The "V-Factor": Distribution, Timing and Correlates of the Great Indian Growth Turnaround: Web Appendix

Chetan Ghate* and Stephen Wright†

August 31, 2011

*Corresponding Author. Address: Planning Unit, Indian Statistical Institute, 7 SJS Sansanwal Marg, New Delhi 110016, India. Tel:91-11-4149-3938; Fax:91-11-41493981. E-mail:cghate@isid.ac.in

†Department of Economics, Birkbeck College, University of London, Malet Street, London W1E 7HX, UK. s.wright@bbk.ac.uk

A Data Sources and Definitions

A.1 Figure 1

Source: Net State Domestic Product (NSDP) is from the Economic Political Weekly Research Foundation (2005) dataset on Indian states. The sectoral definitions and sectors are:"Agriculture" includes agriculture, forestry and fishing; "Mining"; "Manufacturing includes registered and unregistered manufacturing; "Construction"; "Trade" includes trade, hotels and restaurants; "Transport, Electricity" include Transport, Storage and Communication plus Electricity, Gas & Water; "Banking" includes Financing, Insurance, Business Services; "Real Estate"; "Public" includes Public Administration and Defence; and, "Other Services".

All series are at constant 93-94 prices projected back using earlier base years.

A.2 Figure 2

Source: The Net State Domestic Product data have been assembled from various tables in the EPW Research Foundation (2005) dataset, the most comprehensive and up to date dataset on Indian states. The observations have been spliced so that all states have real NSDP figures in constant 1993-1994 prices, divided by state population (interpolated between census dates). Our method of splicing ensures that our measures of state RNSDP are largely immunized from the impact of various changes in state definition.[34]

A.3 Panel dataset Used in Section 3

Our core dataset contains output per capita data for 15 major states (the same list of states as for Figure 2, excluding Jammu and Kashmir) using data from the EPW Research Foundation, for fourteen sectoral headings. All data have been spliced so that the underlying sectoral data are in constant 1993-1994 prices, converted into per capita terms using total state population as for Figure 2. The sectoral series for each state are: 1)Agriculture, 2)Forestry and Logging, 3)Fishing, 4) Mining and Quarrying, 5) Registered Manufacturing 6) Unregistered Manufacturing, 7) Construction,

[34] These changes mainly affect Bihar and, to a lesser extent, Madya Pradhesh and Assam. Details of precise methodology are available from the authors.

8) Electricity, Gas and Water Supply, 9) Transport, Storage and Communication, 10) Trade, Hotels and Restaurants, 11)Banking and Insurance, 12) Real Estate, 13) Public Administration, 14) Other Services.

We eliminate three series from the panel due to clear errors: published data for Electricity, Gas and Water are negative in some years for Assam and Haryana; and published data for real estate in Kerala have clear discontinuities. We also investigate below the implications of omitting some other series that may contain rogue observations.

If we exclude data for Assam, Bihar and Orissa we have a full sectoral breakdown for the remaining 12 states from 1965; if we also exclude Haryana and Punjab we have data for the remaining 10 states from 1960.

A.4 Consumption

To calculate aggregate nominal consumption expenditures by states, we generated a pseudo-panel by utilizing data from various NSS rounds which provide data on nominal monthly mean per capita rural consumption and nominal monthly mean per capita urban consumption These numbers were multiplied by 12 to generate annual figures, and then multiplied by observations for rural and urban population shares. The population data are tabulated from Census figures, with a common compound growth rate applied across decadal observations to impute annual observations for each state. We cross check these figures with population figures obtained by simple extrapolation: (NRSDP/PCNRSDP)*10000000. Both the census figures and extrapolated figures are consistent with each other. Rural Population and Urban Population proportions are then obtained from various rounds of the NSS surveys to give us a full series of rural and urban annual population figures from 1960 - 2005.

To calculate aggregate real consumption expenditures by states, we followed a similar procedure. We generated a pseudo-panel by utilizing data from various NSS rounds on real monthly mean per capita rural consumption (at 1973-74 all India rural prices), real monthly mean per capita urban consumption (at 1973-74 all India urban prices), and population data.

Aggregate annual rural consumption (in crore) is given by: real monthly mean per capita rural consumption $\times 12 \times$ rural population for a given state in a given year.

Aggregate annual urban consumption (in crore) is given by: real monthly mean per capita urban consumption × 12 × urban population for a given state in a given year.

Total state (nominal) real consumption expenditures (in crore) is given by: Aggregate (Nominal) Real Rural Consumption + Aggregate (Nominal) Real Urban Consumption / 10000000.

B Unit Root Tests

Table A1 summarises the results of unit root tests on both the underlying series in the panel, and on the estimated transitory components, calculated as in (3).

[Insert Table A1]

It first reports the panel unit root test as in Im, Pesaran and Shin (2003), which tests the null that all series in the panel have a unit root, and allows for heterogeneity of auto-regressive coefficients under the alternative. The unit root null cannot be rejected for the underlying series, a feature which is accentuated by the result that almost exactly half the individual ADF test statistics are below and above the expected value under the unit root null.

For all three of the estimated transitory components when the factors are estimated by principal components, the null is strongly rejected. This is in itself not an especially strong result, since it is well-known (see, for example, Shin & Snell, 2006), that the null will be rejected if even a quite small number of series being tested (sometimes even a single series) are stationary. More revealing is the distribution of individual ADF statistics, which is shown in Figure A1 for the two models estimated in levels, and in Figure A2 for the model estimated in differences. In all three cases, as Table A1 shows, a much higher proportion of individual test statistics are below the expected value than would be expected under the unit root null, but this feature is clearly very much more evident for our central case using levels estimation and two factors, for which only 3% of individual test statistics are above the expected value. Thus we have particularly strong evidence of stationary transitory components for this, our central case.

[Insert Figure A1]
[Insert Figure A2]

C Data Construction for Figure 4

For Figure 4, we let $\widehat{F}_{1t}$ and $\widehat{F}_{2t}$ be the first and second principal components respectively, (normalized to have zero mean and unit variance, these are the "G-Factor" and "V-Factor" as defined in Figure 4) derived from the sample autocorrelation matrix of y_{it} (or equivalently, from the autocovariance matrix of the series after demeaning and rescaling to have unit sample variance). The series $PC1$ is the cumulated first principal component extracted by the same method from the panel of differenced data as in Bai and Ng (2004).

D Robustness Checks for V-Factor Estimates

D.1 Robustness to changes of time sample

As noted in the main paper, our core analysis is carried out on a balanced panel of data for 15 states. However, as discussed in Appendix A.3, for a subset of ten states we have a longer run of data, back to 1960. A natural robustness check for the dating of the turnaround in the V-Factor is to use the longer datasets, despite the reduction in the cross-sectional dimension (in Appendix G we show that simulation evidence that the gains from increasing T appear to more than offset the losses from decreasing N). Figure A3 shows the results of this experiment. The two alternative estimates of the V-Factor have an identical timing of their minima, and extremely similar paths thereafter. There are somewhat greater differences in earlier years but overall the profiles of all three estimates appear reassuringly similar. It is striking how robust the estimates are both to the inclusion of the additional years and the exclusion of a subset of states.

[Insert Figure A3]

D.2 Robustness to changes of cross-sectional sample

As a further robustness check we also investigate, in our panel from 1970 onwards, the impact of removing certain categories of series from the estimation of the principal components. Table D1 and Figure A4 summarise the impact of these changes.

Table D1 lists the exclusions from the cross-section. The first four exclude data based on state characteristics; the next three exclude series by broad industry type. We also show the impact of excluding series with high levels of volatility, and, for comparison, the impact of prior-filtering data for the short-term impact of fluctuations in rainfall (see next section). The table also shows N, the cross-sectional dimension, the correlation, across the cross-section, between actual changes in growth rates and fitted values implied by the estimated V-Factor and G-Factor, as discussed in Section 3.5, as well as showing the year in which the estimated V-Factor reaches its minimum

[Insert Figure A4]

The first notable feature illustrated by Figure A4 is how similar the broad profiles of the estimated V-Factors are after all these adjustments (as in all other comparisons the estimates are all normalized to have unit mean and variance), despite significant differences in sample both in terms of the change in N, and in terms of the characteristics of the series. All estimates also provide similarly good representations of the shift in growth.

The second notable feature is that, while adjustments for more volatile series have only a modest impact on longer term properties of the estimated V-Factor, they do (unsurprisingly) have some influence on short-run movements. Figure A4 makes it clear that the sharpness of the minimum point in 1987 for the estimated V-Factor using the full cross-section is reduced, or disappears entirely, in any sample that excludes agriculture, forestry and fishing, in particular, and that as a result for these reduced cross section the minimum occurs a year or, at most, two years later. In the light of our simulation results, discussed below in Appendix G, which show that the true minimum point is only reasonably well estimated to within a year or two either side, this should not be viewed as surprising.

Table D1. Impact on estimated V-Factors of excluding series from the panel

D.3 Robustness to rainfall adjustment

As an additional check to adjust for short-run volatility, we prior-filter the data in first differenced form by regressing on a constant and the change in log rainfall over the previous year, and then replace each of the underlying series with the cumulated error from this regression. In the case of agricultural output in particular we find strongly significant positive impacts of rainfall changes, and hence a reduction in the remaining volatility of the series. The impact of rainfall on other sectors is typically less significant. Figure A4 and Table D1 again show that the impact of the adjustment on the V-Factor estimate is very small.

E Policy Indicators and Data Construction and Sources for Figure 7

The V-Factor is equal to $\widehat{F}_{2t}$ as in Figure 4. The effective tariff rate is constructed consistently with Rodrik and Subramanian (2005, Figure 4.) The central government customs duties collection (in crore) and imports (in crore) are from the Reserve Bank of India statistical tables. The effective tariff rate is approximated as Customs Duties Collection/Imports. The Real Exchange Rate data (REER) and the log openness ratio was assembled from the Reserve Bank of India (RBI) database on the Indian Economy. Duties as a percentage of GDP is defined as customs duty collection (in crore) / GDP at factor cost (in crore). This was also obtained from the RBI dataset. See www.rbi.org.in.

F Data Construction and Sources for State-level Regressors inTable 3

The pro-worker dummy is taken from Aghion et al (2008).

The dummy for landlocked states is equal to unity for all series for Assam, Bihar, Haryana, Madhya Pradesh, Punjab, Rajasthan, Uttar Pradesh, and is zero otherwise

The other state characteristics used in the regressions in Table 3 are taken from a new panel dataset for Indian states assembled by the authors comprising roughly 200 regional economic and social indicators for Indian states. A detailed description of the variables in this dataset, and the data used in Table 3, is available in the data appendix in an earlier working paper version of this paper; Ghate and Wright (2008).

G Simulation Methodology

We simulate a system with an underlying common structural shift which is a parameterised version of (1) to (3), as follows

$$\begin{aligned}
y_{it} &= \beta_{i0} + \beta_{i1}F_{1t} + \beta_{i2}F_{kt} + u_{it}; i = 1..N && (5)\\
\Delta F_{kt} &= g_{k1} + \varepsilon_{kt}; t \leq t_b \\
&= g_{k2} + \varepsilon_{kt}; t > t_b; k = 1, 2 && (6)\\
u_{it} &= \gamma_{i1}Q_{1t} + \gamma_{i2}Q_{2t} + r_{it} && (7)\\
Q_{jt} &= \rho_j Q_{jt-1} + \xi_{jt}; \quad j = 1, 2 && (8)\\
r_{it} &= \rho_i r_{it-1} + \omega_{it}; \quad i = 1..N, && (9)
\end{aligned}$$

In (1) we simulate each of the N series as a sum of factor loadings on two $I(1)$ factors, plus a persistent residual component. The two $I(1)$ factors, F_{1t} (the simulated "G-Factor") and F_{2t} (the simulated "V-Factor") are modelled in (6) as drifting random walks with shifts in growth rates at the break point t_b. The transitory components u_{it} are then in turn driven by two common stationary factors, Q_{1t} and Q_{2t} which capture any remaining mutual correlation in the y_{it} after extraction of the two permanent components, plus a strictly idiosyncratic component, r_{it}. The Q_{jt} are modelled in (8) as stationary AR(1) processes without shifts (we examine below the impact of including or excluding these additional stationary factors). We estimate the process for the two permanent and two stationary factors from the time series properties of the first four principal components of the dataset. The data point to a highly significant shift in growth at $t_b = 1987$ for the "V-Factor" ($g_{21} < 0$; $g_{22} > 0$); with a smaller, but still significant shift for the "G-Factor ($0 < g_{11} < g_{12}$). While conventional tests of significance are suspect due to a data mining critique, the primary objective is to simulate a null model where there is a structural shift in growth that also matches the broad properties of our dataset. The estimation procedure for the factor processes is thus for purposes of calibration, rather than to carry out any direct hypothesis testing. The correlation matrix of the vector of estimated factor innovations $\left[\begin{array}{cc}\widehat{\varepsilon}_t' & \widetilde{\xi}_t'\end{array}\right]'$ is close to diagonal in the data so we simulate the four factor innovations as orthogonal processes.

The factor loadings $\left\{\{\beta_{ik}\}, \{\gamma_{ij}\}\right\}$ are calibrated to match (subject to minor modifications noted below) those of the estimated factor loadings on the principal components in the data. Each element is modelled as an independent draw from a normal distribution with mean and standard deviation given by the cross-sectional mean and standard deviation of the loadings on each of the principal components in the data. The simulated orthogonality of the factor loadings that results from this methodology is consistent with the orthogonality (by construction) of factor loadings derived by the method of principal components.

Finally in (9) we model the residual idiosyncratic components, the r_{it} as AR(1) processes with mutually uncorrelated innovations. The $\{\rho_i\}$ and the $\{\sigma_i\}$, (where $\sigma_i = E\left(\omega_{it}^2\right)$) are modelled as independent draws from uniform distributions that approximate the key cross-sectional properties of these parameters in our dataset. We draw from a uniform, rather than normal distribution, since we need to impose bounds on both sets of parameters, such that $\rho_i \in (-1, 1)$, $\sigma \in (0, \infty)$. We calibrate these distributions to match the cross-sectional means and standard deviations of the estimated parameters in our dataset, subject to these inequalities.

Reassuringly the simulation methodology gives a generally good match of the key properties of the dataset. We make only two minor modifications to ensure that the simulated contribution of the two nonstationary factors to the total variance in the dataset is on average (across simulations) equal to that in the data (since we do not wish to over- or understate the importance of these two factors in our simulations). This is achieved by raising $\overline{\beta_{i1}}$, the cross-sectional mean loading on the "G-Factor" from 0.0266 in the data to 0.032 in the simulations (this ensures a match of the average contribution of the first factor in the simulations), and by reducing $\sigma(\beta_{i2})$, the cross-sectional standard deviation of the loadings on the "V-Factor" from 0.030 in the data to 0.025 (this ensures a match of the average contribution of the second factor in the simulations).[35] Given the approximations involved in our simulations (in particular the distributional assumptions for the parameters), the magnitude of the changes required is reassuringly modest.

Table G1 summarizes the key results of our simulations. The first row shows our base case. In each artificial sample we simulate a balanced panel of 207 series all starting in 1970, where the true break year, t_b is set at 1987, in line with the profile of the V-Factor shown in Figure 4 in the main paper. The results show that if the true data generating process has the same breakpoint, the 2nd principal component in levels would identify the breakpoint in the true V-Factor (simulated as F_{2t}) to within ± 1 year in 60% of replications.[36]; in comparison the cumulated 1st principal component in differences has an equivalent percentage of only 32%. Both approaches are somewhat biased: i.e., if the true breakpoint year were 1987, on average both approaches would estimate it to be 1988. But this bias is to be expected since it arises from the AR(1) processes assumed for the u_{it}, such that the mean lag from the impact of a shift in the factors, given by $\rho_i/\left(1-\rho_i\right)$ is always positive. Based on our dataset, ρ_i ranges from -.15 to .67, hence the simulated mean lags range from zero to roughly 2, hence a bias of around one year is to be expected.

The second row of the table shows that if we simulate a smaller cross section, over a longer sample (as in Figure A3), the loss of precision from a lower cross-section appears to be more than offset by the gain in precision from a longer sample.[37]

The third row of the table shows the impact of excluding the impact of the two additional stationary factors. Using both techniques there is a clear improvement, unsurprisingly so, since all remaining variation in the y_{it} is due to the mutually orthogonal u_{it} terms. The improvement in the performance of the approach in differences is particularly marked, but it remains less reliable than the levels approach; albeit only marginally so. The much greater sensitivity to the exclusion of the stationary factors does

[35] The mean loading on the V-factor is close to zero in the data, and we retain this feature in the simulations.

[36] Note that the proportions shown in the table are when the minimum of the estimated component matches that of F_{2t}. This does not always match the true breakpoint, since, given random variation in the simulated F_{2t}, it does not always reach a minimum in the "true" breakpoint year.

[37] If we increase T and decrease N separately the impacts are, as would be expected to improve and decrease precision respectively.

however indicate a lack of robustness of this approach (we show below that this conclusion is further strongly reinforced by the comparative performance of the two approaches with a stochastic breakpoint).

This improvement in identification of breakpoints in the smaller cross-section over a longer sample is clearly a helpful result in itself, but all the more so if we wish to distinguish between the break point of 1987 identified in our dataset and the earlier breakpoints identified in past research. We note in the main paper that some studies have concluded that there was a break point as early as the late 1970s. In the fourth and fifth rows of the table we simulate an alternative data generating process consistent with this earlier breakpoint. With the shorter sample and a larger cross-section neither of the two approaches would be very successful in identifying such an early breakpoint (i.e. only 9 years into the sample); however the fourth row of the table shows that with a longer sample but a lower cross-section the earlier break point would still be reasonably well estimated. We can use this simulated DGP to assess the probability of estimating a break point in 1987 (as in our dataset), or later, if the true breakpoint were in 1979: using principal components in levels this occurs in only 3% of simulations, suggesting that the technique we use can discriminate well between an earlier and a later breakpoint.

A more general way of assessing how well the two alternative techniques perform in identifying breakpoints is summarized in the last two rows of Table G1 and in Table G2. These show the results of allowing the breakpoint to be a random variable across simulations. The true breakpoint t_b is drawn for each simulation as a uniform random variable ranging between 1982 and 1992. The precision with which the breakpoint is estimated by both techniques falls somewhat, but the proportions of simulations in which the estimated breakpoint is within a year of the true breakpoint are quite similar. Table G2 shows that using the levels approach the estimated breakpoint is quite strongly positively correlated with the true breakpoint across the simulations (with correlation coefficient 0.7) but that it does not typically move one for one: essentially there is some bias (albeit not especially strong) towards finding a breakpoint at or near the mid-point of the sample. In contrast Table G2 shows that the estimated breakpoint using the differences approach is only weakly correlated with the true breakpoint across different simulations

[Insert Tables G1 and G2]

Finally we note that the comparative properties of the simulations summarized above, which focus (for obvious reasons) on the identification of the breakpoint, are not dependent on the assumption that the deterministic component of the "V-Factor" is precisely V-shaped. We have also experimented with an alternative DGP in which the second factor is roughly "U"-shaped - i.e., closer to the shape identified by the differences approach in our dataset, as illustrated in Figure 4. The ranking of the two approaches, expressed in terms of the correlation between the estimated principal component and the true factor, remains the same in all cases. When the true factor is a "U"- rather than a "V"-factor this property is captured fairly well in the majority of simulations by the levels approach: i.e. there is no bias in estimation towards finding "V"- as opposed to "U"-Factors.

Thus we can feel reasonably confident that, even if the breakpoint of the true V-Factor cannot be precisely identified in our dataset, it seems likely to have occurred within a year or two of the estimated breakpoint of 1987. Furthermore, it does appear that the turnaround was relatively rapid: thus a "V"-Factor representation does appear valid.

Figure A1
Ranked ADF Statistics for Transitory Components from Levels Estimation

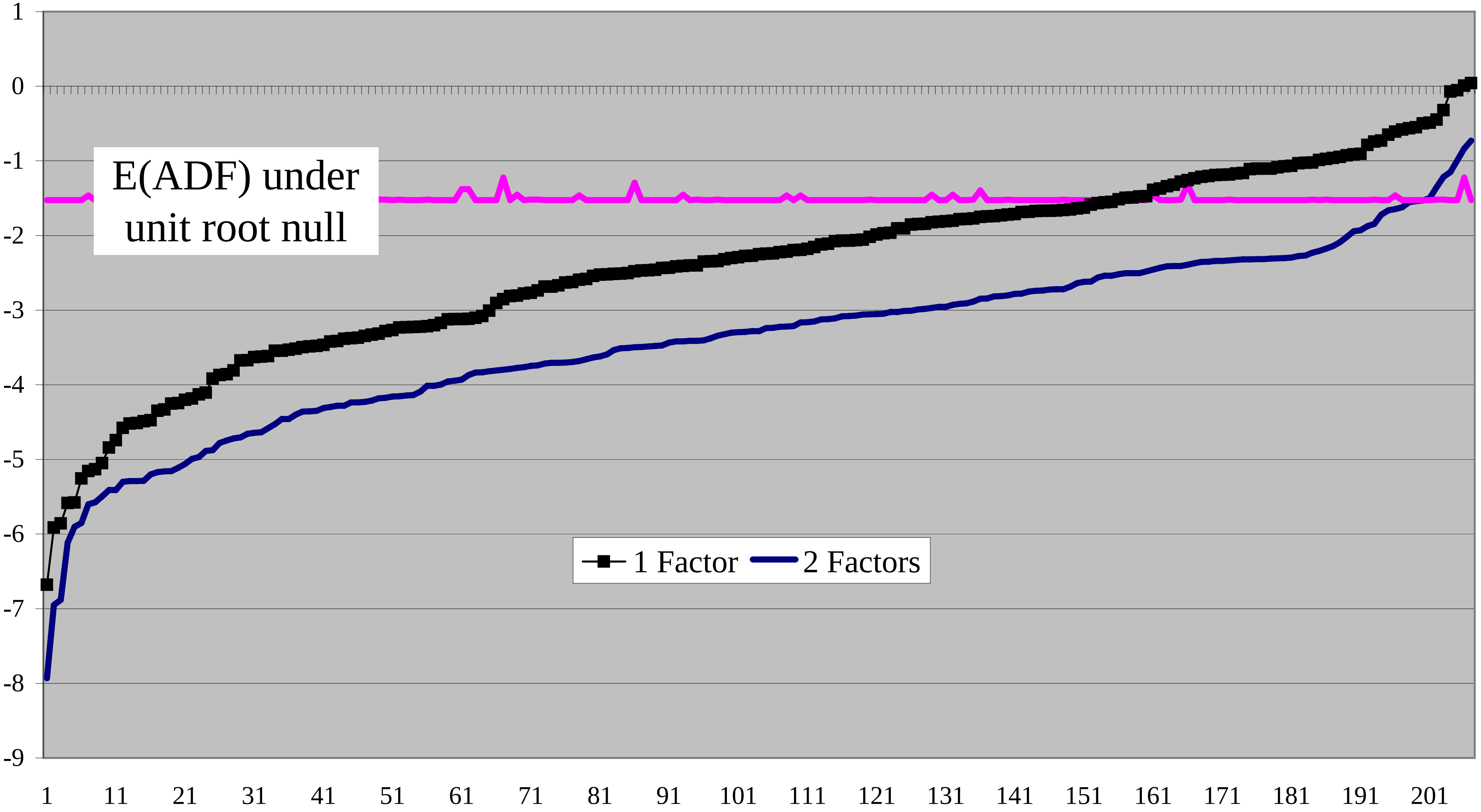

Figure A2
Ranked ADF Statistics for Transitory Components from Estimation in Differences

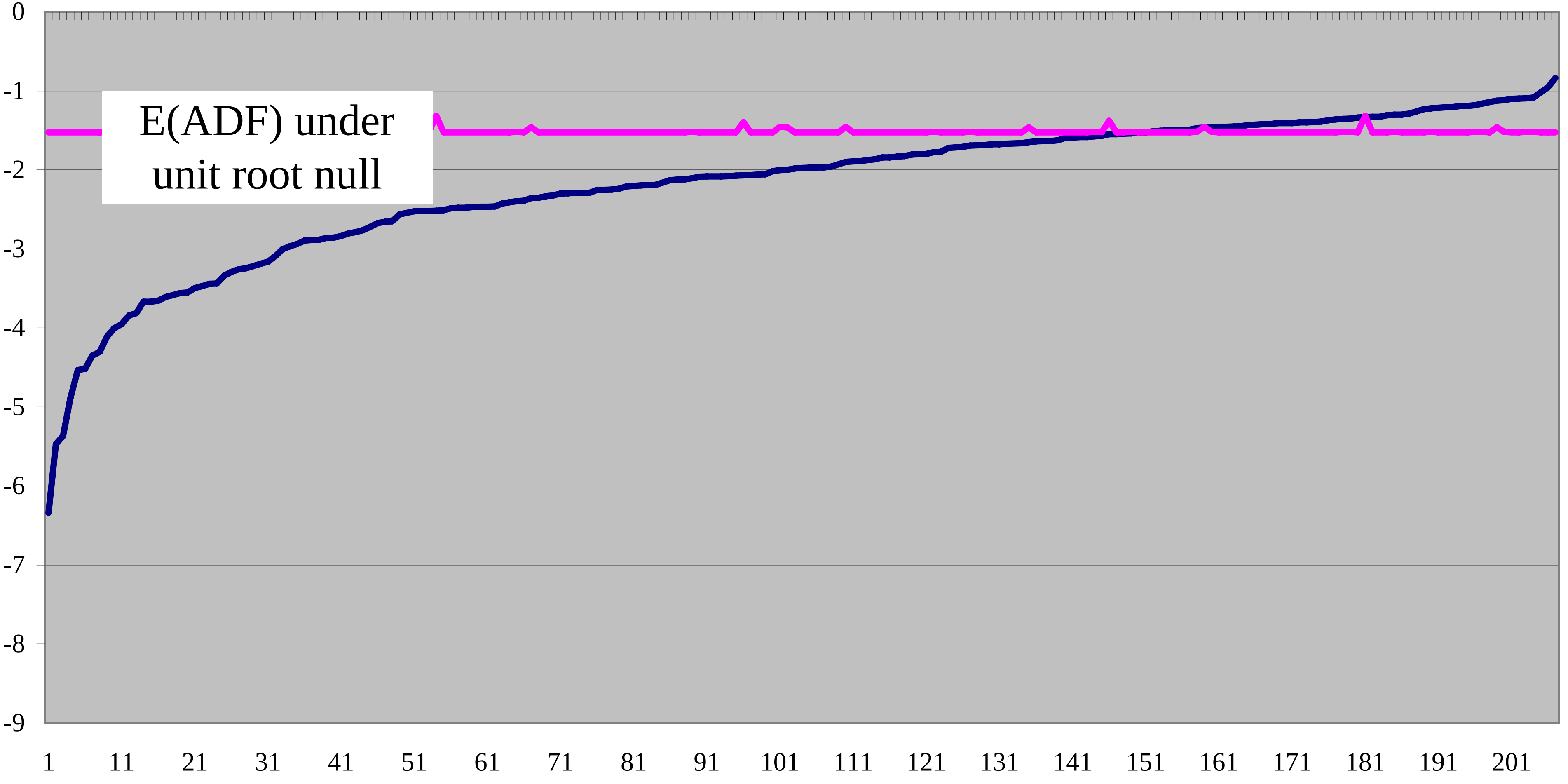

Figure A3
Alternative V Factor Estimates

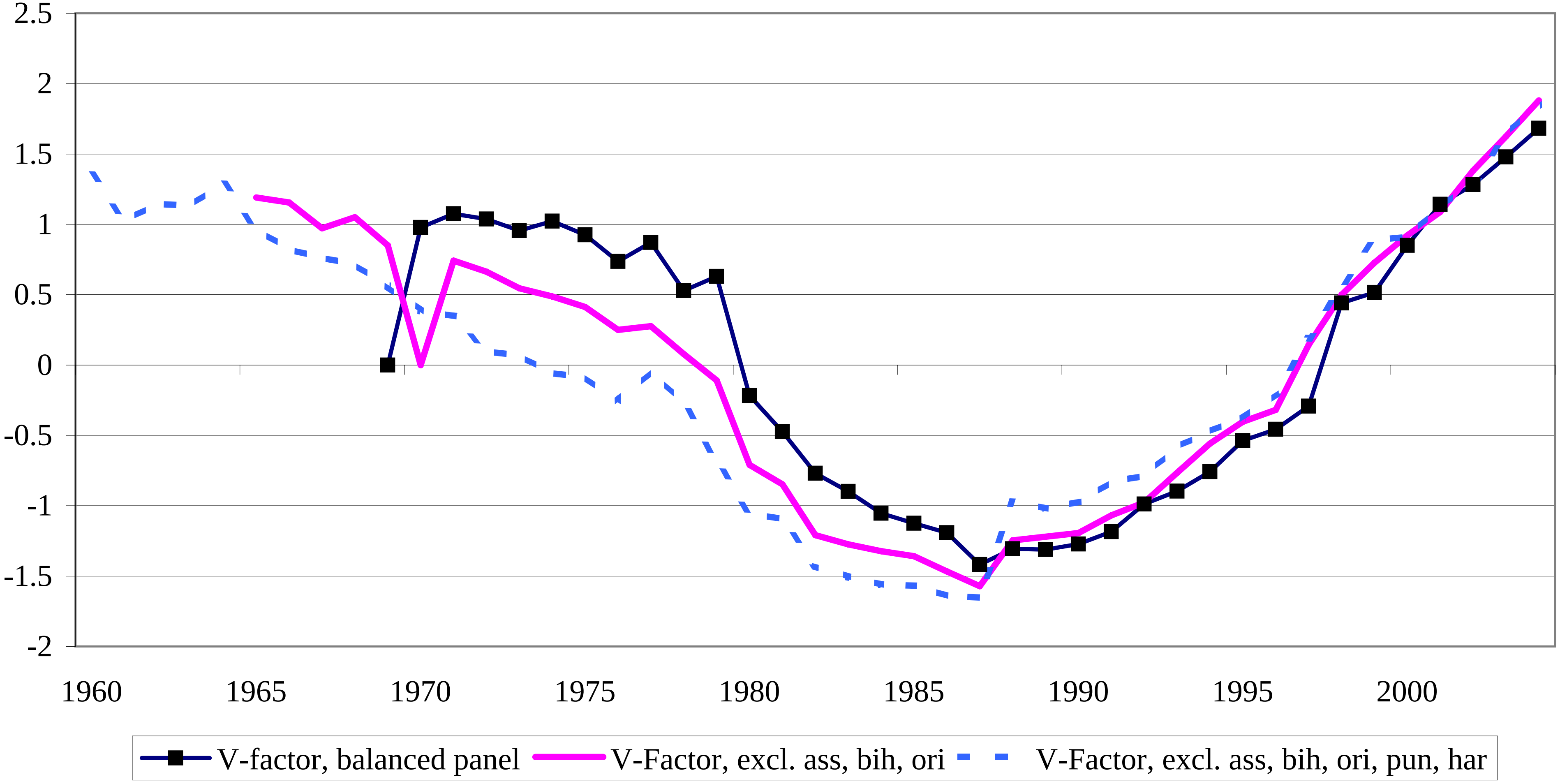

Figure A4. Impact on estimated V-Factors of Excluding Series from Panel

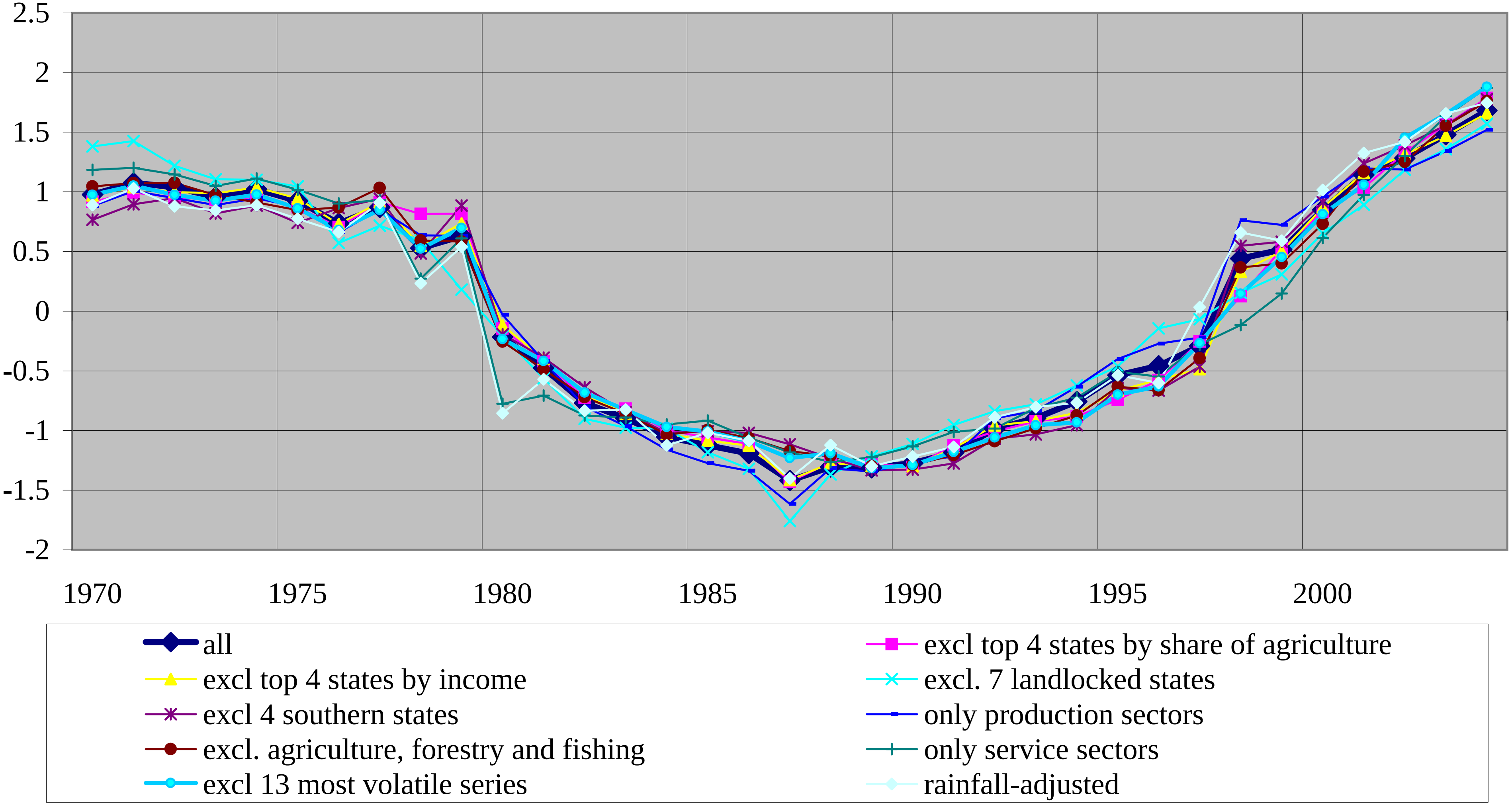

Table A1. Unit Root Tests

	Underlying series	Transitory Components from estimation in…		
		Levels		Differences
		1 Factor	2 Factors	1 Factor
Im *et al* Panel Unit Root Test (*p*-values)	1.000	0.000	0.000	0.000
Proportion of individual ADF tests below mean under unit root null	53%	75%	97%	73%

Table G1. Estimating common breakpoints by principal components: some simulation results

				Levels Approach			Differences Approach		
Start year	N	break point	stationary factors? ("1"=yes)	s.d.	bias	% correct +or- 1 year	s.d.	bias	% correct +or- 1 year
1970	207	1987	1	2.7	-1.0	60%	5.7	-1.0	32%
1960	**139**	1987	1	2.2	0.1	74%	8.0	1.6	24%
1970	207	1987	**0**	1.4	-0.9	72%	2.2	-1.2	64%
1970	207	**1979**	1	5.3	-4.3	30%	6.8	-4.3	26%
1960	**139**	**1979**	1	3.5	-1.9	55%	6.5	-1.2	33%
1970	207	**random**	1	2.8	-0.8	64%	6.3	-0.2	32%
1960	**139**	**random**	1	2.5	0.4	69%	8.3	2.0	34%

Table G2 Systematic properties of estimated breakpoints when the true breakpoint is a random variable

	Levels Approach	Differences Approach
Correlation with true breakpoint	0.716611341	0.289595116
Slope coefficient on true breakpoint	0.8528597	0.166217551