M2 EEE Machine Learning for Economics

Factor Models and In-Sample Inference: A Replication and A Simulation

---

Nikita Marini,

Luca Poll

April 19, 2021

Table of Contents

1 Introduction

The results presented in this report are concerned with two closely linked, yet inherently different exercises that consider the properties of different information criteria (ICs) in the context of factor models.

The first part, which we will refer to as the Simulation, aims at replicating (most of) the results contained in Carrasco and Rossi (2016); in particular, the goal is to employ a series of different ICs (Mallows' C_L, Generalized Cross Validation, AIC, BIC, Bai and Ng (2002)'s IC_{p2}, etc.) in order to estimate the correct number of relevant factors in 6 different Data Generating Processes (DGPs), each of which is simulated between 100 and 10,000 times.

After having presented some of the trends and the conclusions derived from the Simulation exercise, we will apply the same ICs to some real-world data, in order to arrive at a consistent estimate of the number of factors underlying the data at hand.

2 Simulation

2.1 Data Generating Processes

In order to test the performance of the various Information Criteria, we first need to generate some simulated data. Following the methodology of the original paper, we simulate six different Data Generating Processes, each of which has a specific factor: for instance, while DGP 1 only contains 4 relevant factors, the number of factors in DGP 5 is equivalent to the N dimension of our sample; DGP 4, on the other hand, does not contain any relevant factors (as a reminder, and to quickly assess the performance of the ICs, we will report the true number of factors in all of our tables, denoted by k^*). All six DGPs therefore take the following form:

$$\underset{(T\times N)}{X} = \underset{(T\times r)}{F}\underset{(r\times N)}{\Lambda'} + \underset{(T\times N)}{\xi} \tag{1}$$

with Λ' a matrix of r independent factor loadings, each generated from a $\mathcal{N}(0,1)$ and ξ a matrix of T vectors of length N representing iid error terms, following a standard normal distribution. Furthermore, all six DGPs are repeated twice: once for a small-sample scenario, where $N = 100$ and $T = 50$ (therefore where the number of dimensions is greater than the number of observations), and once for a large sample characterized by $N = 200$ and $T = 500$.

Having generated the set of covariates, the same $(T \times r)$ matrix of factors F is then used to generate the response variable y, according to the following equation:

$$\underset{(T\times 1)}{y} = \underset{(T\times r)(r\times 1)}{F}\ \underset{}{\theta} + \underset{(T\times 1)}{v} \tag{2}$$

where θ is the vector that determines which factors are indeed relevant in each DGP, as per the examples above, and v is a vector of iid error terms following a $\mathcal{N}(0, \sigma_v^2)$.

Each DGP is then simulated 10,000 times, in the case of the small sample, and 100 times in the case of the large sample[1].

2.2 Estimators

The next step of interest is concerned with the prediction of the different models. Again following the original paper, we code each of the estimators presented therein, in order to obtain our fitted values $\hat{y}$. In particular, we employ the following:

$$\hat{y} = M_T^{\alpha} y = \frac{1}{T} \sum_{j=1}^{N \wedge T} \hat{q}_j \hat{\psi}_j \hat{\psi}_j' y \tag{3}$$

where M_T^{α} is the *hat matrix*, obtained for a given value of the tuning parameter α, and $\hat{\psi}_j$ are the $(T \times 1)$ orthonormalized eigenvectors of the matrix XX'/T[2]. Finally, $\hat{q}_j \equiv q(\alpha, \lambda_j^2)$ and its form depends on the estimator considered; notably:

[1] this discrepancy is due to the large amount of computing power necessary to estimate the model for the large sample; the robustness of the results is therefore much greater in the small sample case, than in the large one.

[2] notice that, in order to normalize them so that $\hat{\psi}_j \hat{\psi}_j / T = 1$, we have multiply the standard normalized eigenvectors (i.e., vectors of norm 1) by $\sqrt{T}$.

▷ for ridge: $q(\alpha, \lambda_j^2) = \frac{\lambda_j^2}{\lambda_j^2+\alpha}$

▷ for Landweber Fridman (LF): $q(\alpha, \lambda_j^2) = 1 - (1 - d\lambda_j^2)^{\frac{1}{\alpha}}$

▷ for principal components (PC) with k components: $q(\alpha, \lambda_j^2) = \mathbb{1}(j \leq k)$

where λ_j^2 is the j^{th} eigenvalue associated to the $(N \times 1)$ orthonormalized eigenvector $\hat{\phi}_j$ of the matrix $X'X/T$, such that $\lambda_1^2 > \lambda_2^2 > ... > \lambda_N^2$. The d in the Landweber Fridman estimator is such that $0 < d < 1/\hat{\lambda}_1^2$ and, in accordance with the paper, is taken to be $0.018/\hat{\lambda}_1^2$.

Lastly, we also consider the Partial Least Squares (PLS) estimator but were not able to recast it as an inverse problem and therefore adopted the following expression[3]:

$$M_T^{\alpha,PLS} = XV_k(V_k'X'XV_k)^{-1}V_k'X'y, \tag{4}$$

with V_k an $(N \times k)$ matrix defined as:

$$V_k = \big(X'y, (X'X)X'y, ..., (X'X)^{k-1}X'y\big) \tag{5}$$

2.3 Information Criteria

As mentioned above, each one of the estimates of M_T^{α} crucially depends on the choice of the tuning parameter α (which becomes k in the case of PC and PLS), which will in turn depend on the underlying structure of the data. In order to obtain the optimal $alpha/k$, we employ a number of Information Criteria, each of which we expect might perform better or pooorer than the others depending on the DGP at hand, as well as on the dimensions of the sample. The ICs that we focus our attention on are the following:

[3]the complication arose from the fact that to obtain the set $I_k^+ = \{(j_1, ..., j_k) : N \wedge T \geq j_1 > ... > j_k \geq 1\}$ it was necessary to compute a $\binom{N \wedge T}{k}$ permutation, which was computationally intensive as the dimensions of the dataset increased.

▷ Generalized Cross Validation (GCV):

$$\hat{\alpha} = \underset{\alpha \in A_T}{\arg\min} \ \frac{T^{-1}||y - M_T^{\alpha} y||^2}{(1 - T^{-1} tr(M_T^{\alpha}))^2};$$

▷ Mallows' C_L:

$$\hat{\alpha} = \underset{\alpha \in A_T}{\arg\min} \ T^{-1}||y - M_T^{\alpha} y||^2 + 2\hat{\sigma}_{\varepsilon}^2 T^{-1} tr(M_T^{\alpha});$$

▷ Leave One Out Cross Validation (LOOCV)[4]:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ \frac{T||y - M_T^{\alpha} y||^2}{1 - tr(M_T^{\alpha})};$$

▷ Bai and Ng (2002)'s IC_{p2}:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ ln(V(k, \hat{F}^k)) + k \left(\frac{N+T}{NT}\right) lnC_{NT}^2;$$

▷ Bai and Ng (2002)'s PC_{p2}:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ V(k, \hat{F}^k) + k\hat{\sigma}^2 \left(\frac{N+T}{NT}\right) lnC_{NT}^2;$$

▷ Traditional AIC:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ ln(V(k, \hat{F}^k)) + k \left(\frac{2}{T}\right);$$

▷ Traditional BIC:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ ln(V(k, \hat{F}^k)) + k \left(\frac{ln(T)}{T}\right);$$

▷ Bai and Ng (2002)'s AIC_1:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ V(k, \hat{F}^k) + k\hat{\sigma}^2 \left(\frac{2}{T}\right);$$

▷ Bai and Ng (2002)'s BIC_1:

$$\hat{k} = \underset{k \in 1,...,N \wedge T}{\arg\min} \ V(k, \hat{F}^k) + k\hat{\sigma}^2 \left(\frac{ln(T)}{T}\right)$$

where $tr(\cdot)$ is the trace operator, $||\cdot||$ is the Euclidean norm, $V(k, \hat{F}^k)$ contains the sum of squared residuals as defined in Bai and Ng (2002), $\hat{\sigma}_{\varepsilon}^2$ is a consistent estimator of the variance of the error terms in the model $y = X\delta^{\alpha} + \varepsilon$, with δ^{α} representing one of the four estimators presented above, and $C_{NT}^2 \equiv N \wedge T$. As the original paper did not specify the form of the AIC and BIC criteria, we opted to include two versions for each of them, which mostly differ in the inclusion of a variance term, $\hat{\sigma}^2$.

[4]for the PLS estimator, only.

As far as the tuning parameter α is concerned, the ranges over which performed our simulations are the same as the ones reported on footnote 12 in Carrasco and Rossi (2016), while the range of k was taken to be 1 to the value of r_{max} reported for each of the DGPs in the paper.

2.4 Results

In this section we present the results of our simulations and discuss which information criterion performs best in the different scenarios. It is important to point out that the results that we will discuss only concern the identification of the true number of underlying factors in each DGP, r, and the in-sample mean squared error (MSE), while we do not discuss the performance in terms of out-of-sample forecasting (the RMSFER column in the original paper).

2.4.1 Original paper

The departing benchmark is given by the results reported in Carrasco and Rossi (2016). In the large sample case (i.e., $N = 200, T = 500$) the authors report that, for the first two DGPs, where the number of factors is either very small or very large (4 and 50, respectively for DGP 1 and DGP 2), Generalized Cross Validation (GCV) does just as well as compared to the AIC and BIC criteria as applied to the PC estimator. However, if compared to the Bai and Ng (2002)'s IC_{p2}, GCV is superior when the true number of factors is very large, as in DGP 2. This result is attributed by the authors to the violation of one of the assumptions that justify the use of the IC_{p2} criterion. Lastly, Mallows' C_L performs worse than all of the other criteria, for both of the DGPs considered. When the number of factors is low and only 1 or none are relevant (DGPs 3 and 4, respectively) the AIC and BIC criteria outperform GCV; Mallows' C_L, however, is the only one that returns an optimal value of k of 0 for DGP 4 and in general does better than GCV even when applied to ridge and LF (albeit for these, the estimated number of factors remains far from its true value). DGP 5, which is characterized by a number of factors equal to

N and gradually decreasing eigenvalues is best estimated by employing Mallows, which outperforms all the other criteria (with the sole exception of Bai and Ng (2002)'s IC_{p2}, which does not however perform any data reduction and selects the maximum number of principal components, 200). Lastly, the near factor model of DGP 6 finds in AIC and BIC the best ICs to estimate its true structure, while both Mallows and GCV perform rather poorly. Finally, we find that the LOOCV algorithm, as applied to the PLS estimator, always reports a very low number of estimated factors, even when the true number is very high.

In summary, the results in Carrasco and Rossi (2016) show that AIC, BIC, and IC_p2 are rather trustworthy ICs for a large variety of specifications, but their performance can be improved upon by Mallows' C_L criterion, when the number of factors is large and eigenvalues decline gradually as well as in the case where no factor structure is present; and by GCV, when the true number of factors is large.

The results above hold in large part for the small sample case as well ($N = 100, T = 50$), with the noteworthy exceptions of an improvement of the performance of LOOCV in the case where the number of factors is high (DGP 2) and a deterioration of the performance of Mallows *vis-*$\tilde{A}$ *-vis* GCV when the number of relevant factors is very low (DGPs 3 and 4).

If we focus our attention on the Mean Squared Error, we notice however that the best results are delivered by two of the estimators for which both GCV and Mallows delivered poor results in terms of model selection, namely ridge and LF. This result can be interpreted as a sign of the fact that these two estimators are not particularly suited for identifying the factor structure of a data generating process, but are nonetheless very useful when the interest lays in obtaining good in-sample prediction.

2.4.2 Replication

Table 1 reports our average estimates for the tuning parameter (α in the case of ridge and LF, and k for PC and PLS) and the corresponding Degrees of Freedom (DoF, when applicable) for all of the ICs listed in Section 2.3. The second line of each DGP reports

the standard deviation across the 100 Monte Carlo simulations of the parameters above.

Focusing our attention on the performance of GCV and Mallows as applied to the PC estimator, we notice no significant difference between the two in our simulation, although Mallows seems to perform slightly better when the true number of factors is low (DGPs 1 and 6). Like in the original paper, we find that GCV is to be preferred to Bai and Ng (2002) when the number of factors is large, as in DGP 2. However, it must also be pointed out that the BIC_1 criterion (which includes a penalty for the variance of the residuals of the factor model), also identifies the true process correctly. In DGP 3, where the number of factors is 5, all criteria for PC perform equally and estimate the correct number of factors. In DGP 4, however, where the factors are uncorrelated to the y process, GCV and Mallows perform better than the other criteria, although they still estimate 2 relevant factors. Contrarily to the paper, our results are not able to replicate the good performance of Mallows for DGP 5, while when the number of relevant factors is low, we do indeed observe that Mallows performs better than GCV, even for ridge and LF.

Table 3, on the other hand, focuses on the small sample case. Contrarily to the paper, our estimates using the Bai and Ng (2002)'s IC_{p2} do not suffer from the high-dimensionality bias, at least when the number of true factors is small (DGPs 1, 3, 4, 6) and r is estimated very accurately. It is interesting to notice, however, that a slightly different specification of the same IC, which includes a variance term and is denoted by PC_{p2} (cf Section 2.3 for the exact expression) overestimates the true r in those very same cases, but does a much better job for DGP 2 and 5, where the number of factor is significantly higher. The addition of a penalty for the variance therefore helps improving the in-sample prediction when the true number of factors is thought to be relatively high. In accordance with the paper, we find that PC with BIC and GCV performs very well in most cases, although the traditional AIC (and, as just stated, PC_{p2}) are better when r is large.

Finally, the MSEs reported in Tables 2 and 4 highlight a result that is in common with the original paper; namely that the ridge estimator (and in some cases the LF, too) does a very good job in delivering a very low error in in-sample prediction. The red circles in

the tables indicate what is/are the estimator/s reporting the lowest MSE for each DGP. Indeed, in Table 4, with the sole exception of DGP 5, the ridge estimator tuned via GCV results in very low average predicted errors and, for the DGP 1, 2, and 3 is actually the best one out of the ones considered.

3 Application

The second part of this project is dedicated to an application of the methods discussed above. In particular, rather than looking at simulated data and comparing the estimates to the true underlying process, we bring the different estimators and the relative ICs to practice, by using them to obtain a consistent estimate of the number of factors that determine the data observed.

3.1 Setting

The data chosen for the application is the one used in Stock and Watson (2002). It consists of 215 monthly time series spanning the period January 1959-December 1998, therefore determining an X matrix of dimensions $N = 215 \times T = 480$, which puts us in a scenario similar to that of the large sample simulations. The 215 covariates represent different macroeconomic data indicators and we use them to predict the average monthly return of the S&P500 index over the same time period, obtained separately.

3.2 Results

Having chosen our data, we now estimate the in-sample Mean Squared Error of each of the four estimators presented above (ridge, Landweber Fridman, Principal Components, and Partial Least Squares) and choose the optimal tuning parameter via the same algorithms presented in Section 2.3.

	GCV						Mallows					AIC	AIC_1	IC_{p2}	PC_{p2}	BIC	BIC_1	LOOCV
	Ridge		LF		PC		Ridge		LF		PC	PC						PLS
	r	α	DoF	α	DoF	k	α	DoF	α	DoF	k	k						k
DGP 1	4	6.94	28.41	0.0037	13.34	4.63	20.00	12.73	0.0042	7.20	3.85	14.00	14.00	4.00	14.00	14.00	5.95	1.00
(s.e.)	-	1.66	5.54	0.0013	16.69	1.59	0.00	0.04	0.0000	0.29	0.44	0.00	0.00	0.00	0.00	0.00	0.73	0.00
DGP 2	50	1.99	92.79	0.0020	47.33	50.74	2.00	92.58	0.0020	47.24	49.10	200.00	124.76	200.00	200.00	200.00	50.00	1.00
(s.e.)	-	0.05	0.88	0.0000	0.52	2.07	0.00	0.18	0.0002	0.99	1.64	0.00	0.83	0.00	0.00	0.00	0.00	0.00
DGP 3	5	2.96	50.69	0.0005	15.58	5.00	20.00	13.80	0.0014	8.08	5.00	5.00	5.00	5.00	5.00	5.00	5.00	1.00
(s.e.)	-	0.33	3.66	0.0002	6.22	0.00	0.00	0.04	0.0003	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DGP 4	5	184.42	7.02	0.0603	2.19	2.00	188.46	5.77	0.0604	1.98	2.00	15.00	15.00	5.00	15.00	15.00	6.79	1.00
(s.e.)	-	53.10	9.50	0.0285	4.65	2.31	45.91	5.54	0.0283	4.08	2.31	0.00	0.00	0.00	0.00	0.00	0.61	0.00
DGP 5	200	7.03	36.81	0.0057	34.39	13.06	10.77	24.64	0.0078	10.75	10.41	200.00	115.17	200.00	200.00	200.00	17.77	1.20
(s.e.)	-	5.84	13.36	0.0148	33.44	10.33	5.77	7.79	0.0159	15.49	7.18	0.00	0.87	0.00	0.00	0.00	0.68	0.40
DGP 6	1	1.53	72.91	0.2002	6.35	7.00	3.30	45.16	0.2002	6.35	5.55	11.00	11.00	1.00	11.00	10.95	3.21	2.00
(s.e.)	-	0.32	8.02	0.0000	0.21	3.40	0.88	8.80	0.0000	0.21	3.45	0.00	0.00	0.00	0.00	0.41	0.74	0.00

Table 1: Estimated tuning parameters and Degrees of Freedom for the sample of dimensions $N = 200, T = 500$

	Ridge		LF		PLS	PC							
	GCV	Mallows	GCV	Mallows	LOOCV	AIC	AIC_1	IC_{p2}	PC_{p2}	BIC	BIC_1	GCV	Mallows
DGP 1	0.919	1.005	0.983	1.008	1.087	0.985	0.985	1.004	0.985	0.985	1.000	1.000	1.005
(s.e.)	0.069	0.065	0.088	0.064	0.087	0.062	0.062	0.063	0.062	0.062	0.062	0.063	0.063
DGP 2	1.000	1.001	2.192	2.211	12.801	0.803	1.004	0.803	0.803	0.803	1.203	1.197	1.319
(s.e.)	0.072	0.072	0.418	0.437	1.984	0.066	0.073	0.066	0.066	0.066	0.085	0.083	0.306
DGP 3	0.013	0.023	0.014	0.017	0.124	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015
(s.e.)	0.001	0.002	0.001	0.002	0.063	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
DGP 4	0.975	0.980	1.001	1.002	0.997	0.969	0.969	0.990	0.969	0.969	0.986	0.992	0.992
(s.e.)	0.076	0.069	0.067	0.067	0.070	0.061	0.061	0.061	0.061	0.061	0.061	0.062	0.062
DGP 5	166.798	176.130	169.326	187.336	191.381	110.012	141.669	110.012	110.012	110.012	183.537	184.271	186.362
(s.e.)	16.054	14.071	28.213	17.006	13.362	9.297	10.889	9.297	9.297	9.297	12.421	15.365	13.793
DGP 6	1.237	1.437	1.901	1.901	0.988	1.598	1.598	1.697	1.598	1.599	1.669	1.614	1.625
(s.e.)	0.101	0.120	0.122	0.122	0.071	0.098	0.098	0.110	0.098	0.098	0.103	0.102	0.101

Table 2: Mean Squared Error for the sample of dimensions $N = 200, T = 500$

		GCV					Mallows					AIC	AIC_1	IC_{p2}	PC_{p2}	BIC	BIC_1	LOOCV
		Ridge		LF		PC	Ridge		LF		PC	PC						PLS
	r	α	DoF	α	DoF	k	α	DoF	α	DoF	k	k						k
DGP 1	4	6.11	16.35	0.0027	8.43	4.73	10.00	10.54	0.0071	4.70	3.46	14.00	8.49	4.00	14.00	4.00	4.00	1.99
(s.e.)	-	2.97	8.11	0.0010	4.34	2.03	0.00	0.13	0.0017	0.63	0.80	0.00	0.72	0.00	0.00	0.00	0.00	0.11
DGP 2	50	0.53	47.62	0.0013	31.75	19.39	1.00	45.61	0.0155	9.19	7.52	25.00	25.00	1.01	25.00	25.00	25.00	2.00
(s.e.)	-	0.47	2.04	0.0010	3.16	6.11	0.00	0.21	0.0025	1.03	4.68	0.00	0.00	0.09	0.00	0.00	0.00	0.03
DGP 3	5	2.01	25.98	0.0005	12.27	5.00	10.00	11.48	0.0019	6.49	4.99	5.00	5.00	5.00	5.00	5.00	5.00	2.00
(s.e.)	-	0.96	7.00	0.0002	4.73	0.02	0.00	0.12	0.0006	0.47	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.04
DGP 4	5	85.59	7.17	0.0322	2.76	2.05	96.04	4.73	0.0342	1.56	1.83	15.00	9.17	5.00	15.00	5.00	5.00	1.91
(s.e.)	-	34.06	7.68	0.0124	5.37	2.36	18.35	3.00	0.0097	1.80	1.90	0.00	0.72	0.00	0.00	0.00	0.00	0.29
DGP 5	200	10.97	13.32	0.0205	5.22	3.56	14.71	8.36	0.0248	2.79	2.66	25.00	11.84	2.49	25.00	4.06	4.26	2.00
(s.e.)	-	5.46	9.52	0.0165	6.08	4.06	1.52	1.84	0.0145	2.79	2.76	0.00	0.74	0.52	0.00	0.67	0.53	0.03
DGP 6	1	4.94	17.83	0.2196	6.48	3.61	9.58	8.28	0.2336	2.66	2.73	11.00	6.23	1.00	11.00	1.00	1.00	2.00
(s.e.)	-	3.58	10.54	0.3981	12.76	2.90	1.29	1.79	0.3846	0.92	2.20	0.00	0.75	0.00	0.00	0.00	0.00	0.02

Table 3: Estimated tuning parameters and Degrees of Freedom for the sample of dimensions $N = 100, T = 50$

	Ridge		LF		PLS	PC							
	GCV	Mallows	GCV	Mallows	LOOCV	AIC	AIC_1	IC_{p2}	PC_{p2}	BIC	BIC_1	GCV	Mallows
DGP 1	0.567	0.737	0.796	1.066	0.901	0.751	0.866	0.960	0.751	0.960	0.960	0.915	1.007
(s.e.)	0.249	0.150	0.207	0.204	0.188	0.176	0.190	0.199	0.176	0.199	0.199	0.214	0.230
DGP 2	0.062	0.106	2.717	23.942	7.962	8.043	8.043	47.931	8.043	8.043	8.043	11.260	27.200
(s.e.)	0.068	0.045	1.985	4.273	1.847	2.703	2.703	9.773	2.703	2.703	2.703	7.257	9.955
DGP 3	0.007	0.023	0.014	0.044	0.067	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
(s.e.)	0.004	0.005	0.004	0.012	0.045	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
DGP 4	0.794	0.863	0.907	0.944	0.868	0.699	0.815	0.900	0.699	0.900	0.900	0.928	0.937
(s.e.)	0.284	0.199	0.251	0.202	0.182	0.168	0.182	0.191	0.168	0.191	0.191	0.216	0.210
DGP 5	57.514	69.384	80.300	88.524	50.306	45.127	70.876	93.160	45.127	88.771	88.236	84.750	88.149
(s.e.)	25.011	15.062	24.972	20.791	14.197	12.754	16.234	19.085	12.754	18.528	18.439	22.964	21.048
DGP 6	0.850	1.311	1.603	1.747	0.245	1.310	1.526	1.867	1.310	1.867	1.867	1.564	1.634
(s.e.)	0.479	0.275	0.572	0.341	0.077	0.299	0.329	0.381	0.299	0.381	0.381	0.393	0.380

Table 4: Mean Squared Error for the sample of dimensions $N = 100, T = 50$

As mentioned in the previous section, our setting is similar to that of the large sample in terms of dimensions. There, our results suggested that, if the true number of factors is small, then Mallows will perform better than GCV, while if the number of factors is large, GCV should be preferred to Bai and Ng (2002)'s IC_{p2}.

Table 5 displays the application of the different estimators with the respective information criteria to the Stock and Watson (2002) data. We report the optimal tuning parameter that is chosen via the respective information criteria along with the Mean Squared Error and the degrees of freedom in the case where the tuning parameter is α. The r_{max} (the maximum number of factors considered in the choice of the tuning parameter) was set to 100. We observe similar patterns as in the simulation with the large sample. Namely, we see that the number of factors indicated by GCV and Mallows is larger when the model is estimated by ridge than by Landweber Friedman. Strikingly, however, we see contrary to the simulation, that the GCV applied to the PC model select a higher number of factors than when estimated via ridge or LF. In the simulation we found that GCV and Mallows applied to the Principal Components model indicated fewer factors than other models in most DGPs, a pattern that we do not observe in the application. While the optimal PC parameter estimated via Mallows indicates 16 factors, GCV suggests 99. Another pattern we observed from the simulation in the large sample was that if we have few relevant factors, AIC and AIC_1 chose the same number of factors. In casu AIC is twice as large as AIC_1 which might be seen as a hint that there are more than a few relevant factors. As we can see from the graphical representation of the IC's minimizing choices of the optimal number of factors in the appendix, Bai and Ng (2002)'s IC_{p2}, Mallows and BIC_1 identify 15, 16 and 17 relevant factors. Given that these methods perform well in different contexts and in particular when the true number of factors is large (as in the DGP 2 , where $r = 50$), we may assume the true number of relevant factors to indeed be close to these estimates.

Lastly, and in line with the simulations, however, we do observe that the best in-sample prediction is achieved by the LF and the ridge estimators, where the number of factors selected via GCV and is equal to 97 and 87, respectively.

Estimator	IC	α or k	DoF	MSE
ridge	GCV	0.17677	97	0.0000618913
ridge	Mallows	1.66162	41	0.0001204942
LF	GCV	0.00030	87	0.0000655240
LF	Mallows	0.00202	35	0.0001053630
PLS	LOOCV	2		0.0001360271
PC	Bai and Ng (2002)'s IC_{p2}	15		0.0001216831
PC	Mallows	16		0.0001166881
PC	BIC_1	17		0.0001144964
PC	AIC_1	50		0.0000872261
PC	GCV	99		0.0000657498
PC	AIC	100		0.0000656057
PC	BIC	100		0.0000656057

Table 5: Estimated tuning parameters and Mean Squared Errors for each estimator-IC considered

References

Bai, J. and S. Ng (2002). "Determining the Number of Factors in Approximate Factor Models". In: *Econometrica* 70.1, pp. 191–221. DOI: https://doi.org/10.1111/1468-0262.00273.

Carrasco, M. and B. Rossi (2016). "In-Sample Inference and Forecasting in Misspecified Factor Models". In: *Journal of Business & Economic Statistics* 34.3, pp. 313–338. DOI: 10.1080/07350015.2016.1186029.

Stock, J. H. and M. W. Watson (2002). "Macroeconomic Forecasting Using Diffusion Indexes". In: *Journal of Business & Economic Statistics* 20.2, pp. 147–162. DOI: 10.1198/073500102317351921.

A Figure Appendix

Graphical Visualization of the Application Results

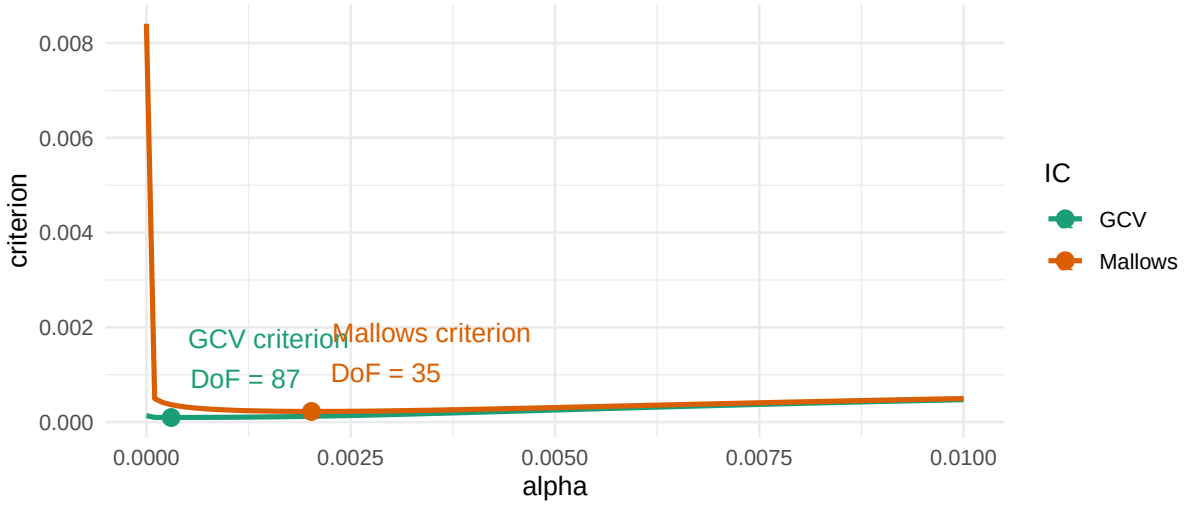

Figure 1: LF

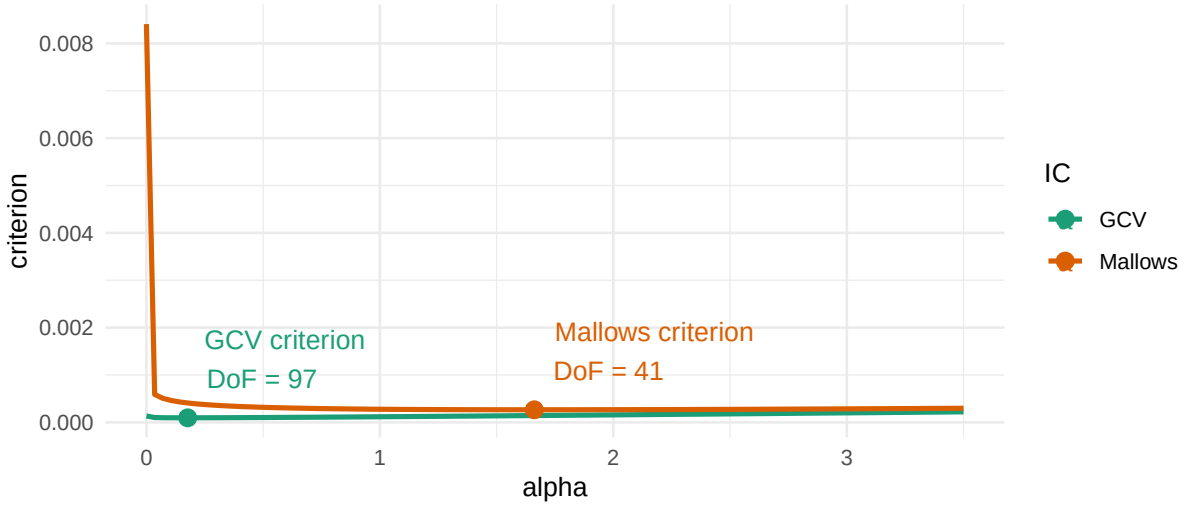

Figure 2: Ridge

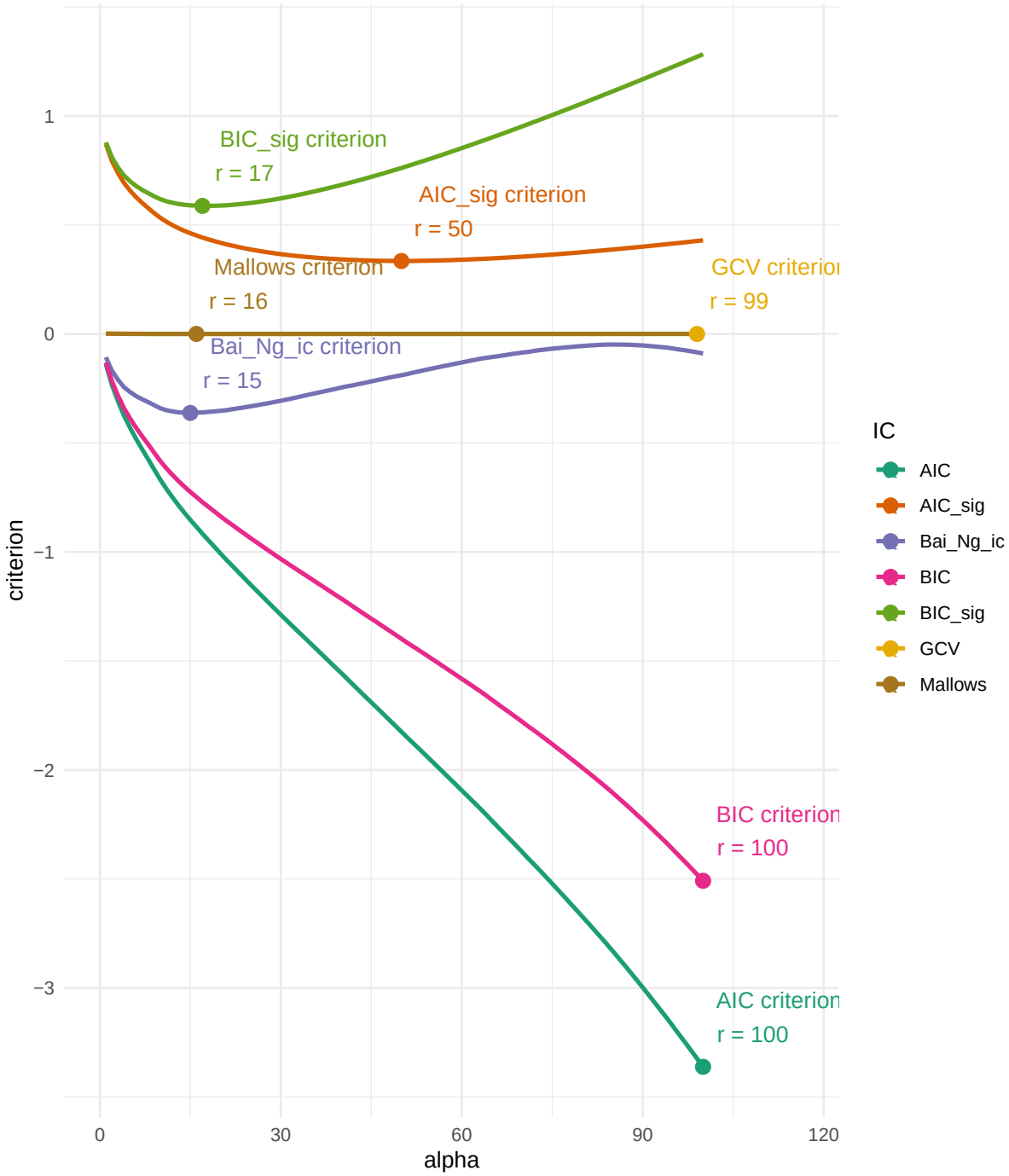

Figure 3: PC all criteria

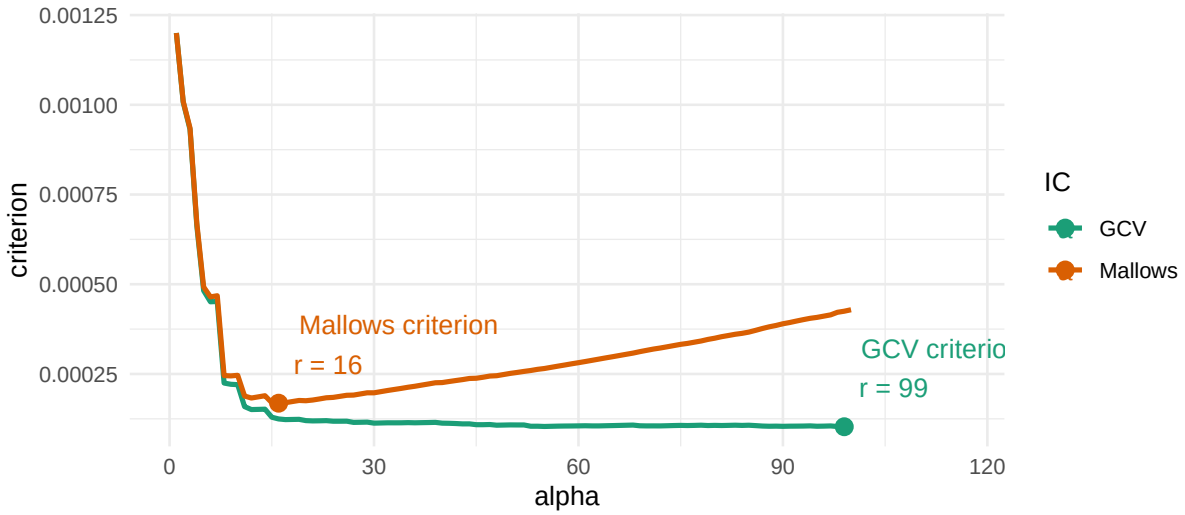

Figure 4: PC GCV and Mallows criteria

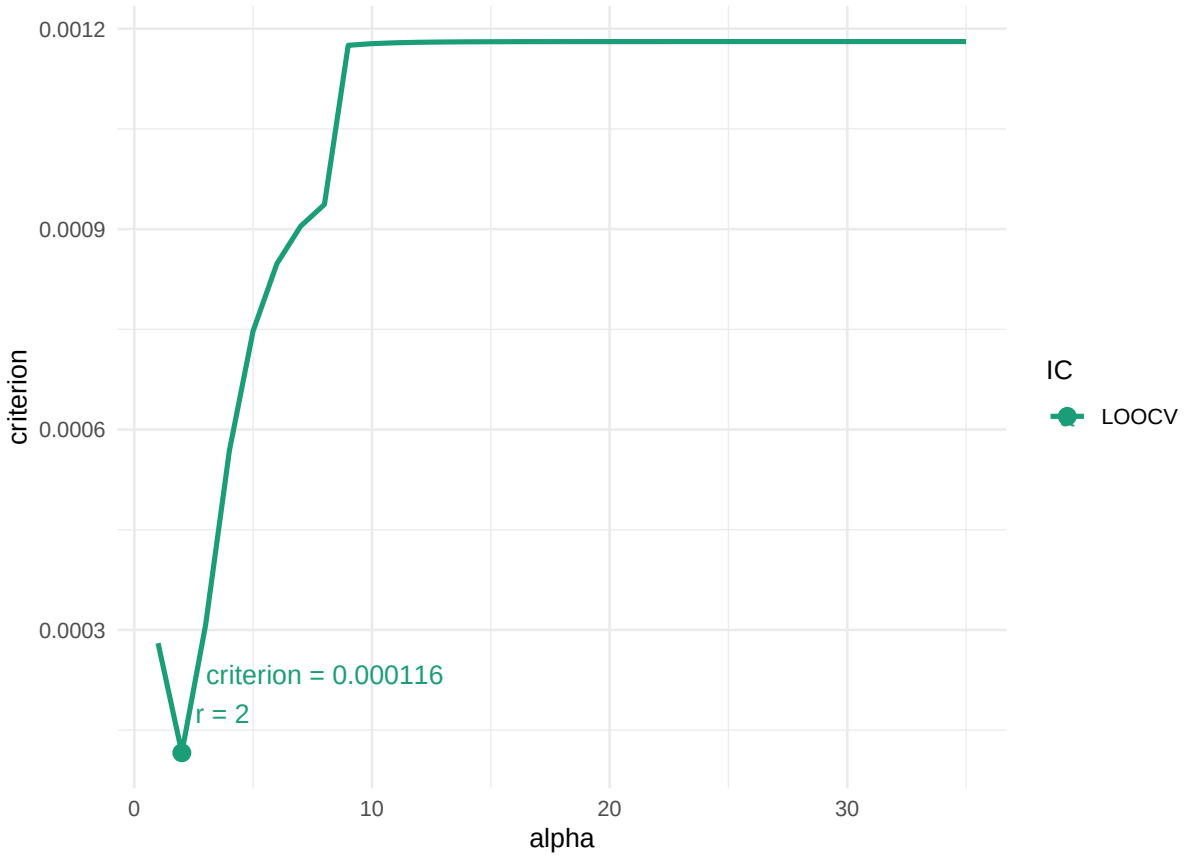

Figure 5: PLS