DEAKIN
UNIVERSITY AUSTRALIA

Faculty of Business and Law
School of Accounting, Economics and Finance

School Working Paper

Financial Econometrics Series

SWP 2014/15

# A practical Note on the Determination of the Number of Factors Using Information Criteria with Data-Driven Penalty

J. Westerlund and S. Mishra

The working papers are a series of manuscripts in their draft form. Please do not quote without obtaining the author's consent as these works are in their draft form. The views expressed in this paper are those of the author and not necessarily endorsed by the School or *IBIS*World Pty Ltd.

# A Practical Note on the Determination of the Number of Factors using Information Criteria with Data-Driven Penalty*

Joakim Westerlund
Deakin University
Australia

Sagarika Mishra
Deakin University
Australia

February 3, 2014

**Abstract**

As is well known, when using an information criterion to select the number of common factors in factor models the appropriate penalty is generally indetermine in the sense that it can be scaled by an arbitrary constant, $c$ say, without affecting consistency. In an influential paper, Hallin and Liška (Determining the Number of Factors in the General Dynamic Factor Model, *Journal of the American Statistical Association* **102**, 603–617, 2007) proposes a data-driven procedure for selecting the appropriate value of $c$. However, by removing one source of indeterminacy, the new procedure simultaneously creates several new ones, which make for rather complicated implementation, a problem that has been largely overlooked in the literature. By providing an extensive analysis using both simulated and real data, the current paper fills this gap.

**JEL Classification:** C12; C13; C33.

**Keywords:** Panel data; Common factor model; Information criterion; Data-driven penalty.

## 1 Introduction

Consider the following common factor model:

$$X_{i,t} = \lambda_i' F_t + e_{i,t}, \tag{1}$$

where $t = 1, ..., T$ and $i = 1, ..., N$ index the time series and cross-section units, respectively, $F_t$ is an $r_0 \times 1$ vector of common factors, $\lambda_i$ is a conformable vector of loading coefficients,

*The authors would like to thank Simon Svensson for valuable input.

and $e_{i,t}$ is idiosyncratic. Models of this type has recently become very popular with a huge number of applications, especially in macroeconomics and financial economics. An important question in this literature is how many factors there are, and several information criteria have been proposed as solutions (see, for example, Bai and Ng, 2002; Hallin and Liška, 2007). As with other information criteria (IC), these include a penalty term for overfitting. For the IC to be consistent, it is crucial that the penalty term goes to zero at an appropriate rate as $N,\ T \to \infty$. But the penalty is not uniquely determined by theory and can be scaled arbitrarily without affecting consistency. That is, if an IC is consistent for penalty $p(N,T)$, then it is consistent also for penalty $p_c(N,T) = cp(N,T)$, where $c > 0$ is a constant tuning parameter (see Hallin and Liška, 2007). This degree of freedom creates a problem in finite samples; how should the researcher scale the penalty given the sample at hand? Although consistent for any $c$, in small samples the choice is not irrelevant, as a larger (smaller) penalty will lead to a smaller (larger) number of estimated factors.

Recognizing this problem, Hallin and Liška (2007) recently proposed a data-driven procedure for determining $c$, henceforth referred to as "HL". The procedure works as follows.[1] The researcher is supposed to suggest several possible values of $c$, and create several differently sized subsamples by removing observations from the original sample. For each candidate value of $c$, the researcher then estimates the number of factors in each subsample, using the IC with the penalty scaled by the specific $c$. This gives one estimate of $r$ for each ($c$, subsample) pair. The HL choice of $c$ is the value corresponding to the lowest variance of the estimated values of $r$ across subsamples. The idea is that only when $c$ is appropriately set should the IC tend to estimate the same number of factors regardless of the sample size (provided that it is not "too small").

While the above indeterminacy of IC is an old and well-known problem, as far as we are aware, HL is the first proposal to that end that is data-driven. It has therefore received much attention in the literature. The overall impression is the procedure works well and that it leads to improved precision in the estimation of $r$ relative to the naive approach of just setting $c = 1$. Indeed, as Hallin and Liška (2007, page 614) conclude: "We have shown how to take advantage of the fact that penalty functions in [IC] defined up to a multiplicative constant. We evaluated the performance of [HL] through simulation and found that it outperforms

[1] Note that here "HL" refers to the Hallin and Liška (2007) procedure for selecting $c$ (see their Section 4) and not the IC that they propose (see their Section 3).

existing methods." Alessi et al. (2010, page 1812) reach a similar conclusion; "The appeal of [HL] is threefold: (i) it builds on a well known criterion, whose theoretical properties have been proved and are preserved; (ii) it improves the finite sample performance of the original criterion; (iii) it is easy to implement."

However, while appealing and certainly promising, HL suffers from one major drawback that seems to have been largely overlooked in the previous literature, namely, that by removing one source of indeterminacy, HL simultaneously creates several new ones. Specifically, the implementation of HL necessitates three choices that are likely to be very important for performance; (C1) the maximum number of factors, (C2) the appropriate subsamples, and (C3) the candidate values of $c$. As an example, consider (C3). If the grid considered for $c$ is "too rough", the procedure is likely to miss the variance-minimizing value of $c$, whereas if the grid is "too fine", then this will cause increased random variation making the variance-minimizing value of $c$ more difficult to discern. It is therefore important that the grid is set just right. Unfortunately, the Monte Carlo studies of Hallin and Liška (2007) and Alessi et al. (2010) do not provide any guidance in this regard. The current paper can be seen as a reaction to this. The purpose is to offer applied researchers some simple rules of thumb that ensure good performance across a range of data generating processes (DGPs). This is done within the large class of IC proposed by Bai and Ng (2002), which are among the most popular IC by far. We also show, using as an illustration the commonly used data set of Stock and Watson (2005), how, if properly implemented, HL can lead to improved inference. In particular, in contrast to existing results suggesting the presence of six to 10 factors (see, for example, Alessi et al., 2010; Breitung and Eickmeier, 2011; Stock and Watson, 2005), we find only one factor (with a potentially break in variance).

# 2 Determining $r_0$

## 2.1 Background

A large number of IC exists; however, the IC of Bai and Ng (2002) are most popular, and in this paper we therefore only consider these IC. The estimation approach on which the IC of Bai and Ng (2002) are based is principal components (PC) (see, for example, Bai, 2003).[2] The PC estimator of $F = (F_1, ..., F_T)'$ based on $r$ factors, denoted $\hat{F}^r = (\hat{F}_1^r, ..., \hat{F}_T^r)'$, is $\sqrt{T}$ times

[2]See Pesaran (2006) for an alternative approach based on the cross-section averages of the observed data.

the matrix consisting of the eigenvectors corresponding to the $r$ largest eigenvalues of the $T \times T$ matrix $XX'$, where $X = (X_1, ..., X_T)'$ is $T \times N$ and $X_t = (X_{1,t}, ..., X_{N,t})'$ is $N \times 1$. The corresponding matrix of estimated loadings is given by $(\hat{\lambda}^r)' = (\hat{\lambda}_1^r, ..., \hat{\lambda}_N^r) = T^{-1}(\hat{F}^r)'X$. The resulting sum of squared residuals (divided by $NT$) is given by

$$V_{N,T}(r) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{i,t} - (\hat{\lambda}_i^r)' \hat{F}_t^r)^2. \tag{2}$$

A typical IC for selecting $r$ takes the form

$$IC_{N,T}(r) = V_{N,T}(r) + rp(N,T), \tag{3}$$

where the penalty satisfies the following two conditions: (i) $p(N,T) \to 0$; (ii) $\min\{N,T\}$ $p(N,T) \to \infty$ (see Bai and Ng, 2002, Theorem 2). These conditions are needed to ensure that the penalty vanishes at an appropriate rate such that asymptotically under- and overparameterized models are not chosen. For example, if $p(N,T) = \ln(\min\{N,T\}) / \min\{N,T\}$, then (ii) is obviously satisfied, and since $\ln(x) \leq x$ for any $x \geq 1$, so is (i). Unfortunately, theory is not more precise than this. Therefore, if $p(N,T)$ satisfies (i) and (ii), then so does $p_c(N,T)$, and there is no way of discriminating between the two.

For a given choice of penalty, the estimator of $r_0$ is given by

$$\hat{r}_{N,T} = \arg \min_{0 \leq r \leq r_{max}} IC_{N,T}(r), \tag{4}$$

where $r_0 \leq r_{rmax}$ ($r_{rmax} < \min\{N,T\}$). According to Bai and Ng (2002, Theorem 2), under conditions (i), (ii) and regularity, $P(\hat{r}_{N,T} = r_0) \to 1$ as $N$, $T \to \infty$. Therefore, whether the penalty is $p(N,T)$ or $p_c(N,T)$ has no effect on the consistency of $\hat{r}_{N,T}$.

## 2.2 The HL procedure

Ideally, if two penalties have the same rate of convergence, in small samples we should choose the one leading to the most parsimonious model. Indeed, for a given sample size, $p_c(N,T)$ can be made arbitrarily small or large relative to $p(N,T)$ by just varying the value taken by $c$.

The question is how to select $c$ so as to ensure accurate estimation of $r_0$? While setting $c$ "too small" leads to underpenalization and therefore overestimation of $r_0$, setting $c$ "too large" leads to overpenalization and underestimation of $r_0$. It is therefore important that $c$ set in between these extremes. This is the idea of Hallin and Liŝka (2007). To formally

describe their procedure, let us denote by $\hat{r}_{c,N,T}$ the minimizer of $IC_{c,N,T}(r) = V_{N,T}(r) + rp_c(N,T)$. Let $\hat{r}_{c,N_1,T_1}, ..., \hat{r}_{c,N_J,T_J}$ be the estimates obtained from considering subsamples of size $(N_1, T_1), ..., (N_J, T_J)$, where $0 < N_1 < ... < N_J = N$ and $0 < T_1 \leq ... \leq T_J = T$. The variation of these estimates is key:

- If $c > 0$ is "too small", then $j \mapsto \hat{r}_{c,N_j,T_j}$ will increase until it reaches $r_{max}$;
- If $c$ is "too large", then $j \mapsto \hat{r}_{c,N_j,T_j}$ is decreasing towards zero;
- Since $c \mapsto \hat{r}_{c,N_j,T_j}$ is nonincreasing, this means that somewhere between these extreme values of $c$ there will be a range for which $j \mapsto \hat{r}_{c,N_j,T_j} \approx r_0$ is stable.

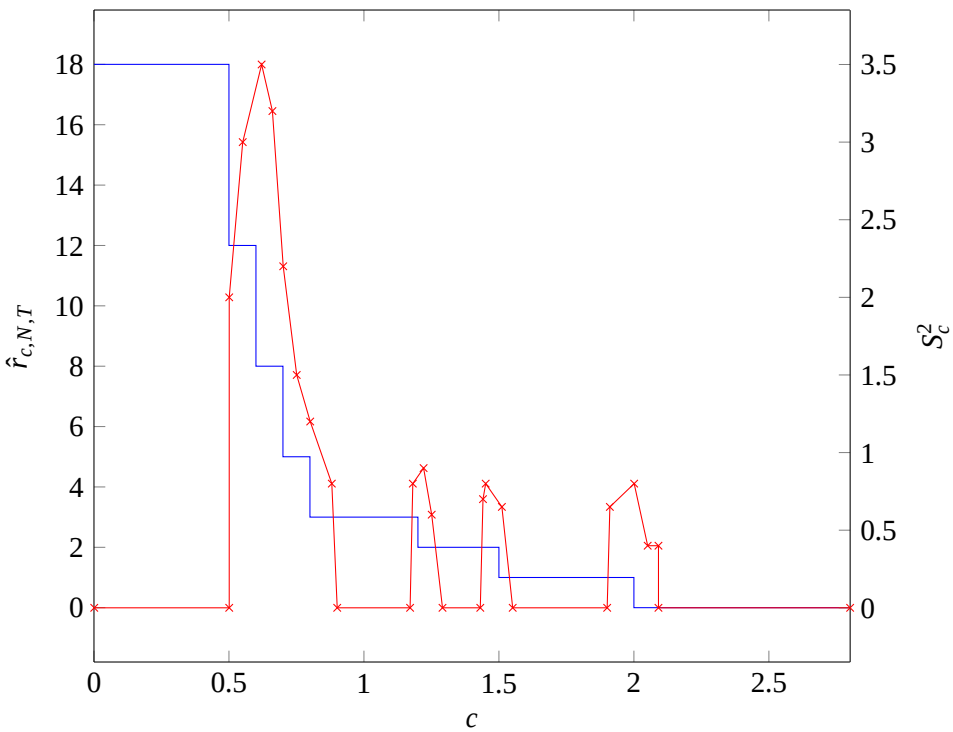
Thus, the main difference between, on the one hand, the extremes and, on the other hand, the mid-range values of $c$ is the variability in $j \mapsto \hat{r}_{c,N_j,T_j}$; the variability should be relatively low whenever $r_{c,N_j,T_j}$ is close to $r_0$. As a measure of this we look at the variance of $\hat{r}_{c,N_1,T_1}, ..., \hat{r}_{c,N_J,T_J}$,

$$S_c^2 = \frac{1}{J}\sum_{j=1}^{J}\left(\hat{r}_{c,N_j,T_j} - \frac{1}{J}\sum_{i=1}^{J}\hat{r}_{c,N_i,T_i}\right)^2. \tag{5}$$

Suppose that $c \in C \subset [c_{min}, c_{max}]$. Values of $c$ corresponding to $S_c^2 = 0$ are promising candidates. But we can expect there to be several values of $c$ for which $S_c^2 = 0$. In particular, if $c = 0$, then the penalty will be zero and therefore $\hat{r}_{c,N_1,T_1} = ... = \hat{r}_{c,N_J,T_J} = r_{max}$, implying $S_c^2 = 0$. Hallin and Liŝka (2007) therefore suggest using $\hat{r}_{\hat{c},N,T}$ as an estimator of $r_0$, where $\hat{c}$ belongs to the second interval for which $S_c^2 = 0$. A typical situation is illustrated in Figure 1 where $r_0 = 3$ and $r_{max} = 18$. There are two lines; while the marked red line depicts $c \mapsto S_c^2$, the solid blue line depicts $c \mapsto \hat{r}_{c,N,T}$. In this example, the second interval with $S_c^2 = 0$ corresponds to $\hat{r}_{\hat{c},N,T} = 3$, where $\hat{c} \in [0.9, 1.1]$.

**Remark 1.** It should be pointed out that the above discussion supposes finite values of $N$ and $T$. If $N$, $T \to \infty$, unless $c = 0$, $P(\hat{r}_{c,N,T} = r_0) \to 1$. Hence, while in finite samples there may be a tendency for $\hat{r}_{N,T}$ to wander away from $r_0$, as the sample size increases, it will eventually come back and tend towards $r_0$. Since $P(\hat{r}_{c_{min},N,T} \leq \hat{r}_{\hat{c},N,T} \leq \hat{r}_{c_{max},N,T}) = 1$ for all $N$ and $T$, this implies that the HL procedure does not affect consistency.

Figure 1: Graphical illustration of the HL procedure when $r = 3$.




*Note*: The marked red and solid blue lines represent $S_c^2$ and $\hat{r}_{c,N,T}$, respectively.

## 2.3 Implementation issues

While the HL procedure "resolves" the indeterminacy caused by $c$, it also introduces several tuning parameters, and it is not obvious how these should be set. In particular, for a given penalty $p(N, T)$ the researcher is required to make three choices;

C1. the maximum number of factors to consider, $r_{max}$,

C2. the appropriate subsamples $(N_1, T_1), ..., (N_J, T_J)$, including the minimum sample size, $(N_1, T_1)$, and

C3. the candidate values of $c$.

Hallin and Liŝka (2007), and Alessi et al. (2010) report some Monte Carlo results for the HL procedure when applied to different IC. Unfortunately, these results do not offer any guidance regarding the above tuning parameter choices. In fact, as we allude to in the below, the values opted for in these two studies are not even close to one another.

Consider C1. In time series, rules such as $8\lfloor(T/100)^{1/4}\rfloor$, where $\lfloor x\rfloor$ is the integer part of $x$, are often used to select the maximum dimension of a model (see, for example, Ng and Perron, 1995); unfortunately, no such rules are available for the current problem. One possibility is to replace $T$ by $\min\{N,T\}$ and use $8\lfloor(\min\{N,T\}/100)^{1/4}\rfloor$ as a maximum, which would then have the same asymptotic justification as in the time series case (Bai and Ng, 2002); however, this is typically not done in the literature. Instead, since $r_0$ is likely to be a fairly small number, it is common to just fix $r_{max}$ to an arbitrary number. For example, while Hallin and Liŝka (2007) set $r_{max} = 19$, Alessi et al. (2010) set $r_{max} = 10$.

The subsamples to consider (C2) are likely to be key in determining the small-sample performance of the procedure. As with C1, the existing results are not enough to gauge the decision. In particular, while Hallin and Liŝka (2007) set $N_j = 120 + 10j$ and $T_j = 70 + 10j$, Alessi et al. (2010) set $N_j = \lfloor 3N/4\rfloor + j - 1$ with $T$ held fixed. Intuitively, varying $N$ is more important than varying $T$. The reason is that the consistency of $\hat{r}_{N,T}$ ($\hat{r}_{c,N,T}$) is based on the fact that if there are $r_0$ factors, then the $r_0$ largest eigenvalues of the $N \times N$ covariance matrix $\Sigma_X = E(X_t X_t')$ are diverging as $N \to \infty$. Hence, if $T$ is "small" (relative to $N$), one may consider subsampling $N$ only, keeping $T$ fixed (see Hallin and Liŝka, 2007). The main purpose of the subsampling is therefore to capture the effect of an increase in the sample size, and in view of this the exact size of the subsample increments, $(N_j - N_{j-1})$ and $(T_j - T_{j-1})$, are maybe not that important. The minimum sample size, $(N_1, T_1)$, on the other hand, is likely to be quite important; if $(N_1, T_1)$ is "too small", then the variation in $\hat{r}_{c,N_1,T_1}, ..., \hat{r}_{c,N_J,T_J}$ may not reflect divergence (but rather poor small-sample performance), whereas if $(N_1, T_1)$ is "too large", then the variation might not be enough to identify the second stability interval.

Where the subsample increments are likely to be relatively unimportant, the grid considered for $c$ (C3) is key. In particular, the grid should not be too rough, as this will lead to discretization bias. Consider again the example in Figure 1. If $c \in \{0.5, 1, ..., 3\}$, then the procedure will be unable to detect the first instability interval, and therefore the second stability interval picked up by the procedure will not be the "correct" one. Specifically, the stability interval is likely to come later had the grid been finer. On the other hand, a too fine grid will not only make the procedure unnecessarily slow, but will also induce unnecessary variation in the results. In Hallin and Liŝka (2007), and Alessi et al. (2010), $c \in \{0.01, 0.02, ..., c_{max}\}$, where $c_{max} = 3$ in the former study and $c_{max} = 5$ in the latter.

In sum, while promising, by eliminating the indeterminacy caused by $c$, the HL procedure also introduces tuning parameter choices for which there are presently no guidelines. The above discussion gives a rough indication of what is important and what is not; however, in practice this is not likely to be enough. In Section 3 we therefore use Monte Carlo simulation as a means to shed some light on the issue.

# 3 Monte Carlo simulations

The DGP is very similar to the ones considered by Alessi et al. (2010), Hallin and Liŝka (2007), and Bai and Ng (2002), and is given by (**??**), where

$$e_{i,t} = \rho e_{i,t-1} + \varepsilon_{i,t} + \sum_{k=i+1}^{K} \beta(\varepsilon_{i-k,t} + \varepsilon_{i+k,t}), \tag{6}$$

with $e_{1,0} = ... = e_{N,0} = 0$, $K = 10$, $(F_t', \lambda_i', \varepsilon_{i,t})' \sim N(0, \text{diag}(I_{2r_0}, \sigma_\varepsilon^2))$, and $r_0 = 3$. A large number of DGPs were considered. However, in interest of space, here we only report the results for the following DGPs that differ depending on whether $e_{i,t}$ is serial- and cross-correlation free, serially correlated, or cross-correlated:

DGP1. $\rho = \beta = 0$ and $\sigma_\varepsilon^2 = 1$;

DGP2. $\rho = \beta = 0$ and $\sigma_\varepsilon^2 = 4$;

DGP3. $\rho = 0.5$, $\beta = 0$ and $\sigma_\varepsilon^2 = 1$;

DGP4. $\rho = 0$, $\beta = 0.1$ and $\sigma_\varepsilon^2 = 1$.

For simplicity, we set $c \in C = \{\Delta c, 2\Delta c, ..., c_{max}\}$, such that the number of candidate values for $c$ is $c_{max} / \Delta c$. Hence, by setting $\Delta c$, we can control the fineness of the grid considered for $c$. We also set $(N_j - N_{j-1}) = \Delta N$ and $(T_j - T_{j-1}) = \Delta T$ for all $j$, such that by setting $(N_1, T_1)$ and $J$ we can control $\Delta N = (N - N_1) / J$ and $\Delta T = (T - T_1) / J$. The number of replications is set to 500.

The specific IC considered is motivated by the results Bai and Ng (2002), who show that $IC_{p1}$, $IC_{p2}$, $PC_{p1}$ and $PC_{p2}$ tend to perform best. We have considered all four IC. However, in agreement with the results of Bai and Ng (2002) and Alessi et al. (2010), we find that $IC_{p1}$ and $IC_{p2}$ ($PC_{p1}$ and $PC_{p2}$) tend to perform very similarly, and we therefore only report the results for $IC_{p1} = \ln[V_{N,T}(r)] + rp_{IC_{p1}}$ and $PC_{p1} = V_{N,T}(r) + rp_{PC_{p1}}$, where $p_{IC_{p1}} =$

$(N+T)\ln[NT/(N+T)]/(NT)$ and $p_{PC_{p1}} = V_{N,T}(r_{max})p_{IC_{p1}}$. The HL versions of these IC are denoted in an obvious notation as $IC_{c,p1}$ and $PC_{c,p1}$, respectively. For simplicity, the dependence of $IC_{p1}$, $PC_{p1}$, $IC_{c,p1}$ and $PC_{c,p1}$ on $N$, $T$ and $r$ is henceforth suppressed.

### 3.1 C1

Foreshadowing the results we obtain in Section C2, here we set $(N_1, T_1) = 0.5(N, T)$ and $J = 6$. Also, as demonstrated in Section C3, while $IC_{p1}$ tend to perform best when $c_{max} = 2$, for $PC_{p1}$ the best performance is obtained by setting $c_{max} = 6$. In this section, we therefore only consider these specifications. In both cases, $\Delta c = 0.1$ and $r_0 = 3$. Three different specifications of $r_{max}$ are considered; (i) $r_{max} = 2 < r_0 = 3$, (ii) $r_{max} = r_0 = 3$ and (iii) $r_{max} = 8\lfloor(\min\{N,T\}/100)^{1/4}\rfloor > r_0$.

The results are reported in Table 1. The first thing to note is that, provided that $r_{max} > r_0$, $IC_{p1}$ and $PC_{p1}$ tend to be dominated by $IC_{c,p1}$ and $PC_{c,p1}$, respectively. Note in particular how $PC_{p1}$ always overestimates $r_0$. HL therefore leads to IC that are relatively less sensitive to the choice of $r_{max} \geq r_0$, which is, of course, a great advantage in practice. The exception is when $r_{max} = r_0$, in which case $IC_{p1}$ and $PC_{p1}$ tend to perform best. We also see that there are some instances when $r_{max} = r_0$ leads to worse performance when compared to $r_{max} = 8\lfloor(\min\{N,T\}/100)^{1/4}\rfloor$. Hence, in order to minimize the risk of underestimation, it seems important not to set $r_{max}$ "too close" to $r_0$. As expected, all criteria perform poorly when $r_{max} < r_0$.

### 3.2 C2

In this section we focus on the choice of subsamples. The choice has three dimensions; (a) the number of subsamples, $J$, (b) whether the subsampling is should be over $N$, $T$, or both, and (c) how to set the minimum sample, $(N_1, T_1)$. In view of the results of Section C1, we set $\Delta c = 0.1$, $r_0 = 3$ and $r_{max} = 8\lfloor(\min\{N,T\}/100)^{1/4}\rfloor > r_0$. Also, the results for $IC_{c,p1}$ ($PC_{c,p1}$) are based on setting $c_{max} = 2$ ($c_{max} = 6$). The results for the case when $N = T = 60$ are reported in Tables 2 and 3, and may be summarized as follows.

(a) While the results when $N_1 = 0.15N$ and/or $T_1 = 0.15T$ (Table 2) are almost completely flat in $J$, when $N_1 = 0.5N$ and/or $T_1 = 0.5T$ (Table 3) there is a slight improvement as $J$ increases. Hence, while increasing $J$ may lead to better performance, the potential

improvement is rather marginal. The exception is $PC_{c,p1}$ in DGP3 when subsampling $N$, in which case the improvement is quite substantial.

(b) Contrary to intuition, there is no direct advantage to subsampling both $N$ and $T$ when compared to subsampling only one of the two dimensions. In fact, the best results are generally obtained by subsampling $T$ only, although the gain in performance is typically quite marginal. The fact that subsampling $T$ seems to work better than subsampling $N$ is unexpected given the discussion of Section 2.3.

(c) The appropriate minimum sample to use depend heavily on the DGP considered. Specifically, while in DGP1 and DGP3 the results do not vary much depending on the size of the minimum subsample, in DGP2 and DGP4 there is substantial variation. While in DGP2 the best performance is obtained by setting $N_1 = 0.15N$ and/or $T_1 = 0.15T$, in DGP4 the best performance is obtained by setting $N_1 = 0.5N$ and/or $T_1 = 0.5T$. Hence, unless we are in DGP4, it is not obvious that a larger minimum subsample leads to better performance.

The best performance across the four DGPs is obtained by setting $(N_1, T_1)$ and $J$ either relatively large or relatively small. Thus, by having fewer and larger subsamples we can compensate for a small minimum. However, given the potential adverse consequences of using a too small sample, the opposite strategy seems preferable.

Looking across $IC_{c,p1}$ and $PC_{c,p1}$ we see that the best performance is generally obtained by using $IC_{c,p1}$, which in turn tend to outperform original $IC_{p1}$. The exception is in DGP2 when $N_1 = 0.15N$ and/or $T_1 = 0.15T$. The worst performance is obtained by using original $PC_{p1}$, which consistently overestimates $r_0$. Thus, among the two types of IC, HL matter most when using $PC_{p1}/PC_{c,p1}$.

The results reported in Tables 2 and 3 are for $N = T = 60$; however, the conclusions remain the same even considering samples of different size.

### 3.3 C3

In this case we set $(N_1, T_1) = 0.5(N, T)$, $J = 6$ and $r_{max} = 8\lfloor(\min\{N, T\}/100)^{1/4}\rfloor$, which allows us to focus on $\Delta c$ and $c_{max}$. As Bai and Ng (2002) show, $IC_{p1}$ and $PC_{p1}$ have very different small-sample properties. In particular, while the former tends to underestimate $r_0$, the latter tends to overestimate $r_0$, suggesting that $c_{max}$ is going to be important. We begin

by considering $IC_{p1}$ and $IC_{c,p1}$. The results for $N = T = 40$ and $N = T = 60$ are reported in Tables 4 and 5, respectively. The information content of these tables may be summarized as follows.

- Unless $\Delta c \geq 0.5$, the value of $\Delta c$ has almost no effect on the results. This is important, because the value of $\Delta c$ has a significant impact on the overall speed of the HL procedure. Hence, one would not want to set $\Delta c$ "too small". Alessi et al. (2010) and Hallin and Liŝka (2007) set $\Delta c = 0.01$, which, in view of our results, is not necessary.

- $c_{max}$ can have a quite substantial effect on the performance of the procedure. In particular, it seems important not to set $c_{max}$ "too large". The value 2 seems to work very well. This is true not only for the cases covered by Tables 4 and 5, but also for all unreported experiments we ran. In particular, it holds also when $r_0 \neq 3$. The reason for the good performance when $c_{max} = 2$ can be inferred by looking at the results for $\hat{r}$ and $\hat{c}$, suggesting that the main concern is to not over-penalize, which is more likely to happen if $c_{max}$ is relatively large.[3]

- Unless $\Delta c = 1$, $IC_{c,p1}$ leads to relatively good performance when compared to $IC_{p1}$ in the practically relevant DGP3 and DGP4 when $e_{i,t}$ is serially and cross-sectionally correlated, respectively. The worst performance is obtained in DGP2, which is not totally unexpected as a larger value of $\sigma_\varepsilon^2$ makes $r_0$ more difficult to discern.

- As $N$ and $T$ increases the performances of $IC_{p1}$ and $IC_{c,p1}$ are equalized. The exception (for $\Delta c \leq 0.1$) is DGP4 where $IC_{p1}$ does not perform at all. There is also a tendency for $IC_{c,p1}$ based on $c_{max} \in \{4, 6\}$ to gain relative to the same IC based on $c_{max} = 2$.

The results for $PC_{p1}$ and $PC_{c,p1}$ are reported in Tables 6 ($N = T = 40$) and 7 ($N = T = 60$). While the results are very different from those reported in Tables 4 and 5 for $IC_{p1}$ and $IC_{c,p1}$, the conclusions regarding the candidate values of $c$ are very similar. The exception is $c_{max}$; while previously setting $c_{max} = 2$ led to the best performance, now this specification leads to worst performance. This is partially expected, because, as already mentioned, while $IC_{p1}$ tends to tends to underestimate $r_0$, $PC_{p1}$ tends to overestimate $r_0$. $PC_{c,p1}$ therefore requires a relatively large penalty. This, in turn, makes the procedure less sensitive to the value of $\Delta c$;

[3]This is in agreement with the results of Bai and Ng (2002) showing how $IC_{p1}$ tend to underestimate $r_0$ when $N$ and / or $T$ are "small".

however, $\Delta c \in \{0.05, 0.1\}$ still leads to the pest performance. The overall best performance (especially when $N = T = 60$) is obtained by setting $c_{max} = 6$ and $\Delta c \in \{0.05, 0.1\}$.

# 4 Empirical illustration

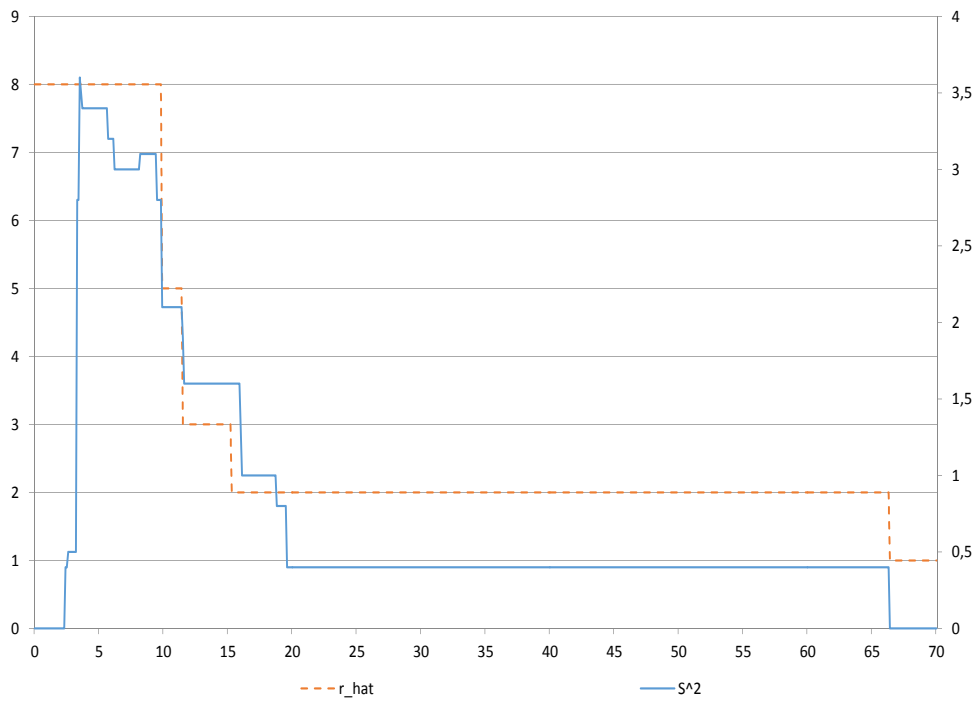
In this section we illustrate the above results using as an example the Stock and Watson (2005) dataset, which has attracted considerable interest in the common factor literature.[4] A very incomplete list of users of this data set include Breitung and Eickmeier (2011), Alessi et al. (2010), and Hallin and Liŝka (2007). The dataset has also been revised and extended in several directions (see, for example, Stock and Watson, 2008). In all these papers the appropriate number of common factors to use has been a key issue. Different approaches have been used; however, most studies estimate between six and 10 factors, which, as Stock and Watson (2005) note themselves in their Section 4.3 called "Why So Many Factors?", seems excessive given the estimates reported in the bulk of the empirical factor model literature. Our empirical illustration is motivated by this observation. The purpose is to illustrate the importance of C1–C3 in applied work.

The data comprise 132 monthly macroeconomic indicators of the US economy from January 1960 to December 2003. The analysis requires some pre-treatment of the data. We proceed exactly as in Stock and Watson (2005). Non-stationary raw data (which were already available to us in seasonally adjusted form) are differenced until they are stationary. We also remove outliers and normalize the series to have means of zero and unit variances. The reader is referred to Stock and Watson (2005) for details on the treatment of the data.

In view of the Monte Carlo results reported in Section 3, suggesting that $IC_{c,p1}$ generally outperforms $PC_{c,p1}$, in this section we only consider $IC_{p1}$ and $IC_{c,p1}$, which are implemented using the following tuning parameter values: $J = 6$, $(N_1, T_1) = 0.5(N, T)$, $r_{max} = 8\lfloor(\min\{N, T\}/100)^{1/4}\rfloor = 8$ and $\Delta c = 0.1$. As expected given the results reported in Section 3, the value of $c_{max}$ turned out to be very important. In particular, no matter how we set $r_{max}$, $IC_{p1}$ and $IC_{c,p1}$ always ended up estimating the maximum, which is not an uncommon finding. With $IC_{p1}$ this is rather problematic in the sense that there is not much else to do but to pick a relatively large number and hope that this will eliminate the effects of the cross-correlation (as is done in, for example, Stock and Watson, 2005). With $IC_{c,p1}$, however,

[4] The data are available at http://www.princeton.edu/~mwatson/wp.html.

Figure 2: Empirical results for $IC_{c,p1}$.

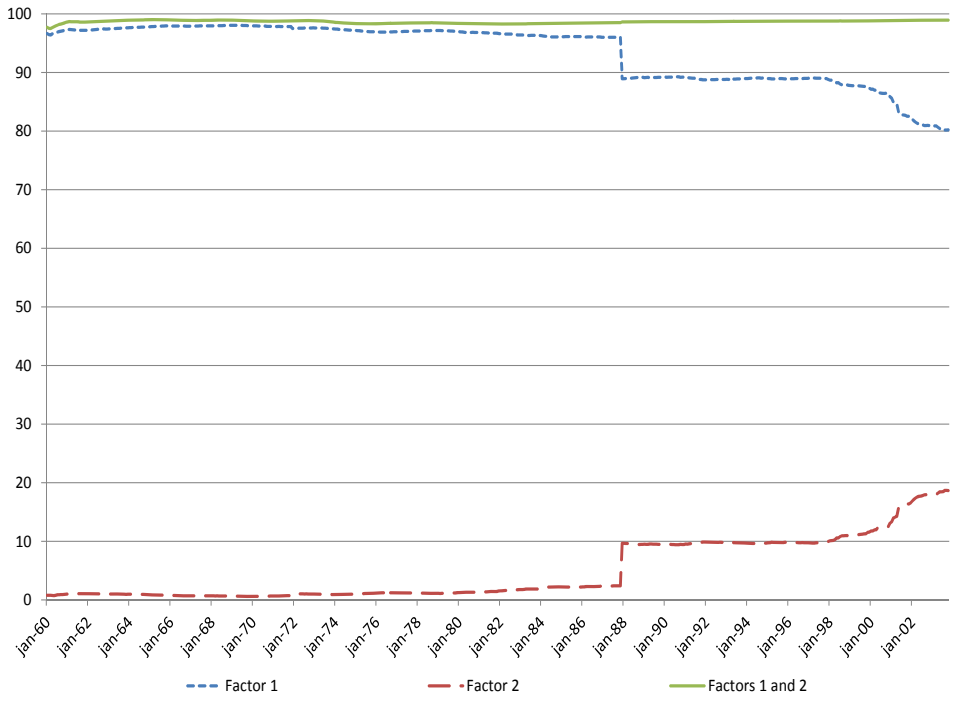



*Note*: The horizontal, left vertical and right vertical axes display $c$, $\hat{r}_{c,N,T}$ and $S_c^2$, respectively.

there is more information to be had. In our case, for a wide range of values of $c_{max}$, we ended up estimating $\hat{c} = c_{max}$, and for this value of $\hat{c}$ we obtained $\hat{r}_{c,N,T} = r_{max}$. This suggests that the penalty is too low and therefore that $c_{max}$ should be increased. However, as we explain in Section 3.3, $c_{max}$ must also not be set too high. We therefore adopted a sequential procedure, in which $c_{max}$ was increased by a small amount until the second stability interval is reached.

In Figure 2 we again plot $S_c^2$ and $\hat{r}_{c,N,T}$ as a function of $c$. The second stability interval occurs very late, at $c = 66.4$, and for this value of $c$ we estimate $\hat{r}_{\hat{c},N,T} = 1$ factor. This is partly in agreement with the full sample results of Hallin and Liŝka (2007), suggesting the presence of either one or four factors, but it is inconsistent with the six to 10 factors found in many other studies. Consider as an example the study of Alessi et al. (2010), who report six factors. While in the empirical part of their paper the authors do not provide any implementation details, it seems reasonable to assume that they proceed in the same way as in the Monte Carlo part, in which $c_{max} = 5$. In view of Figure 1 such a value of $c_{max}$ is clearly not enough. In fact, for this value we are still estimating the maximum number of factors (8), and it is not

Figure 3: Importance of the estimated factors.




*Note*: The horizontal and vertical axes display time and the percentage of variance explained, respectively.

until $c = 9.9$ that $\hat{r}_{c,N,T}$ start to go down. This is consistent with the finding of, for example, Breitung and Eickmeier (2011), and Stock and Watson (2005), that the IC tend to be flat for $\hat{r}_{N,T} \in [6, 10]$.

As an illustration of the importance of the first two principal components of the data and how it has changed over time, in Figure 3 we plot the percentage of the total variation in the data that can be explained by the estimated common component based on two factors. The following observations can be made. First, the percentage of variation explained by the first common factor is very high. It never goes below 80% and in most of the times it is over 95%. A single factor is therefore sufficient to account for most of the variation in the data. This is confirmed by our (unreported) results based on using six rather than one factor, as is done in Alessi et al. (2010). With six factors, the marginal contribution to the explained variation of the data of the sixth factor is no more than 0.0006%, which means that it can be safely ignored, as can factors seven to 10 had we included 10 factors. Second, the importance of the first common factor drops quite markedly in December 1987, which is largely consistent

with the start of the so-called "the Great Moderation", which refers to the general decline in volatility experienced by US output growth and inflation (see, for example, Bernanke (2004)). Third, while not nearly as important as the first, the second factor is still quite interesting, as it seems to be able to explain most of the variation that the first factor cannot explain. In fact, the "importance profiles" of the first and second factors are almost exact mirror images of each other. Fourth, the percentage of variation explained by both factors is very high and never falls below 97%.

The fact that the importance of the first and second factors form mirror images of each other is worthy of further discussion. In a recent work, Breitung and Eickmeier (2011) consider a model that is similar to (**??**), but where $\lambda_i$ is subject to structural change. The following specification captures this: $\lambda_{i,t} = D_t\lambda_i^1 + (1 - D_t)\lambda_i^2$, where $D_t$ is a dummy variable that takes on the value one for $t = 1, ..., T_B$ and zero otherwise. Letting $\lambda_t = (\lambda_{i,t}, ..., \lambda_{N,t})'$ with similar definitions of $\lambda^1$ and $\lambda^2$, $e_t = (e_{1,t}, ..., e_{N,t})'$, $\Sigma_e = E(e_t e_t')$ and $\Sigma_F = E(F_t F_t')$, such that $X_t = \lambda_t F_t + e_t$, it is not difficult to show that

$$E(X_t X_t') = \lambda_t \Sigma_F \lambda_t' + \Sigma_e = D_t \lambda^1 \Sigma_F (\lambda^1)' + (1 - D_t)\lambda^2 \Sigma_F (\lambda^2)' + \Sigma_e. \tag{7}$$

Therefore, since $\lambda^1\Sigma_F(\lambda^1)'$ and $\lambda^2\Sigma_F(\lambda^2)'$ are both assumed to be of full rank $r_0$, the rank of $E(X_t X_t')$ will in general be $2r_0$. Hence, if there is a break in the loadings, then the number of factors will be inflated by a factor of two. Interestingly, there is a slightly different version of the above model that is perhaps more intuitive in the present context, namely, that it is $\Sigma_F$ and not $\lambda_i$ that is subject to a break. Specifically, letting $\Sigma_{F,t} = D_t\Sigma_F^1 + (1 - D_t)\Sigma_F^2$, we obtain

$$E(X_t X_t') = \lambda \Sigma_{F,t} \lambda' + \Sigma_e = D_t \lambda \Sigma_F^1 \lambda' + (1 - D_t)\lambda \Sigma_F^2 \lambda' + \Sigma_e. \tag{8}$$

Of course, $\lambda^1\Sigma_F(\lambda^1)' = \lambda^1 H^{-1} H \Sigma_F H' (H^{-1}\lambda^1)' = \lambda\Sigma_F^1\lambda'$ for any invertible matrix $H$ (with a similar result applying to $\lambda^2\Sigma_F(\lambda^2)'$), suggesting that the covariance structures in (**??**) and (**??**) are observationally equivalent. This means that we can apply the test approach of Breitung and Eickmeier (2011) to test if what we have estimated here as two factors are actually one factor with a breaking variance reflecting the Great Moderation. Application of their Wald, likelihood ratio (LR) and Lagrange multiplier (LM) tests produces three test values for each of the 132 macroeconomic indicators in the sample, which is too much information to digest. Therefore, as a summary measure, we compute the average rejection frequency, which should be close to the 5% level under the null hypothesis of no break. The rejection frequencies for the Wald, LR and LM tests are given by 35%, 31% and 58%, respectively,

which are much higher than expected had the null been true. We therefore conclude that the factor variance has in fact undergone a structural break in December 1987. Hence, even if $IC_{c,p1}$ estimates just one factor, if we allow for two, we find that there is actually just one factor but with a break in variance.

The above results are interesting not only in their own right, but also for what they imply for the results reported in other studies. In particular, it suggests that the common problem of overfitting the number of common factors can be (partly) avoided by using HL.

# 5 Conclusion

The results reported in the current paper suggests that if properly implemented HL can lead to quite substantial gains in performance relative to the original IC of Bai and Ng (2002). Implementation-wise we reach the following conclusions:

- As long as $r_{max} > r_0$, the choice of $r_{max}$ is relatively unimportant. The only exception is if $r_{max}$ is "too close" to $r_0$; hence, it is important that $r_{max}$ is set large enough. In our case, $r_{max} = 8\lfloor(\min\{N,T\}/100)^{1/4}\rfloor$ seemed to work very well. This is true when using both simulated and real data.

- The results are rather insensitive to the choice of subsamples. The best overall performance was obtained by setting $(N_1, T_1)$ and $J$ relatively large. Among the combinations considered here, $(N_1, T_1) = 0.5(N, T)$ and $J \in \{6, 11\}$ worked well.

- The candidate values of $c$ have a significant impact on the performance of the IC, and stand out as the most important choice by far. In the experiments considered here setting $\Delta c \in \{0.05, 0.1\}$ led to good performance. The appropriate value of $c_{max}$ depends heavily on the IC considered. In the simulations, while for $IC_{p1}$ the best performance is obtained by setting $c_{max} = 2$, for $PC_{p1}$ the best performance is obtained by setting $c_{max} = 6$. However, from the real data illustration we learned that the appropriate value of $c_{max}$ can also be highly dependent on the DGP, and that in such cases a reasonable strategy involved sequentially increasing the value of $c_{max}$.

## References


Alessi, L., M. Barigozzi and M. Capasso (2010). Improved Penalization for Determining the Number of Factors in Approximate Factor Models. *Statistics and Probability Letters* **80**, 1806–1813.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica* **71**, 135–171.

Bai, J., and S. Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* **70**, 191–221.

Bernanke, B. (2004). The Great Moderation **20**.

Breitung, J., and S. Eickmeier (2011). Testing for Structural Breaks in Dynamic Factor Models. *Journal of Econometrics* **163**, 71–84.

Hallin, M., and R. Liška (2007). Determining the Number of Factors in the General Dynamic Factor Model. *Journal of the American Statistical Association* **102**, 603–617.

Ng, S., and P. Perron (1995). Unit Root Tests in ARMA Models With Data Dependent Methods for Selection of the Truncation Lag. *Journal of the American Statistical Association* **90**, 268–281.

Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica* **74**, 967–1012.

Stock, J. H., and M. W. Watson (2005). Implications of Dynamic Factor Models for VAR Analysis. NBER Working Paper No. 11467.

Stock, J. H., and M. W. Watson (2008). Forecasting in Dynamic Factor Models Subject to Structural Instability. In Castle, J., and N. Shephard (Eds.), *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*. Oxford University Press, Oxford.

Table 1: Results for C1.

| | | Correct selection frequency | | | | Average $\hat{r}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | $N, T$ | $IC_{c,p1}$ | $IC_{p1}$ | $PC_{c,p1}$ | $PC_{p1}$ | $IC_{c,p1}$ | $IC_{p1}$ | $PC_{c,p1}$ | $PC_{p1}$ |
| | | | | $r_{max} = 2$ | | | | | |
| 1 | 40 | 0.0 | 0.0 | 0.0 | 0.0 | 2.00 | 2.00 | 2.00 | 2.00 |
| 2 | 40 | 0.0 | 0.0 | 0.0 | 0.0 | 1.06 | 2.00 | 1.02 | 2.00 |
| 3 | 40 | 0.0 | 0.0 | 0.0 | 0.0 | 2.00 | 2.00 | 1.97 | 2.00 |
| 4 | 40 | 0.0 | 0.0 | 0.0 | 0.0 | 2.00 | 2.00 | 1.99 | 2.00 |
| 1 | 60 | 0.0 | 0.0 | 0.0 | 0.0 | 2.00 | 2.00 | 2.00 | 2.00 |
| 2 | 60 | 0.0 | 0.0 | 0.0 | 0.0 | 1.39 | 2.00 | 1.13 | 2.00 |
| 3 | 60 | 0.0 | 0.0 | 0.0 | 0.0 | 2.00 | 2.00 | 2.00 | 2.00 |
| 4 | 60 | 0.0 | 0.0 | 0.0 | 0.0 | 2.00 | 2.00 | 2.00 | 2.00 |
| | | | | $r_{max} = 3$ | | | | | |
| 1 | 40 | 99.4 | 100.0 | 98.2 | 100.0 | 2.99 | 3.00 | 2.98 | 3.00 |
| 2 | 40 | 0.4 | 92.0 | 0.0 | 100.0 | 1.25 | 2.92 | 1.75 | 3.00 |
| 3 | 40 | 95.8 | 100.0 | 91.4 | 100.0 | 2.96 | 3.00 | 2.91 | 3.00 |
| 4 | 40 | 98.0 | 100.0 | 96.4 | 100.0 | 2.98 | 3.00 | 2.96 | 3.00 |
| 1 | 60 | 100.0 | 100.0 | 100.0 | 100.0 | 3.00 | 3.00 | 3.00 | 3.00 |
| 2 | 60 | 15.0 | 99.8 | 2.6 | 100.0 | 1.62 | 3.00 | 1.76 | 3.00 |
| 3 | 60 | 100.0 | 100.0 | 100.0 | 100.0 | 3.00 | 3.00 | 3.00 | 3.00 |
| 4 | 60 | 100.0 | 100.0 | 100.0 | 100.0 | 3.00 | 3.00 | 3.00 | 3.00 |
| | | | | $r_{max} = 8\lfloor(\min\{N,T\}/100)^{1/4}\rfloor$ | | | | | |
| 1 | 40 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 6.00 |
| 2 | 40 | 33.2 | 92.0 | 52.6 | 0.0 | 1.86 | 2.92 | 2.41 | 6.00 |
| 3 | 40 | 96.8 | 49.8 | 95.0 | 0.0 | 2.98 | 3.78 | 3.08 | 6.00 |
| 4 | 40 | 54.0 | 0.4 | 9.2 | 0.0 | 3.55 | 5.08 | 4.48 | 6.00 |
| 1 | 60 | 92.6 | 99.8 | 100.0 | 0.0 | 2.87 | 3.00 | 3.00 | 7.00 |
| 2 | 60 | 100.0 | 96.2 | 96.8 | 0.0 | 3.00 | 3.04 | 2.97 | 7.00 |
| 3 | 60 | 54.2 | 0.0 | 98.2 | 0.0 | 3.73 | 5.59 | 3.04 | 7.00 |
| 4 | 60 | 54.2 | 0.0 | 2.8 | 0.0 | 3.73 | 5.59 | 5.40 | 7.00 |

*Notes*: $\hat{r}$ refer to the estimated value of $r$, the number of common factors. DGP1–DGP4 refer to the cases when the idiosyncratic component is $N(0,1)$, $N(0,4)$, serially correlated, and cross -correlation correlated, respectively. $IC_{p1}$ and $PC_{p1}$ ($IC_{c,p1}$ and $PC_{c,p1}$) refer to the original (HL-based) IC.

Table 2: Results for C2 when $N = T = 60$, $N_1 = 0.15N$ and/or $T_1 = 0.15T$.

| | | Correct selection frequency | | | | Average $\hat{r}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $J$ | DGP | $IC_{c,p1}$ | $IC_{p1}$ | $PC_{c,p1}$ | $PC_{p1}$ | $IC_{c,p1}$ | $IC_{p1}$ | $PC_{c,p1}$ | $PC_{p1}$ |
| | | Subsampling $N$: $(N_1, T_1) = (0.15N, T)$ | | | | | | | |
| 2 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 72.0 | 99.8 | 38.2 | 0.0 | 2.69 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 93.0 | 0.0 | 91.6 | 0.0 | 3.07 | 5.59 | 3.08 | 7.00 |
| 6 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 68.8 | 99.8 | 38.2 | 0.0 | 2.66 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 94.2 | 0.0 | 91.6 | 0.0 | 3.06 | 5.59 | 3.08 | 7.00 |
| 11 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 68.2 | 99.8 | 38.2 | 0.0 | 2.65 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 94.8 | 0.0 | 91.6 | 0.0 | 3.05 | 5.59 | 3.08 | 7.00 |
| | | Subsampling $T$: $(N_1, T_1) = (N, 0.15T)$ | | | | | | | |
| 2 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 73.2 | 99.8 | 38.2 | 0.0 | 2.70 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 94.2 | 0.0 | 91.6 | 0.0 | 3.06 | 5.59 | 3.08 | 7.00 |
| 6 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 68.4 | 99.8 | 38.2 | 0.0 | 2.64 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 95.6 | 0.0 | 91.6 | 0.0 | 3.04 | 5.59 | 3.08 | 7.00 |
| 11 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 66.6 | 99.8 | 38.2 | 0.0 | 2.61 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 95.8 | 0.0 | 91.6 | 0.0 | 3.04 | 5.59 | 3.08 | 7.00 |
| | | Subsampling $N$ and $T$: $(N_1, T_1) = 0.15(N, T)$ | | | | | | | |
| 2 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 15.6 | 99.8 | 38.2 | 0.0 | 1.59 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 99.8 | 0.0 | 91.6 | 0.0 | 3.00 | 5.59 | 3.08 | 7.00 |
| 6 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 15.4 | 99.8 | 38.2 | 0.0 | 1.59 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 99.8 | 0.0 | 91.6 | 0.0 | 3.00 | 5.59 | 3.08 | 7.00 |
| 11 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 15.4 | 99.8 | 38.2 | 0.0 | 1.59 | 3.00 | 2.33 | 7.00 |
| | 3 | 100.0 | 96.2 | 100.0 | 0.0 | 3.00 | 3.04 | 3.00 | 7.00 |
| | 4 | 99.8 | 0.0 | 91.6 | 0.0 | 3.00 | 5.59 | 3.08 | 7.00 |

*Notes*: See Table 1 for an explanation.

Table 3: Results for C2 when $N = T = 60$, $N_1 = 0.5N$ and / or $T_1 = 0.5T$.

| | | Correct selection frequency | | | | Average $\hat{r}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $J$ | DGP | $IC_{c,p1}$ | $IC_{p1}$ | $PC_{c,p1}$ | $PC_{p1}$ | $IC_{c,p1}$ | $IC_{p1}$ | $PC_{c,p1}$ | $PC_{p1}$ |
| | | Subsampling $N$: $(N_1, T_1) = (0.5N, T)$ | | | | | | | |
| 2 | 1 | 100.0 | 100.0 | 95.6 | 0.0 | 3.00 | 3.00 | 3.07 | 7.00 |
| | 2 | 95.0 | 99.8 | 96.0 | 0.0 | 2.92 | 3.00 | 3.07 | 7.00 |
| | 3 | 92.2 | 96.2 | 23.2 | 0.0 | 3.08 | 3.04 | 4.74 | 7.00 |
| | 4 | 17.2 | 0.0 | 0.2 | 0.0 | 4.54 | 5.59 | 5.56 | 7.00 |
| 6 | 1 | 100.0 | 100.0 | 98.8 | 0.0 | 3.00 | 3.00 | 3.02 | 7.00 |
| | 2 | 97.0 | 99.8 | 98.4 | 0.0 | 2.96 | 3.00 | 3.03 | 7.00 |
| | 3 | 95.8 | 96.2 | 41.0 | 0.0 | 3.05 | 3.04 | 4.14 | 7.00 |
| | 4 | 40.0 | 0.0 | 1.0 | 0.0 | 4.04 | 5.59 | 5.54 | 7.00 |
| 11 | 1 | 100.0 | 100.0 | 98.8 | 0.0 | 3.00 | 3.00 | 3.02 | 7.00 |
| | 2 | 98.4 | 99.8 | 98.6 | 0.0 | 2.98 | 3.00 | 3.02 | 7.00 |
| | 3 | 96.2 | 96.2 | 52.6 | 0.0 | 3.04 | 3.04 | 3.85 | 7.00 |
| | 4 | 45.2 | 0.0 | 2.0 | 0.0 | 3.93 | 5.59 | 5.43 | 7.00 |
| | | Subsampling $T$: $(N_1, T_1) = (N, 0.5T)$ | | | | | | | |
| 2 | 1 | 100.0 | 100.0 | 96.4 | 0.0 | 3.00 | 3.00 | 3.07 | 7.00 |
| | 2 | 95.6 | 99.8 | 96.8 | 0.0 | 2.93 | 3.00 | 3.07 | 7.00 |
| | 3 | 99.2 | 96.2 | 89.0 | 0.0 | 3.02 | 3.04 | 3.24 | 7.00 |
| | 4 | 29.2 | 0.0 | 0.4 | 0.0 | 4.43 | 5.59 | 5.71 | 7.00 |
| 6 | 1 | 100.0 | 100.0 | 98.4 | 0.0 | 3.00 | 3.00 | 3.03 | 7.00 |
| | 2 | 97.4 | 99.8 | 98.8 | 0.0 | 2.96 | 3.00 | 3.02 | 7.00 |
| | 3 | 100.0 | 96.2 | 95.0 | 0.0 | 3.00 | 3.04 | 3.10 | 7.00 |
| | 4 | 29.2 | 0.0 | 0.2 | 0.0 | 4.52 | 5.59 | 5.76 | 7.00 |
| 11 | 1 | 100.0 | 100.0 | 98.8 | 0.0 | 3.00 | 3.00 | 3.02 | 7.00 |
| | 2 | 97.8 | 99.8 | 99.0 | 0.0 | 2.96 | 3.00 | 3.01 | 7.00 |
| | 3 | 100.0 | 96.2 | 96.0 | 0.0 | 3.00 | 3.04 | 3.07 | 7.00 |
| | 4 | 32.0 | 0.0 | 0.4 | 0.0 | 4.47 | 5.59 | 5.68 | 7.00 |
| | | Subsampling $N$ and $T$: $(N_1, T_1) = 0.5(N, T)$ | | | | | | | |
| 2 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 93.4 | 99.8 | 99.8 | 0.0 | 2.89 | 3.00 | 3.00 | 7.00 |
| | 3 | 99.4 | 96.2 | 92.8 | 0.0 | 3.01 | 3.04 | 3.18 | 7.00 |
| | 4 | 22.0 | 0.0 | 1.4 | 0.0 | 4.43 | 5.59 | 5.69 | 7.00 |
| 6 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 92.6 | 99.8 | 96.8 | 0.0 | 2.87 | 3.00 | 2.97 | 7.00 |
| | 3 | 100.0 | 96.2 | 98.2 | 0.0 | 3.00 | 3.04 | 3.04 | 7.00 |
| | 4 | 54.2 | 0.0 | 2.8 | 0.0 | 3.73 | 5.59 | 5.40 | 7.00 |
| 11 | 1 | 100.0 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 7.00 |
| | 2 | 91.0 | 99.8 | 96.2 | 0.0 | 2.85 | 3.00 | 2.96 | 7.00 |
| | 3 | 100.0 | 96.2 | 99.0 | 0.0 | 3.00 | 3.04 | 3.02 | 7.00 |
| | 4 | 61.0 | 0.0 | 6.0 | 0.0 | 3.59 | 5.59 | 5.17 | 7.00 |

*Notes*: See Table 1 for an explanation.

Table 4: Results for $IC_{p1}$ and C3.

| | Correct selection frequency | | | | Average $\hat{r}$ | | | | Average $\hat{c}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | $c_{max}=2$ | $c_{max}=4$ | $c_{max}=6$ | $c=1$ | $c_{max}=2$ | $c_{max}=4$ | $c_{max}=6$ | $c=1$ | $c_{max}=2$ | $c_{max}=4$ | $c_{max}=6$ |
| | | | | | $\Delta c=0.05,\ N=T=40$ | | | | | | |
| 1 | 100.0 | 99.4 | 99.4 | 100.0 | 3.00 | 2.99 | 2.99 | 3.00 | 1.06 | 1.07 | 1.08 |
| 2 | 36.8 | 36.8 | 36.8 | 92.0 | 1.94 | 1.92 | 1.92 | 2.92 | 1.30 | 1.31 | 1.31 |
| 3 | 96.8 | 81.0 | 81.0 | 49.8 | 2.99 | 2.69 | 2.69 | 3.78 | 1.49 | 1.63 | 1.63 |
| 4 | 49.2 | 37.8 | 37.8 | 0.4 | 3.63 | 3.41 | 3.41 | 5.08 | 1.46 | 1.59 | 1.59 |
| | | | | | $\Delta c=0.1,\ N=T=40$ | | | | | | |
| 1 | 100.0 | 98.8 | 98.8 | 100.0 | 3.00 | 2.98 | 2.98 | 3.00 | 1.10 | 1.11 | 1.11 |
| 2 | 33.2 | 33.2 | 33.2 | 92.0 | 1.86 | 1.84 | 1.84 | 2.92 | 1.36 | 1.36 | 1.36 |
| 3 | 96.8 | 80.0 | 80.0 | 49.8 | 2.98 | 2.66 | 2.66 | 3.78 | 1.52 | 1.69 | 1.69 |
| 4 | 54.0 | 39.6 | 39.6 | 0.4 | 3.55 | 3.28 | 3.28 | 5.08 | 1.51 | 1.68 | 1.68 |
| | | | | | $\Delta c=0.5,\ N=T=40$ | | | | | | |
| 1 | 99.6 | 58.0 | 53.6 | 100.0 | 3.00 | 2.22 | 2.09 | 3.00 | 1.73 | 2.52 | 2.58 |
| 2 | 1.4 | 1.0 | 1.0 | 92.0 | 1.12 | 1.07 | 1.07 | 2.92 | 1.86 | 2.02 | 2.13 |
| 3 | 96.2 | 65.4 | 65.4 | 49.8 | 2.96 | 2.35 | 2.35 | 3.78 | 1.71 | 2.11 | 2.12 |
| 4 | 81.0 | 41.8 | 41.8 | 0.4 | 3.15 | 2.38 | 2.36 | 5.08 | 1.84 | 2.43 | 2.44 |
| | | | | | $\Delta c=0.05,\ N=T=60$ | | | | | | |
| 1 | 100.0 | 100.0 | 100.0 | 100.0 | 3.00 | 3.00 | 3.00 | 3.00 | 0.84 | 0.84 | 0.84 |
| 2 | 94.4 | 94.2 | 94.2 | 99.8 | 2.91 | 2.90 | 2.90 | 3.00 | 0.88 | 0.88 | 0.88 |
| 3 | 99.4 | 99.2 | 99.2 | 96.2 | 3.01 | 3.00 | 3.00 | 3.04 | 1.17 | 1.17 | 1.17 |
| 4 | 48.6 | 46.2 | 46.0 | 0.0 | 3.77 | 3.73 | 3.73 | 5.54 | 1.39 | 1.43 | 1.44 |
| | | | | | $\Delta c=0.1,\ N=T=60$ | | | | | | |
| 1 | 100.0 | 100.0 | 100.0 | 100.0 | 3.00 | 3.00 | 3.00 | 3.00 | 0.86 | 0.86 | 0.86 |
| 2 | 93.0 | 92.8 | 92.8 | 99.8 | 2.88 | 2.88 | 2.88 | 3.00 | 0.91 | 0.91 | 0.91 |
| 3 | 99.8 | 99.6 | 99.6 | 96.2 | 3.00 | 3.00 | 3.00 | 3.04 | 1.19 | 1.20 | 1.20 |
| 4 | 56.0 | 53.0 | 52.6 | 0.0 | 3.65 | 3.60 | 3.59 | 5.54 | 1.45 | 1.51 | 1.52 |
| | | | | | $\Delta c=0.5,\ N=T=60$ | | | | | | |
| 1 | 100.0 | 83.0 | 0.8 | 100.0 | 3.00 | 2.75 | 1.03 | 3.00 | 2.00 | 3.98 | 4.90 |
| 2 | 15.0 | 0.0 | 0.0 | 99.8 | 1.59 | 1.04 | 1.04 | 3.00 | 1.98 | 2.26 | 2.32 |
| 3 | 100.0 | 92.4 | 90.6 | 96.2 | 3.00 | 2.86 | 2.81 | 3.04 | 1.55 | 1.73 | 1.76 |
| 4 | 90.2 | 81.6 | 76.6 | 0.0 | 3.10 | 2.95 | 2.83 | 5.54 | 1.83 | 2.10 | 2.16 |

*Notes*: $c$ refers to the multiplicative constant on the penalty. $c_{max}$ and $\Delta c$ refer to the maximum $c$ and step size in $c$ considered in the HL procedure, respectively. $\hat{r}$ and $\hat{c}$ refer to the estimated values of $r$ and $c$, respectively. DGP1–DGP4 refer to the cases when the idiosyncratic component is $N(0,1)$, $N(0,4)$, serially correlated, and cross-correlation correlated, respectively. While the columns labeled "$c_{max}=2$", "$c_{max}=4$" and "$c_{max}=6$" refer to the HL procedure, the column labeled "$c=1$" refers to the original (non-data-driven) IC.

Table 5: Results for $PC_{p1}$ and C3.

| | Correct selection frequency | | | | Average $\hat{r}$ | | | | Average $\hat{c}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | $c_{max}=2$ | $c_{max}=4$ | $c_{max}=6$ | $c=1$ | $c_{max}=2$ | $c_{max}=4$ | $c_{max}=6$ | $c=1$ | $c_{max}=2$ | $c_{max}=4$ | $c_{max}=6$ |
| | | | | | $\Delta c=0.05,\ N=T=40$ | | | | | | |
| 1 | 30.6 | 100.0 | 100.0 | 0.0 | 3.81 | 3.00 | 3.00 | 6.00 | 2.00 | 3.82 | 4.12 |
| 2 | 32.2 | 77.0 | 53.2 | 0.0 | 3.78 | 2.77 | 2.42 | 6.00 | 2.00 | 3.83 | 4.59 |
| 3 | 0.0 | 91.2 | 94.4 | 0.0 | 5.94 | 3.12 | 3.08 | 6.00 | 2.00 | 3.96 | 5.58 |
| 4 | 0.0 | 2.6 | 8.4 | 0.0 | 5.76 | 4.61 | 4.51 | 6.00 | 2.00 | 3.26 | 3.50 |
| | | | | | $\Delta c=0.1,\ N=T=40$ | | | | | | |
| 1 | 30.6 | 100.0 | 100.0 | 0.0 | 3.81 | 3.00 | 3.00 | 6.00 | 2.00 | 3.83 | 4.15 |
| 2 | 32.2 | 76.6 | 52.6 | 0.0 | 3.78 | 2.77 | 2.41 | 6.00 | 2.00 | 3.84 | 4.62 |
| 3 | 0.0 | 91.6 | 95.0 | 0.0 | 5.94 | 3.11 | 3.08 | 6.00 | 2.00 | 3.96 | 5.60 |
| 4 | 0.0 | 2.8 | 9.2 | 0.0 | 5.76 | 4.59 | 4.48 | 6.00 | 2.00 | 3.29 | 3.56 |
| | | | | | $\Delta c=0.5,\ N=T=40$ | | | | | | |
| 1 | 30.6 | 100.0 | 100.0 | 0.0 | 3.81 | 3.00 | 3.00 | 6.00 | 2.00 | 3.91 | 4.34 |
| 2 | 32.2 | 74.4 | 44.2 | 0.0 | 3.78 | 2.74 | 2.30 | 6.00 | 2.00 | 3.92 | 4.92 |
| 3 | 0.0 | 94.6 | 98.0 | 0.0 | 5.94 | 3.06 | 3.02 | 6.00 | 2.00 | 3.99 | 5.76 |
| 4 | 0.0 | 4.2 | 16.2 | 0.0 | 5.76 | 4.41 | 4.22 | 6.00 | 2.00 | 3.61 | 4.09 |
| | | | | | $\Delta c=0.05,\ N=T=60$ | | | | | | |
| 1 | 99.8 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 6.00 | 2.00 | 2.63 | 2.63 |
| 2 | 99.6 | 99.4 | 98.8 | 0.0 | 3.01 | 3.00 | 2.99 | 6.00 | 2.00 | 2.63 | 2.63 |
| 3 | 0.0 | 95.4 | 95.4 | 0.0 | 5.21 | 3.07 | 3.07 | 6.00 | 2.00 | 3.80 | 4.10 |
| 4 | 0.0 | 1.0 | 4.0 | 0.0 | 5.96 | 4.83 | 4.79 | 6.00 | 2.00 | 2.89 | 2.96 |
| | | | | | $\Delta c=0.1,\ N=T=60$ | | | | | | |
| 1 | 99.8 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 6.00 | 2.00 | 2.65 | 2.65 |
| 2 | 99.8 | 99.4 | 98.8 | 0.0 | 3.00 | 2.99 | 2.99 | 6.00 | 2.00 | 2.65 | 2.66 |
| 3 | 0.0 | 96.4 | 96.4 | 0.0 | 5.21 | 3.05 | 3.05 | 6.00 | 2.00 | 3.82 | 4.13 |
| 4 | 0.0 | 1.2 | 4.6 | 0.0 | 5.96 | 4.81 | 4.76 | 6.00 | 2.00 | 2.93 | 3.01 |
| | | | | | $\Delta c=0.5,\ N=T=60$ | | | | | | |
| 1 | 99.8 | 100.0 | 100.0 | 0.0 | 3.00 | 3.00 | 3.00 | 6.00 | 2.00 | 2.85 | 2.85 |
| 2 | 99.8 | 98.8 | 96.2 | 0.0 | 3.00 | 2.99 | 2.96 | 6.00 | 2.00 | 2.87 | 2.90 |
| 3 | 0.0 | 99.4 | 99.4 | 0.0 | 5.21 | 3.01 | 3.01 | 6.00 | 2.00 | 3.94 | 4.38 |
| 4 | 0.0 | 6.6 | 19.6 | 0.0 | 5.96 | 4.45 | 4.28 | 6.00 | 2.00 | 3.41 | 3.77 |

*Notes*: See Table 4 for an explanation.