept can be extended easily to the case $p > 1$ and to non-zero $\mu$. For the basic experiments, $p = 1$ and $\mu = 0$ is retained. Unless otherwise indicated, all draws are mutually independent. For example, $\varepsilon_t$ and $\varepsilon_s$ are independent for $s \neq t$ in all experiments, thus assuming strict white noise for error terms.

This prior distribution is not exhaustive on the space of admissible models. It excludes derogatory Jordan forms and complex roots. Derogatory Jordan forms are ‘rare’ in the sense that they occupy a lower-dimensional manifold. Complex roots are covered in an extension that was used for some experiments.

In this variant, 50% of the $\Phi$ matrices were drawn from $\Phi = TJcT^{-1}$ instead of
(4) and (5), with

\[
J_c = \begin{pmatrix}
  r \cos \phi & r \sin \phi \\
  -r \sin \phi & r \cos \phi
\end{pmatrix}.
\] (6)

It is known from matrix algebra that \(J_c\) is obtained from an original diagonal \(2 \times 2\) matrix \(J\) with conjugate complex elements that is transformed by a complex matrix

\[
T_c = \begin{pmatrix}
  1 + i & 1 - i \\
  1 - i & 1 + i
\end{pmatrix}
\] (7)

such that \(J_c = T_c J T_c^{-1}\). The prior distribution for \(J_c\) is constructed by drawing \(r\) from a \(U(0, 1)\) distribution and \(\phi\) from a \(U(0, \pi)\) distribution. The specification for the priors for \(T\) is unchanged. For these experiments with complex roots, the prior distribution for \(H_3\) was also modified accordingly.

In \(H_1\), both variables \(x\) and \(y\) have a finite expectation and revert to it geometrically. In geometric terms, the \((x, y)\)-plane has a unique equilibrium point \((\bar{x}, \bar{y})\), which in the case \(\mu = 0\) collapses to \((0, 0)\). While many researchers may rule out the stable model for series as presented in Figure 1 on grounds of intuition, such a decision is not so obvious when the variables are observed for extremely long time spans. For example, interest rates are observed to cluster in the single-digit region over a range of several centuries.

The second model class \(H_2\) consists of cointegrating vector autoregressions, with the cointegrating vector defined as the difference of the two variables \((1, -1)\). For interest rates of different maturity, this difference is the ‘yield spread’. For saving and loan rates, it is the mark-up of banks. For first-order vector autoregressions of this type, a prior distribution is obtained from the error-correction representation

\[
\begin{pmatrix}
  \Delta x_t \\
  \Delta y_t
\end{pmatrix} = \mu + \Pi \begin{pmatrix}
  x_{t-1} \\
  y_{t-1}
\end{pmatrix} + \begin{pmatrix}
  \varepsilon_{1t} \\
  \varepsilon_{2t}
\end{pmatrix}
\] (8)

with a matrix \(\Pi\) of rank one. The matrix \(\Pi\) can be represented in the form

\[
\Pi = \begin{pmatrix}
  \alpha_1 \\
  \alpha_2
\end{pmatrix}
\] (1, -1) .
\] (9)
The elements $\alpha_1$ and $\alpha_2$ are chosen in such a way that explosive modes in the system are avoided. Again, this condition is more readily imposed on the Jordan representation $\Pi = TJT^{-1}$ with diagonal matrix $J = \text{diag}(\lambda, 0)$ for $\lambda \in (-2, 0)$. The specification

$$T = \begin{pmatrix} 1 & 1 \\ a & 1 \end{pmatrix}$$

(10)

with $a \sim N(0, 1)$ covers a wide variety of admissible matrices $\Pi$. The implied form of $\Pi$ is

$$\Pi = \frac{1}{1-a} \begin{pmatrix} \lambda & -\lambda \\ a\lambda & -a\lambda \end{pmatrix}$$

(11)

and satisfies the general form (9). The class $H_2$ again assumes $\varepsilon_{jt} \sim N(0, 1)$ for the disturbances and, in the basic specification, $\mu = 0$. The thus defined prior $\pi_2$ is more difficult to generalize to higher $p$ than $\pi_1$.

The third model class $H_3$ are threshold cointegrating models. The basic form of these models is

$$\begin{pmatrix} \Delta x_t \\ \Delta y_t \end{pmatrix} = \mu + \begin{pmatrix} \alpha_{11} \\ \alpha_{21} \end{pmatrix} (1, -1) \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{12} \\ \alpha_{22} \end{pmatrix} (1, 0) \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} I\{|x_{t-1} - x^*| > \delta\} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

(12)

The symbol $I\{\cdot\}$ denotes the indicator function on the set $\{\cdot\}$. There are two cointegrating vectors. The first one, $(1, -1)$, is always active whereas the second one, $(1, 0)$, is only activated at ‘large’ values of the trigger variable, in our case $x_{t-1}$. An obvious variant is obtained by replacing the second vector by $(0, 1)$ and the trigger variable by $y_{t-1}$. For simplicity, we do not focus on the choice of the trigger variable, nor do we consider a variation of the trigger lag (‘delay’), as is common in the literature on threshold time series models. In empirical applications, apart from $x_{t-1}$ and $y_{t-1}$ and longer lags, also weighted averages $\lambda x_{t-1} + (1-\lambda) y_{t-1}$ may deserve consideration. While such modifications are minor and do not affect the main results of this paper, trigger variables that are
stationary under $H_2$, such as $\Delta x_{t-1}$ or $x_{t-1} - y_{t-1}$, change the properties of $H_3$ and cannot be treated in the presented framework.

The model is ergodic and both variables $x$ and $y$ have finite expectation (see Appendix). The stable ‘outer’ linear model

$$
\begin{pmatrix}
\Delta x_t \\
\Delta y_t
\end{pmatrix} = \mu + \begin{pmatrix}
\alpha_{11} & \alpha_{12} \\
\alpha_{21} & \alpha_{22}
\end{pmatrix} \begin{pmatrix}
1 & -1 \\
1 & 0
\end{pmatrix} \begin{pmatrix}
x_{t-1} \\
y_{t-1}
\end{pmatrix} + \begin{pmatrix}
\varepsilon_{1t} \\
\varepsilon_{2t}
\end{pmatrix} \quad (13)
$$

has a well-defined mean vector $(x_o, y_o)$. The typical behavior is obtained when $x_o$ is contained in $C = \{|x - x^*| < \delta\}$. Then, the mean is targeted for ‘large’ values of $x$ and, because of cointegration, also for large values of $y$, that imply large values of $x$. If the band $C$ is reached, i.e., if $x$ is ‘small’ again, $x_o$ is no longer interesting. Instead, the dynamic behavior of the variables resembles cointegrated processes, until the band is left and the cycle starts anew. Whenever $x_o$ falls outside the band, typical trajectories will remain near $(x_o, y_o)$ for long time spans. Only atypically large errors will shift them into $C$, where cointegrated behavior takes over. For $x_o \in C$, the set $C \cap \{(x, x) | x \in \mathbb{R}\}$ can be regarded as an ‘equilibrium’, with the error-correction vector possibly suitably shifted up or down by restrictions on $\mu$. For $x_o \in C$, the point $(x_o, y_o)$ constitutes a further element of the equilibrium or attractor set.

Hence, the threshold model allows for substantial variation in behavior. In concordance with the other models, we do not elicit informative priors but rather define non-informative reference structures with stochastic parameters. To this aim, we adhere to the following variant of the principle of insufficient reason.

Consider the traditional statistical problem of testing a point value against a finite interval. In that case, we would assume weights of $0.5$ for each hypothesis and a uniform prior on the interval for the ‘alternative’. Treating the present problem in an analogous manner, we use a normal distribution for $x^*$ and a half-normal distribution for $\delta$. These laws are sufficiently flat around $0$ to mimic the behavior of the typical uniform and normal laws. However, as a consequence of these assumptions, many processes show trajectories with little indication of threshold behavior. In fact, many trajectories closely resemble those drawn
from $H_1$. ‘Non-revealing’ trajectories occur if the threshold criterion is never
activated in the assumed sample length. Statistical criteria cannot be expected
to classify such cases correctly. Thus, it may be more difficult to discriminate
$H_3$ from $H_1 \cup H_2$ than to discriminate between $H_1$ and $H_2$.

2.3 The discriminatory statistics

In order to discriminate among the three candidate models, two discriminatory
statistics were employed. The statistic $S_2$ is designed to be powerful in discrim-
inating $H_1$ and $H_2$. In the notation of section 2.1, it would be labelled $S_{21}$. $S_2$
is defined as the smaller canonical root for $(\Delta x_t, \Delta y_t)$ and $(x_{t-1}, y_{t-1})$. As Jo-
hansen (1995) pointed out, this root makes part of the likelihood-ratio test for
hypotheses that concern the cointegrating rank of vector autoregressions. If the
larger canonical root is zero, $(x, y)$ forms a bivariate integrated process without
a stable mode. If only the smaller canonical root is zero, $(x, y)$ is a cointegrated
process with a stationary linear combination $\beta_1 x + \beta_2 y$. If the smaller canonical
root is also non-zero, $(x, y)$ is a stationary process with all modes being stable.
Because the rank-zero model is not acceptable for the investigated data, only
the smaller root is in focus.

An alternative statistic to $S_2$ would be a variant that uses the information
on the potential cointegrating vector $(1, -1)$. While that statistic would be
more powerful in discriminating $H_1$ and $H_2$ in the given form, we feel that some
flexibility regarding the exact form of the equilibrium vector is appropriate. In
other words, the decision maker is not supposed to know the vector exactly.

The statistic to appear on the $x$-axis, $S_1$, is designed to discriminate $H_1$ and
$H_3$. In the notation of section 2.1, it would be labelled $S_{13}$, though it may also
be useful in discriminating $H_2$ and $H_3$. $S_1$ is an approximate likelihood-ratio
test statistic for the stationary vector autoregression $H_1$ versus the quite spe-
cial threshold-cointegrating model with $\delta$ and $x^*$ being determined over a grid
of fractiles of $x$. In detail, $\delta$ is varied from the halved interquartile range to
the halved distance between the empirical 0.05 and 0.95 fractiles, with $x^*$ thus
assumed in the center of the range. For assumed models of type (12), the error sum of squares is minimized over a grid with step size of 0.05. As $T \to \infty$, this grid should be refined. Conditional on $x^*$ and $\delta$, estimating (12) is an ordinary least-squares problem with a corresponding residual covariance, whose log determinant can be compared to that of the unrestricted linear autoregression. The residual under the assumption of model class $H_j$ is denoted by $\hat{\epsilon}_t^{(j)}$. For $j = 1, 2$, this residual is calculated using the maximum-likelihood estimator. For $j = 3$, the approximate maximum likelihood estimator over the outlined grid is used.

The residual covariance matrix is denoted by $\hat{\Sigma}_j = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_t^{(j)}\hat{\epsilon}_t^{(j)'}$. With this notation, the second statistic is defined as $S_1 = \ln \left( \text{det} \hat{\Sigma}_3 \right) - \ln \left( \text{det} \hat{\Sigma}_1 \right)$.

Note that $S_1$ can be positive or negative. $S_1 < 0$ for strong nonlinear threshold effects, as $H_3$ attains a better fit. $S_1 > 0$ for stable autoregressions. The hypotheses $H_1$ and $H_3$ are not nested. The statistic $S_1$ optimizes the likelihood ratio over a limited range, which is the prevalent approach for nonlinearity testing (see, e.g., Caner and Hansen, 2001). An alternative is the semi-parametric test by Tsay (1998) that relies on sorting the observations according to the source of nonlinearity. The fully parametric nature of our decision problem, however, suggests the use of $S_1$. Another alternative would be a likelihood ratio $S_{23}$, which may be motivated by the fact that the hypotheses $H_2$ and $H_3$ are nested. We do not consider $S_{23}$, as it is an unusual statistic whose properties have not been fully explored.

The null fractiles were generated as follows. For $S_2$, Jordan priors were used on a cointegrating autoregression—the classical lower-dimensional ‘null’ hypothesis $H_2$. The cointegrating rank was fixed at one, whereas the cointegrating vector was not specified in constructing $S_2$. The null fractiles differ from those that were tabulated by Johansen (1995), as those were calculated under the hypothesis of multivariate random walks. For $S_1$, Jordan priors on stable vector autoregressions were used, according to the classical ‘null’ hypothesis $H_1$. 

14
2.4 Smoothing

The technique of discretizing the statistics in bins corresponds to a rectangular smoothing kernel in density estimation. Similarly, simulated boundaries of the decision areas often have a rough appearance, even for a high number of replications. It was found that smoothing the original numbers in the bins across neighboring bins is not as reliable as smoothing approximate posterior probabilities for the hypotheses. This effect is probably due to the scaling of numbers.

As a smoothing kernel, an inverse absolute function

\[ w(i, j) = \frac{W}{1 + |i - i_0| + |j - j_0|}, \quad i_0 - n_w \leq i \leq i_0 + n_w, \]
\[ j_0 - n_w \leq j \leq j_0 + n_w \quad (14) \]

was used over an \((2n_w + 1) \times (2n_w + 1)\) submatrix of the complete matrix of bins. This submatrix is centered at the location \((i_0, j_0)\), where the value is to be estimated. \(W\) is set such that the sum of kernel weights \(\sum w(i, j)\) over the submatrix equals one. Generally, the number \(n_w\) was selected as the minimum number, at which smooth boundary curves were obtained. The maps show some deliberate variation of \(n_w\), in order to demonstrate its influence on the results.

3 Simulation results: the maps

Figures 2 and 3 show decision maps resulting from \(3 \times 10^6\) simulations of process trajectories of length \(T = 50\) with stochastic parameters according to the prior distributions that were described in Section 2.2, i.e., \(10^6\) simulations for each model. These basic simulations use the simplified priors with \(\mu = 0\). In Figure 2, no smoothing was performed \((n_w = 0)\), while Figure 3 used \(n_w = 3\). One sees that the main effect of smoothing is a better separation of the preference areas for \(H_1\) and \(H_3\), which is mainly achieved by eliminating the scattered preference specks for the rather inhomogeneous hypothesis \(H_3\). Further increases of \(n_w\) distort the main boundaries of \(H_1\) and \(H_2\) and of \(H_1\) and \(H_3\).

The main features of the decision map are somewhat surprising. The threshold statistic \(S_1\) appears to be valuable in discriminating cointegrated and stable
Figure 2: Decision map for processes without deterministic part. $T = 50$, $n_w = 0$. The blue area prefers $H_2$, the yellow area prefers $H_1$ and the red area prefers $H_3$. The same color code is used for the other figures.
Figure 3: Decision map for processes without deterministic part. $T = 50$, $n_w = 3$. 
models, while it was designed to point out threshold structures. The Johansen-type statistic $S_2$ separates linear cointegrated models from cases of threshold cointegration, while it was designed to test for potential cointegration in stable vector autoregressions. A closer look reveals that $H_3$ models are indeed characterized by small values of $S_1$, as expected, and hence are most numerous in the left part of the chart. However, $H_2$ models also incur small values of the statistic $S_1$, as a threshold structure with the critical value pushed away from the starting values will achieve a better fit to data than an unrestricted vector autoregression for cointegrated models. In other words, linear cointegration results as a limiting case of threshold cointegration. The posterior probability for $H_2$ is more concentrated than that for $H_3$, hence $H_2$ dominates the left part of the diagram. Similarly, rejection of the smaller canonical root being zero may point to a stable linear model without unit roots but it may also point to a threshold model, which is stationary and ergodic though non-linear. The joint evidence of two non-zero canonical roots and a better fit by a restricted structure yields the preference area for $H_3$ in the north-west.

The map implies a crude empirical guideline. First, use the threshold statistic $S_1$. For values larger than the 0.3 fractile, a linear stable model is suggested. For smaller values, use the lesser canonical root $S_2$. If this root is ‘significant’ at 0.05, consider the threshold model, otherwise opt for linear cointegration.

In Figure 4, the sample size has increased to $T = 100$, while the other simulation parameters were retained. The smoothing bandwidth was kept at $n_w = 3$. The nominal significance level of the canonical root has decreased to 0.02, whereas the critical fractile of the threshold statistic decreases to 0.2. These features are in line with expectations regarding large-sample performance. There is now more evidence on a preference for threshold models in a wedge between the other two models, i.e., for threshold statistics in the fractile range $(0.2, 0.3)$, particularly if $S_2$ is ‘not too small’. The scattered appearance of the $H_3$ area reflects the varying shape of trajectories of length $T = 100$, which generally does not permit a safe classification. Many of these $H_3$ trajectories are indeed very similar or identical to $H_1$ trajectories from the stable model.
Figure 4: Decision map for processes without deterministic part. $T = 100$, $n_w = 3$.

Figure 5 shows the map for $T = 200$. The critical fractile of $S_1$ for $H_1 \cup H_3$ versus $H_2$ decisions shifts in to 0.12, while the implied significance level of $S_2$ for $H_2$ versus $H_3$ decisions has fallen to 0.01. The boundary is flanked by an unconnected preference area for $H_3$. Metaphorically speaking, it looks as if the dark $H_2$ curtain were drawn from a window that reveals a landscape that is populated by $H_1$ as well as $H_3$ models. Due to the large variation in appearance of $H_3$ trajectories, the typical shape of the decision map is likely to persist in even much larger samples. As a consequence, the empirical guideline
for sample sizes around $T = 200$ is to first have a look at the $S_1$ statistic. If it is situated between the 0.12 and 0.2 fractiles of its null distribution, $H_3$ or $H_1$ deserve consideration. Additional information from subject matter theory may help in the decision between these two classes. If $S_1$ is larger than its 0.2 null fractile, $H_1$ is recommended. If $S_1$ is less than its 0.12 null fractile, $S_2$ should be consulted. Highly significant values of $S_2$ imply a preference for $H_3$, otherwise $H_2$ is selected.

Figure 5: Decision map for processes without deterministic part. $T = 200$, $n_w = 3$.

For the smallest sample size considered, $T = 50$, Figure 6 shows the map for a variant with standard normal constants included in the data generation
mechanism. The assumption $\mu = 0$ in (3), (6), (10) was replaced by $\mu \sim N(0, I_2)$. The existence of a constant was also assumed in calculating the statistics $S_1$ and $S_2$. The effects of this intercept are different for each hypothesis. In $H_1$, only the mean is affected. In $H_2$, a linear trend is added. In $H_3$, a linear trend is generated within the inner regime. These asymmetric effects tend to simplify decisions between the model classes. Hence, the vertical boundary of the $H_2$ preference area is to the left of the comparable one in Figure 3. The critical fractile for this decision is around 0.1. The preference area for threshold models $H_3$ has grown considerably and shows a connected pattern in the upper part of the map between the 0.1 and 0.3 fractiles of $S_1$. The influence of the canonical root $S_2$ on the decision has disappeared with respect to the $H_2$ class and is rather secondary for the $H_1$ versus $H_3$ decision. As a rough guideline, this map suggests that threshold cointegration should be considered whenever $S_1$ is between its lower decile and lower quartile and $S_2$ is not in its lower tail. It is difficult to explain the $H_3$ preference in the north-east corner or the S–shape in the right-hand boundary of the main $H_3$ preference area. These features may be caused by specific properties of the prior distributions or may be artifacts.

For $T = 100$, the map of Figure 7 is obtained. The critical fractile of $S_1$ for the decision $H_1 \cup H_3$ versus $H_2$ has shifted in to about 0.05. By contrast, the boundary between the preference areas for $H_1$ and $H_3$ has hardly changed. The effect of the ‘drawn curtain’ is felt again, such that the $H_3$ preference area stretches down to the $x$–axis, i.e., to low $S_2$ values. The spot in the north-east persists. The asymptotic behavior suggested by Figure 7 is corroborated for $T = 200$ in Figure 8. The critical fractile for the $S_1$ statistic and the $H_2$ versus $H_1 \cup H_3$ areas decreases to 0.03, whereas the boundary between the hypotheses $H_1$ and $H_3$ remains in place. In this setup, hypothesis $H_2$ corresponds to the only transient model, which simplifies its detection in larger samples. Contrary to what may be an intuitive assumption, discriminating threshold structures from linear autoregressions is not automatically simplified as $T \to \infty$. Note, however, that the resolution of the grid and the range for the grid search were kept constant. For asymptotic properties such as consistency, such parameters
Figure 6: Decision map for processes with standard normal constant term. $T = 50, n_w = 3$. 
Figure 7: Decision map for processes with standard normal constant term. $T = 100$, $n_w = 3$. 
should depend on $T$.

Figure 8: Decision map for processes with standard normal constant term. $T = 200$, $n_w = 3$.

In unreported experiments, various modifications of the deterministic specifications were studied. It was found that the two reported variants—without deterministic part and with constant added to the basic dynamic model—represent important benchmark cases. In the presence of a strong ‘drift’, the decision according to the nonlinearity statistic $S_1$ should be regarded as separating $H_1 \cup H_3$ from $H_2$ and should not be overruled by any value of $S_2$. By contrast, if drifts can be ruled out \textit{a priori}, a clear vertical boundary separates $H_1$ from $H_2 \cup H_3$. This classification rule is confirmed by a further experiment that is reported in Figure 9. The difference from Figure 6 is that $H_2$ processes were generated with
a constant that is restricted to a multiple of the loading vector $\alpha$. Johansen (1995) showed that this restriction implies the absence of drift, thereby removing a part of the asymmetry among the unstable $H_2$ and the stable $H_1$ and $H_3$ in the experiments 6–8. Note that $H_2$ is still the only unstable case, though the divergence of its probability law as $T \to \infty$ comes at a slower pace. For the

![Figure 9: Decision map for processes with standard normal constant term and a no-drift restriction for linear cointegration. $T = 50, n_w = 4$.](image)

case of restricted constants, the behavior of the decision map for $T \to \infty$ can be guessed from the corresponding map for $T = 100$, which is shown as Figure 10. The preference area for $H_3$ in the northwest corner has shrunk, such that the implicit significance level of $S_2$ has decreased to 0.01. Scattered $H_3$ preference
specks appear across a rather ample area.

Figure 10: Decision map for processes with standard normal constant term and a no-drift restriction for linear cointegration. \( T = 100, n_w = 4 \).

The last variant to be reported concerns the complex-roots modification of the Jordan forms of the priors for \( H_1 \) and the outer regime of \( H_3 \), which was mentioned in Section 2.2. 50% of the coefficient matrices for hypotheses \( H_1 \) and the outer regime of \( H_3 \) were drawn from \( TJ_cT^{-1} \) defined in (6) and 50% from the hitherto used \( TJT^{-1} \) with two real roots. For \( H_2 \) and the inner band of \( H_3 \), one root is fixed at unity, which implies that the other root is real. These designs remain unchanged. The deterministic constant \( \mu \) was set at zero, as in the basic experiments shown as Figures 3–5. The differences between Figure 11
and the corresponding map with an all-real design in Figure 4 are due to the cycles caused by the conjugate complex roots in $H_1$ and $H_3$. The main effect appears to be the absence of an extended support area for $H_3$ around the point $(0.2,0.9)$. A tentative explanation is as follows. The complex roots in $H_1$ and $H_3$ increase the average ‘distance’ between the cointegrated model $H_2$ and the other models. The increased accuracy of decisions among $H_1$ and $H_2$ reduces the chance of the ‘compromise’ hypothesis $H_3$ to rule areas of conflict. As in the other maps for models without a constant, the statistic $S_1$ separates the preference areas for $H_1$ and $H_2 \cup H_3$, while the statistic $S_2$ separates $H_2$ and $H_3$, conditional on the first-round decision.

For a final note, one may return to the empirical example shown in Figure 1. The generating mechanism of this data set does not correspond exactly to the design of the maps, for two reasons. Firstly, the sample size is $T = 470$, much above the $T = 200$ that was used as a maximum in this section. Secondly, the autocorrelation function suggests a slightly longer memory in the process than a first-order autoregression, even a nonlinear one. A tentative calculation of the null fractiles according to the version with restricted drift constant and of the statistics $S_1$ and $S_2$ yields the coordinates $(0.01,0.95)$. A simulation of the decision map for this large data set yields a leftward shift of the critical fractile for $S_1$ to around 0.02. Therefore, the evidence favors the model class $H_2$, i.e., the non-ergodic cointegrated models. If this class is regarded as unacceptable, it may be removed from the set of available alternatives. Then, the class $H_3$ with the second largest value of its predictive distribution is supported. Even with all caveats, the exemplary data set confirms the common observation that the boundaries are not hit frequently enough to provide clear evidence of their existence.

4 Summary and conclusion

Decision maps can be useful in their own right. Assuming a researcher analyzes a data set and his or her $a\ priori$ plausible hypotheses correspond to the set-up of
Figure 11: Decision map for processes with zero constant and a 0.5 chance of conjugate complex roots in stable autoregressive coefficient matrices. $T = 100$, $n_w = 3$. 
the decision map simulation. If the coordinates of the decision map are available, two statistics $S_1$ and $S_2$ can quickly be calculated from the data and can be encoded as null fractiles. Otherwise, a simulation may be used for performing the encoding. In the map, preference areas for the conflicting hypotheses are clearly separated by boundaries.

Decision maps are, however, even more important as summary guidelines for the empirical researcher. Vertical or almost vertical boundaries indicate that relying on $S_1$ is almost as valuable for discriminating the hypotheses as the joint evaluation of $S_1$ and $S_2$. Similarly, horizontal boundaries underscore the value of $S_2$ relative to $S_1$. In the present experiments, it was found that the main discriminatory power rests on the threshold statistic $S_1$, not only among hypotheses $H_1$ and $H_2$, i.e., stationarity and cointegration. Correct identification of threshold models turned out to be extremely difficult even for relatively large samples such as $T = 200$. This difficulty is corroborated by tentatively removing $H_2$ from the set of available hypotheses. The implied map shows that $H_2$ and $H_3$ approximately dominate the same area, a vertical band, with $H_2$ more concentrated there. Formally, the decision maps suggest decisions for $H_3$ around the boundary between the preference areas for the other two hypotheses, i.e., for $S_1$ values around the ‘critical values’. Another preference area for $H_3$ is the more ‘natural’ one in the northwest corner, where $S_2$ is in the upper tail of its null distribution. The empirical recommendation by some authors (see LO AND ZIVOT, 2001) to test for cointegration in a linear frame first and then to check for nonlinearity is not generally confirmed. Rather, the maps recommend to test for nonlinearity first. Structures with sufficiently large values of $S_1$ are classified as stationary $H_1$ models. As a second step, a cointegration test is conducted. If cointegration is rejected for a model classified as ‘nonlinear’, a threshold model $H_3$ is indicated. If cointegration cannot be rejected, a cointegrated linear model $H_2$ is supported. Traditional simulation with fixed parametric designs could never unveil this general decision pattern.

The unconnected specks of preference for $H_3$ reflect the fact that threshold processes generate two species of trajectories: typical trajectories with statistics
clustered in the northwest region and atypical trajectories with hardly recognizable threshold effects and statistics almost anywhere in the left part of the $[0,1] \times [0,1]$-plane. Most atypical trajectories stem from designs with small values of $\delta$ and therefore roughly ‘look like’ trajectories from stationary autoregressions. The high risk of incorrectly classifying the generating processes as $H_1$ may incur a relatively modest risk if one proceeds with the incorrect model, as the linear model may be a good workhorse for typical econometric tasks such as prediction. A careful evaluation of this conjecture is a promising task for future work.

Many researchers may be skeptical about the usage of decision maps, particularly when the dynamic specification of short-run nuisance for $H_j, j = 1,\ldots,3$ is slightly simpler than time-series structures that prevail in the literature. In order to counter this argument, more sophisticated priors must be introduced, which unfortunately entails a considerable increase in computer time. For an example of higher-order autoregressions and elements of lag order search via information criteria within the decision-maps framework, see Kunst (2001). Such extensions are possible directions for future research.

Acknowledgments

The data on U.S. interest rates are taken from the International Financial Statistics database. The author thanks Manfred Deistler, Elizaveta Krylova, Sylvia Kaufmann, and an anonymous referee for helpful comments, and Elizabeth Raab for proofreading. The usual proviso applies.

Appendix: Geometric ergodicity of threshold cointegrated models

The recent econometric literature on threshold cointegrated models offers no formal proof of the stability properties of threshold cointegrated models with
stable outer regimes. For the variant that is used as hypothesis $H_3$ in the paper, such proof is provided here.

The ‘threshold cointegration model’ is defined as the nonlinear first-order autoregressive structure

$$
\begin{pmatrix}
\Delta x_t \\
\Delta y_t
\end{pmatrix} = \alpha \beta' \begin{pmatrix} x_{t-1} \\
y_{t-1}
\end{pmatrix} + \begin{pmatrix} \gamma_1 \\
\gamma_2
\end{pmatrix} x_{t-1} I(|x_{t-1} - x^*| > \delta) + \varepsilon_t. \quad (15)
$$

For simplicity, no deterministic terms are used except for the $x$ center $x^*$. The model is equivalent to a stable autoregression for the outer region $\{|x_{t-1} - x^*| > \delta\}$ and to a cointegrating partial stable autoregression for the inner region. We assume the following conditions:

A1: The polynomial $\det\{I_2 - (I_2 + \alpha \beta') z\}$ has no roots inside the unit circle or for $|z| = 1$ but $z \neq 1$. The second-order matrix condition for the Granger representation theorem (see Engle and Granger, 1987, and Johansen, 1995) $\det \alpha' \beta' \neq 0$ applies, where the subscript $\perp$ denotes the orthogonal complement.

A2: The polynomial $\det\{I_2 - (I_2 + \alpha \beta' + \gamma \epsilon_1)\}$ has all roots outside the unit circle, where $\epsilon_1 = (1, 0)'$.

For the errors $\varepsilon_t$ a regularity condition is assumed:

A3: The distribution of the errors $\varepsilon_t$ is absolutely continuous and strictly positive on $\mathbb{R}^2$.

Conditions A1 and A2 guarantee that the model corresponds to the above concept, with A1 essentially due to Engle and Granger (1987) and to Johansen (1988) and A2 a standard assumption of time series analysis. Given A1, note that A2 restricts $\gamma$ and excludes $\beta = (\lambda, 0)'$ for arbitrary $\lambda \neq 0$. A3 implies irreducibility and aperiodicity for all threshold autoregressive models. With these assumptions, the following result by Tong (1990, p. 457) can be applied:

Theorem 1 (Drift criterion, Tong) Let $\{Z_t\}$ be aperiodic and irreducible. Suppose a small set $C$ exists, a non-negative measurable function $g$, and con-
\begin{align}
\text{constants } r > 1, \; \gamma > 0, \text{ and } B > 0 \text{ such that }
E\{rg(Z_t) | Z_t = z\} < g(z) - \gamma, \quad z \notin C \tag{16}
\end{align}

and

\begin{align}
E\{g(Z_{t+1}) | Z_t = z\} < B, \quad z \in C. \tag{17}
\end{align}

Then, \{Z_t\} is geometrically ergodic.

The small set \(C\) is meant to contain the ‘center’ of the stationary distribution, assuming its existence. For a first-order autoregression with stable coefficient, the condition (16) holds for any subset of \(\mathbb{R}^2\) outside the mean, that is, outside a disk around zero if there are no deterministic terms. For the definition of a small set, see Tong (1990, p. 454). A technical complication is to prove that compact sets are small. For all nonlinear autoregressions of the threshold type, this can be shown as in Chan et al. (1985). This result implies the following.

**Theorem 2** Let \(Z_t = (x_t, y_t)'\) for \(t > 0\) obey the model (15) with the conditions A1–A3 and arbitrary fixed starting conditions. Then, \(\{Z_t\}\) is geometrically ergodic.

**Proof:** Decompose the space \(\mathbb{R}^2\) into five disjoint areas such that \(\mathbb{R}^2 = \bigcup_{j=1}^{5} A_j\). We analyze all of them in turn.

1. \(A_1 = \{x < x^* - \delta\}\) The process is locally geometric stable and condition (16) is fulfilled for many functions \(g(z)\), among them all absolute values of linear functions in the arguments \(x\) and \(y\), provided that the implied mean of the autoregression \(Z_t = (I_2 + \alpha \beta' + \gamma e'_i)Z_{t-1} + \varepsilon_t\) is outside \(A_1\). The maximum eigenvalue \(\lambda\) of the regressor matrix is less than one in modulus because of A2, hence any value \(\lambda'\) in the open interval \((|\lambda|, 1)\) can be chosen for \(1/r\). \(\gamma\) can be set to \((\lambda' - \lambda) \min \{g(z) | z \in A_1\}\}. If the implied mean \(\mu\) is inside \(A_1\), the proof must be formulated with respect to \(A_1' = \{x < x^{**} - \delta\}\) for \(x^{**}\) being the \(x\)-component of \(\mu\). The area \(A_1' - A_1\) is appropriately allotted to \(A_3 \cup A_4 \cup A_5\).
2. \( A_2 = \{ x > x^* + \delta \} \) Same as \( A_1 \). Again, \( A_2 \) may be replaced by \( A_2^* \) if necessary.

3. \( A_3 = \{ |x - x^*| < \delta \text{ and } |y| < K \} \) with \( K \) chosen large enough that the inner equilibrium line segment \( \beta' (x, y)' \) is fully contained in \( A_3 \). Note that this construction assumes that the second element of \( \beta \) is non-zero, which is excluded by assumption \( A_2 \), as the system given in \( A_2 \) and defined by the ‘outer regime’ cannot become stable if both cointegrating vectors coincide. Clearly, condition (17) is fulfilled. The implication is unaffected by the change from \( A_1 \) to \( A_1^* \) and the implied change of \( A_3 \) to \( A_3^* \). To show that \( A_3 \) is small under the assumptions \( A_1 - A_3 \), we refer to Tong (1990) who states that, for locally linear models with error distributions satisfying \( A_3 \), all compact sets are small.

4. \( A_4 = \{ |x - x^*| < \delta \text{ and } y > K \} \). Defining \( g(x) \) as the distance to the equilibrium line segment, for example in the Euclidean metric, yields condition (16) for this area. This function is also valid for \( A_1 \) and \( A_2 \). The implication is unaffected by the change from \( A_j \) to \( A_j^* \) for \( j = 1, \ldots, 5 \).

5. \( A_5 = \{ |x - x^*| < \delta \text{ and } y < K \} \). Same as \( A_4 \).

\( A_3 \) is small in the sense of Theorem 1, which completes the proof. \( \blacksquare \)

Note that Theorem 2 gives no result for the case that the inner regime does not cointegrate. In fact, then ‘probability mass escapes’, as trajectories may wander in the direction of the unrestricted variable \( y \). The result by Chan et al. (1985) does not generalize to the multivariate case immediately, when the inner area is not completely bounded. A similar observation holds with respect to the degenerate case where the equilibrium line segment is vertical in the sense that \( \beta = (\lambda, 0)' \). Obviously, the proof is unaffected by changing the signal variable \( x \) to \( y \) or to any linear combination of \( x \) and \( y \) different from \( \beta' X \).
References


