Econometrics Group Project (ECON 5079)

Regression with Many Predictors

Mohammad Sadra Heydari*| Alina Posrednikova †| Marina Ivanova ‡

November 17, 2023

Adam Smith Business School, University of Glasgow

*2929332h@student.gla.ac.uk
†292251P@student.gla.ac.uk
‡2922514I@student.gla.ac.uk

1 Introduction

In the contemporary era, the proliferation of data has reached unprecedented levels, accompanied by a significant enhancement in the processing power of computers. This technological advancement has empowered researchers to conduct experiments and formulate models that were once deemed unattainable. Consequently, it has opened avenues for the introduction and analysis of intricate models, particularly evident in fields such as economics.

Historically, economists have favored simplicity in their models to elucidate the dynamics of socio-economic systems. However, the surge in available data and computational capabilities has prompted the consideration of more complex models. While this expansion of possibilities is promising, it also introduces inherent risks. Economists navigating this new terrain must exercise caution in model selection to ensure the efficacy of their analyses.

This project is dedicated to exploring the pitfalls associated with employing an extensive set of explanatory variables in an Ordinary Least Squares (OLS) model. In Section 1, we delve into the repercussions of omitted variables on OLS models, employing the Monte Carlo algorithm as our analytical tool. Sections 2, 3, and 4 are dedicated to investigating diverse strategies for eliminating unnecessary variables, aiming to refine the model without sacrificing its accuracy. Finally, in the concluding section, we consolidate our findings to present an overarching perspective on the risks and benefits of employing a broad set of explanatory variables in OLS modeling. Through this endeavor, we seek to contribute valuable insights to the ongoing discourse on optimal model selection in economic analyses.

2 Omitted Variable Analysis Using Monte Carlo

Omitted variable bias is an important issue in econometric analysis; taking it into account is necessary to obtain accurate estimates in econometric forecasting. Omitted variable bias occurs in regression analysis when one or more relevant variables are omitted from the model, leading to biased and inconsistent estimates of the coefficients of the included variables. Excluding significant predictors causes the model to explain the effect of the missing variables on those that were included. That is, the effects of the omitted variable are erroneously attributed to the included variables.

In Monte Carlo, we evaluate the impact of omitted variable bias by comparing a model in which all p predictors are used and one in which the corresponding variable is omitted. By comparing the results of the full model and the model with the omitted variable, we can estimate the degree and nature of the bias.

We will now use Monte Carlo since it involves a random sampling method, which is consistent with our data. Moreover, it makes it possible to reproduce and analyze a wide range of existing scenarios.
For omitted variable bias to exist in linear regression, the following conditions must be present:

1. The true regression coefficient should not be equal to zero.

2. The omitted variable must be correlated with the independent variable.

Consider a linear regression of the following form:

$$Y = X\beta + Z\gamma + U$$

where Y is the dependent variable, X and Z are the independent variables, β and γ are the associated coefficients, and U is the error term. Assuming that Z is not included in the regression, we obtain the least squares estimate $\hat{\beta}(X'X)^{-1}X'Y$. Replacing Y we get the following estimate:

$$\hat{\beta}(X'X)^{-1}X'(X\beta + Z\gamma + U) = \beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'U$$

Since $corr(X, U) = 0$, calculating the conditional expectation, we get $\mathbf{E}[\hat{\beta}|X] = \beta + bias$, where the bias is equal to $(X'X)^{-1}\mathbf{E}[X'Z|X]\gamma$. It is not equal to zero if $X'Z \neq 0$.

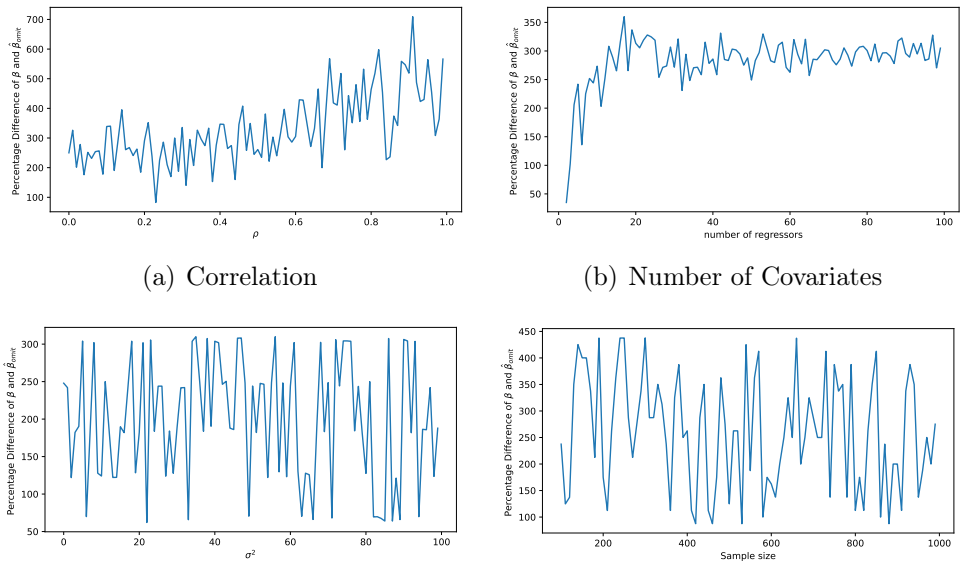
It is important to discuss how model parameters affect bias. To test how the bias value reacts to the changes in model parameters, we are observing "bias percentage difference". The term "bias percentage difference" refers to the percent difference between the true coefficient and the estimated one.

$$BPD = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{\left|\beta_i - \hat{\beta}_i\right|}{\beta_i}\right)$$

An increase in the percentage difference indicates that the model produces estimates that diverge strongly from the true value.

Correlation between regressors (ρ) is the first parameter to be considered. As the correlation between the predictors increases, the probability of encountering multicollinearity increases. Multicollinearity means significant correlations between predictor variables. In the presence of multicollinearity, the exact determination of the individual contribution of each predictor becomes a difficult task, potentially leading to increased bias in coefficient estimates.

It is evident from figure 1(a) that the results are consistent with the understanding that the correlation between predictors can significantly affect the stability



(a) Correlation (b) Number of Covariates

(c) Error Variance (d) Number of observation

Figure 1: Bias Percentage difference of the OLS model with omitted variables for different hyperparameters.

and accuracy of parameter estimates in regression models. Omitting a variable in a highly correlated model leads to instability of estimates. When the correlation between variables is small, the results we obtain having omitted variables are close to the truth.

Figure 1(a) shows that with the growth of ρ, the difference between estimations of the true model and the model with omitted variables increases. The same dynamic is observed for OLS estimations of the true model and the model with omitted variables (figure 2); correlation itself can cause bias. However, figure 3 illustrates that an increase in the correlation coefficient leads to higher values of the model's coefficient of determination. So, in the presence of omitted variables, a high correlation coefficient results in a better fit for the data because all the data can be explained by one variable. In the presence of high correlation, we obtain accurate results.

The second model parameter to be considered is the number of predictors. It can be seen from figure 1(d) that a strict dependence of the bias on the number of regressors cannot be identified. As the number of regressors increases, the dimensionality increases, which can have either a positive or negative effect on the bias. On the one hand, a large number of regressors increases the likelihood of

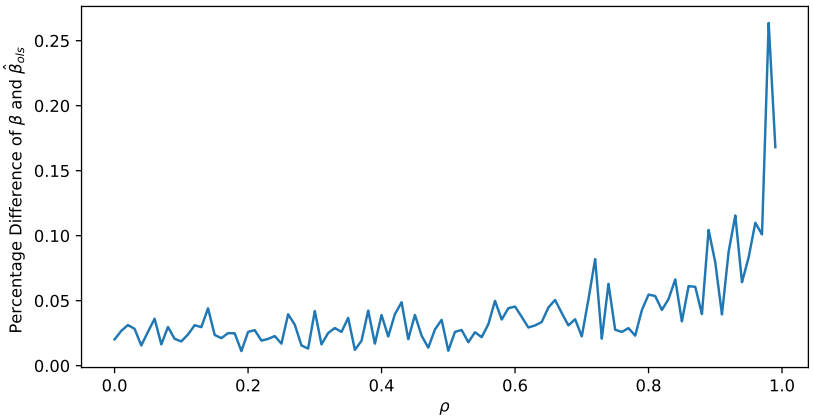


Figure 2: Bias percentage difference for OLS model

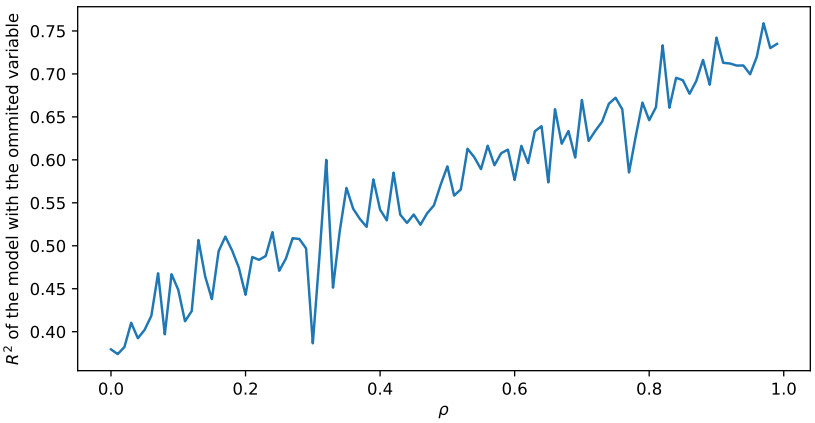


Figure 3: R^2 estimated by the OLS with omitted variables

multicollinearity problems (high correlations between regressors), which can make estimates less stable. On the other hand, when one of the regressors is excluded in a model with a large number of covariates, more variables can take over some of the dependent variables, spreading the bias evenly.

Next, let us consider what effect error variance has on bias. It can be seen from figure 1(c) that changes in variance do not have a significant effect on the bias, but the variance of bias increases with the growth of error variance. It can be concluded that the choice of variance in the Monte Carlo method does not significantly affect the bias. The last parameter to be considered is the number of observations. It is evident from figure 1(d) that sample size, as well as an error variance, does not have a significant effect on the value of bias.

3 Information Theoretic Model Averaging

Information-theoretic model averaging (ITMA) is a statistical method used to select and average regression models. This method consists of combining models and assigning them weights that indicate their probability of being the best model based on their information efficiency, calculated using the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). In cases of model selection uncertainty, ITMA is a good method.

In this section, we select the first 10 variables in our data sets, and like Pesaran and Timmermann 1995, we create models with all possible different permutations of these variables resulting in a total of $2^10 = 1024$ models. Each model is specified with a 10-digit binary number, showing whether each variable is involved in that model or not. For example, '1010000010' refers to a model that contains 1st, 3rd, and 9th variables. We then run OLS to estimate the coefficients and extract features like BIC, and AIC. Then like Kapetanios, Labhard, and Price 2008 we calculate the probability of each model with the difference that instead of using AIC, in this project we used BIC. Therefore the probability of each model would be:

$$\pi_i = \frac{\exp\left(-\frac{1}{2}\Psi_i\right)}{\sum_{j=1}^{2^{10}} \exp\left(-\frac{1}{2}\Psi_j\right)} \qquad \Psi_i = BCI_i - \min_j \{BCI_j\}$$

Note that this formulation is different from one that was suggested in the instruction of the project, that is $\pi_i = \frac{\exp(BIC_i)}{\sum_j \exp(BIC_j)}$ which means as the model performs better, the BIC_i decreases and therefore the π_i decreases and that is the exact opposite of the aim of this method. Additionally, we define the probability of inclusion of a variable such that it is the sum of the probability of all the models that contain that variable.

Dataset	Combination	AIC	BIC	R2	Adj R^2	π
DGP1	1110000000	292.1490	305.3423	0.993	0.993	0.410
DGP2	1110000000	291.0571	304.2503	0.989	0.989	0.382
DGP3	1110000000	287.2928	300.4861	0.982	0.982	0.399

Table 1: Result of ITMA algorithm for data sets generated using correlation coefficient $\rho = 0.9$ and $\sigma^2 = 0.25$. The values are the average of 100 Monte Carlo iterations.

Table 1 shows the result of the model with the highest probability after running the ITMA algorithm. We observe that for our specific hyperparameters, this algorithm was able to identify the true model for all our DGPs. Figure 4 shows the inclusion probability of the first 10 variables in our data sets. As we can see, the probability of inclusion of the true variables is neat to one for our specific DGP setup.

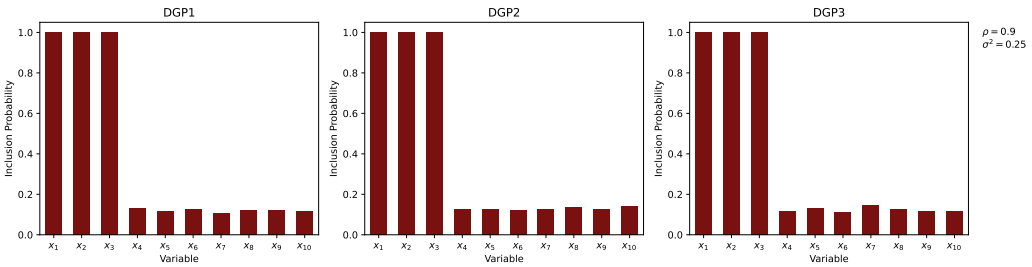


Figure 4: Probability of inclusion for the first 10 variables of our data sets with correlation coefficient $\rho = 0.9$ and $\sigma^2 = 0.25$. The values are the average of 100 Monte Carlo iterations.

The results presented in table 1 and figure 4 suggest that this method is performing well in eliminating the irrelevant variables. In order to check the validity of this statement, we try to analyze the performance of this model for different DGP setups.

In the case of DGP1, it is observed from figure 5 that the probability of the true model does not vary significantly as a result of an increase in the error variance (σ^2). But for the probability of the inclusion of the relevant variables, we can see from figure 6 that it only decreases at high levels of variance. In the cases of DGP2 and DGP3, a high dependence on the magnitude of the error variance is observed (figures 6 & 6(b)). As σ^2 increases, the probability of the true model decreases.

As for the effect of changes in correlation coefficient (ρ) on results for DGP2, it can be noticed that as the correlation grows, ITMA begins to exclude even significant parameters, this trend is especially noticeable with high dispersion (figure 6(c)).

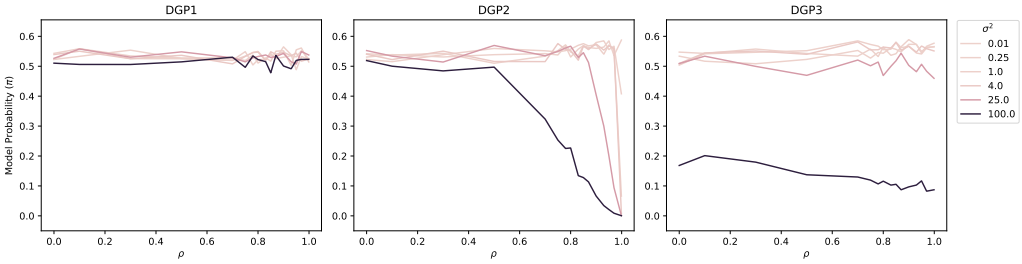


Figure 5: Probability of the true model for different correlation coefficient and error variance

Even when the error variance is small (figure 6(a)), a correlation extremely close to one leads to the exclusion of significant parameters.

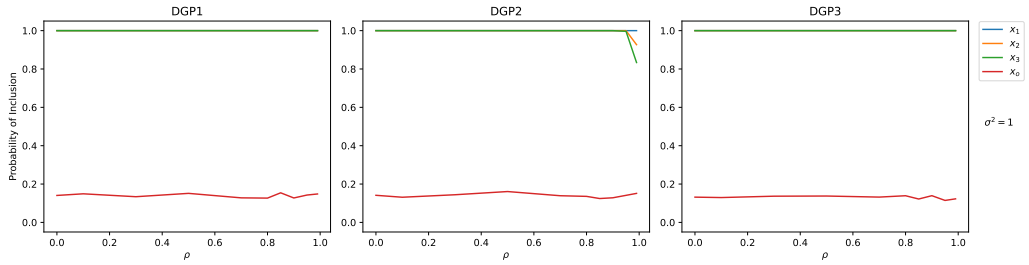
ITMA performs well on DGP3 in low error variance conditions, even if the correlation coefficient is high enough (figure 5). But when the error variance reaches high enough values, the inclusion probability of including variables begins to decrease if the correlation is strong (figure 6(c)). Thus, in DGP3, which has high σ^2, the presence of a significant correlation can lead to lowering the probability of inclusion of relevant variables and increasing the probability of irrelevant variables's inclusion. However, changes caused by changing the correlation coefficient for DGP3 are not as sharp as for DGR2 (figure 5).

In conclusion, ITMA is a great fit for DGP1. Regarding the applicability of this method for DGP2, it is necessary to pay attention to parameters such as error variance and correlation. ITMA is suitable for DGP2 under conditions of low error variance and low correlation. As for DGP3, the ITMA algorithm is suitable for such data only in low error variance conditions.

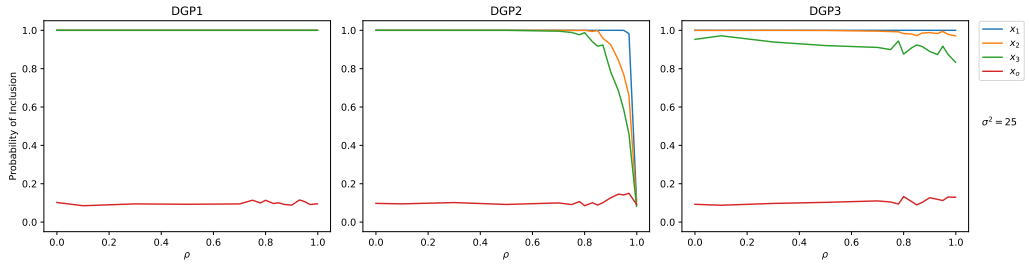
4 Least Abs. Shrinkage and Selection Operator (LASSO)

LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization method in linear regression that performs shrinkage and variable selection to simplify linear regression models and prevent overfitting. The LASSO method applies a penalty on the sum of the modulo values of the coefficients in the model. The LASSO optimization formula can be expressed as follows:

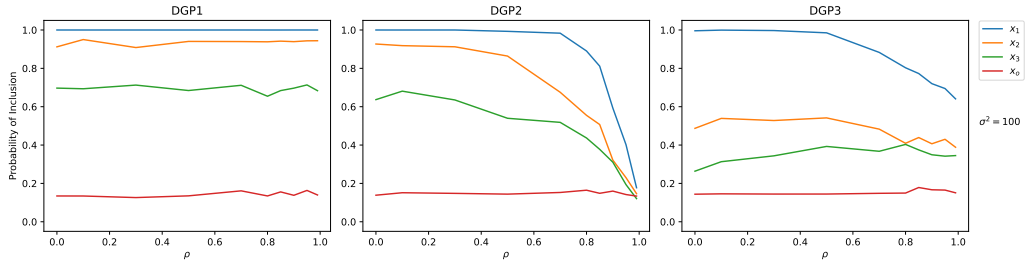
$$\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$



(a) $\sigma^2 = 1$



(b) $\sigma^2 = 25$



(c) $\sigma^2 = 100$

Figure 6: Probability of inclusion for different DGP setups. x_1, x_2, and x_3 are the relevant variables and x_o is the highest probability of all the irrelevant variables. It is observed that as the correlation coefficient increases, the probability of inclusion for true variables decreases.

We can show that if $\lambda \to 0$, then $\hat{\beta}_{lasso} \to \hat{\beta_{ols}}$ and if $\lambda \to \infty$, then $\hat{\beta}_{lasso} \to 0$.

The peculiarity of this regression is that it tends to nullify the coefficients that are more significant through the regularization parameter λ. This parameter plays a key role in LASSO regression. Values of λ denote the strength of the regularization; meaning the higher λ, the more coefficients in the model are equal to zero, leading to a sparser model.

Thus, λ, correctly selected for the research objectives, excludes the least significant variables from the model, thereby not only reducing the computational complexity of the model but also facilitating the process of interpreting the research results. Therefore, parameter λ shows the degree of trade-off between bias and variance of our model. High values allow the model to better generalize data, while low values allow the model to produce more accurate results with minimal bias but greater variance.

The first data generation process (DGP1) is a special scenario in DGP2 in which there is no multicollinearity. This aims to investigate how well the Lasso method performs under conditions where there is no or minimal multicollinearity. Consideration of its performance in different data generation processes can provide insight into its suitability for different scenarios and explain its reliability in dealing with specific conditions. figure 8 shows that the model with zero correlation has a smaller MSE compared to the MSE in models with correlation.

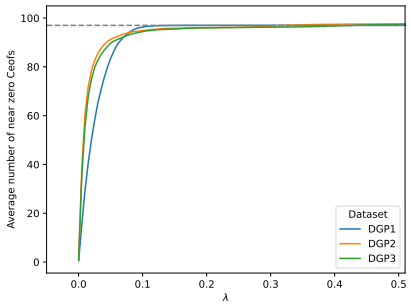


Figure 7: The average number of near-zero coefficients for different regularization parameters (λ). The dashed line in the plot shows the value of 97.

Moreover, it is worth noting that the coefficients obtained for DGP1 using the cross-validation method are close to the true values. From figure 7, it can be seen that as the number of zero coefficients increases, DGP2 and DGP3 show a sharp increase and a faster approach to the value of 97 zero coefficients applicable in the correct model. However, DGP1 reaches the desired lambda value faster compared to other approximations.

Regarding DGP2, our data has collinearity, which leads to some problems with

the Lasso method. Thus, figure 8 shows that as the correlation coefficient increases, Lasso performs worse than when there is no correlation in the model at all. In the presence of multicollinearity, when a number of variables contain similar data, LASSO can select only one variable from those that are highly correlated, turning the remaining coefficients to zero. It is worth noting that LASSO may not always select the same variable for different runs. Moreover, while working with data with a high correlation coefficient, it was observed that growth in variance leads to higher bias (figure 8).

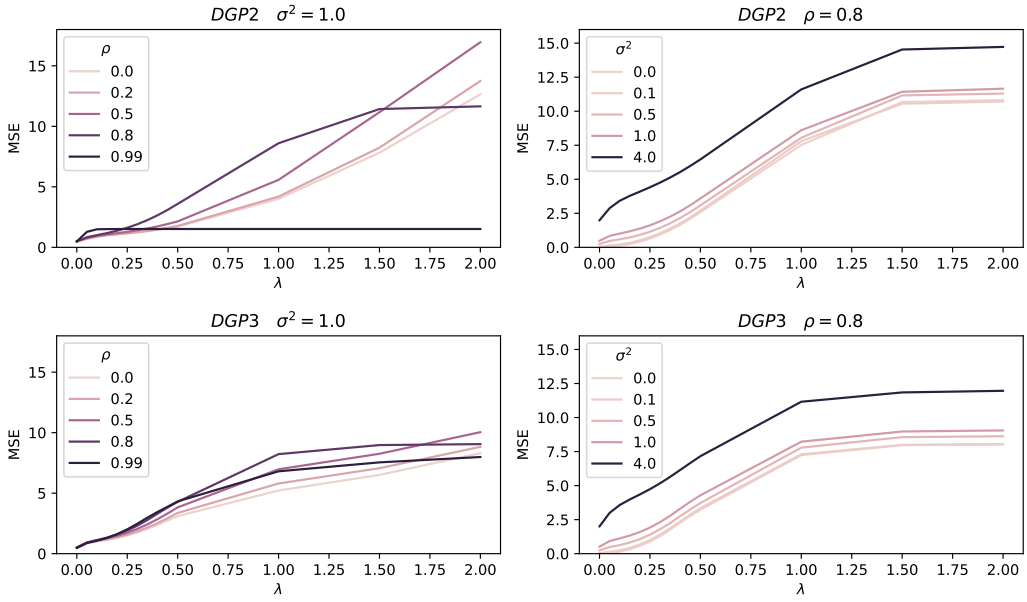
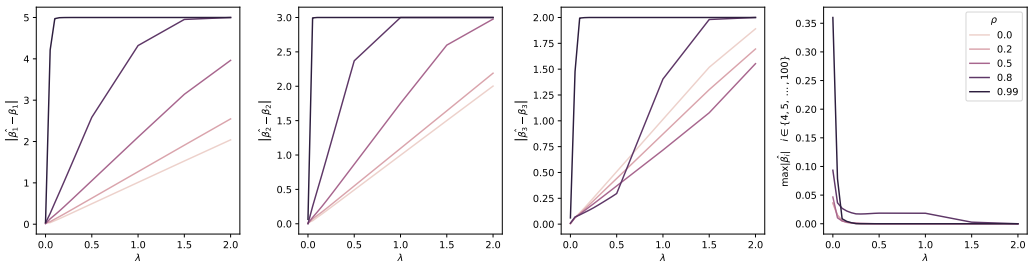


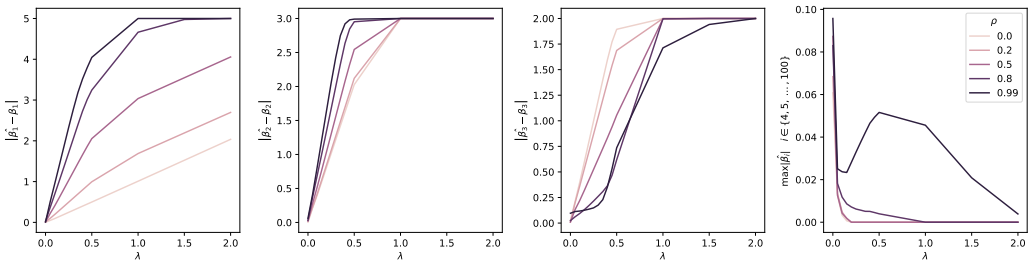
Figure 8: MSE and hyperparameter values for DGP1-3

Thus, the implementation of the LASSO algorithm can be useful in research, especially while working with a large dataset. This method helps to reduce the number of regressors in the model, leaving only the most relevant ones. Increasing the interpretability of the model makes the process of result analysis more clear.

Figure 9 shows the difference between the real value of β and the value of the results of the LASSO model. So, the true models contain three relevant variables β_1, β_2, β_3, which are equal to -5, 3, 2, respectively. When applying LASSO to models using these vectors, it becomes evident that as the regularization parameter λ increases, the associated values tend to converge rapidly to zero, especially for highly correlated vectors. Conversely, this stands as one of the merits of LASSO, as depicted in graph 9, where it is evident that the coefficients corresponding to irrelevant variables decrease rapidly.



(a) DGP2



(b) DGP3

Figure 9: The absolute difference between the real coefficients and the Lasso estimated ones. From left to right, the first three subplots correspond to the relevant coefficients with values -5, 3, and 2 respectively, and the last graph shows the same thing but for irrelevant covariates.

Moreover, when performing LASSO regression with cross-validation, we selected the best model settings i.e λ. It creates a sparse model in which some coefficients are exactly zeros. The cross-validation process divides the dataset into k folds and runs k times, choosing different folders as a test set and performing on the remaining folds as on the training set. 'LassoCV' fits the model with different values of λ. The algorithm chooses the value of λ that results in the best regression performance across all folders. The coefficient of determination is used to evaluate the quality of the model. The final step of 'LassoCV' is training the model on the whole dataset with the usage of the selected λ.Kumar n.d.

Also, it may be useful to standardize the data when using regularization methods. The standardization approach allows the researcher to bring all elements to a single scale, enabling them to use algorithms such as LASSO (Least Absolute Shrinkage and Selection Operator) to obtain high-quality results despite sensitivity to scale.

LASSO implies adding a penalty term to the linear regression cost function. The penalty term affects elements with different weights in the regression with different strengths. Thus, neglecting the standardization stage can lead to a bias in the coefficients obtained using this algorithm, since the impact of the penalty term will be disproportionate.

Regarding the orthogonalization of explanatory variables, although it is found to be useful in dealing with the multicollinearity problem, it is not needed when using LASSO due to its ability to deal with the multicollinearity problem. The penalty term in LASSO contributes to the sparsity of the model by reducing some coefficients to exactly zero.

Using the correlation between variables in LASSO helps to choose between variables when they are highly correlated. Moreover, the orthogonalization of explanatory variables in the LASSO method can lead to:

1. loss of interpretation of variables since orthogonalization often implies linear combinations of the original variables, and therefore makes them less interpretable;

2. an impact on model sparsity. By making new variables, orthogonalization can disrupt LASSO's variable selection process, which involves setting some coefficients to zero;

3. may lead to collinearity problems in the LASSO model.

Thus, orthogonalization of explanatory variables is not necessary since LASSO has the ability to handle correlated variables by using their relationships to select and regularize variables.

In addition, we can use LASSO when we have more predictors than observations, but this can lead to some difficulties, and solving them may require the use of

methods such as cross-validation and the study of alternative regularization methods. Here are some examples of solving the problem "p > n" in LASSO:

- sparse solutions: in conditions when we have more predictors than observations, LASSO wins by its ability to automatically take a subset of predictors, making some coefficients equal to zero;
- size reduction: LASSO hides the sample dimension by selecting the most suitable samples;
- cross-validation for model selection: cross-validation can serve as a useful tool for selecting the optimal regularization parameter (λ) and evaluating model performance even in situations with a large number of predictors.

Thus, LASSO can be effective even in a situation where the number of predictors exceeds the number of observations.

Another advantage of the Lasso method is that "LASSO with preselection" or "LASSO-P" can be useful, for example, when we have a large number of potential predictors. This allows us to make a preliminary selection to reduce the set of variables to the most significant. So, the advantages of preselecting some variables and then implementing the LASSO include:

1. alleviation of computational complexity due to condition preselection;
2. reduction of multicollinearity by preselection, which makes it possible to select less correlated variables before applying LASSO;
3. creating more interpretable variables.

Thus, the preliminary selection of variables before applying LASSO can have a positive effect on the results obtained, especially when dealing with a large number of potential predictors. In conclusion, it can be seen that the Lasso algorithm is well-suited for fitting the coefficients in the models for DGP1 and DGP3. However, this method is not applicable to DGP2, due to the high multicollinearity of the data. Thus, DGP2 can be estimated using Ridge regression (Tibshirani 1996), which takes into account multicollinearity.

5 Principal Component Analysis

Principal component analysis (PCA) is a method of reducing dimensionality with minimal loss of useful information by retaining the first few principal components that explain the most variance, commonly used in statistics and machine learning. It represents an orthogonal linear transformation of a set of correlated variables

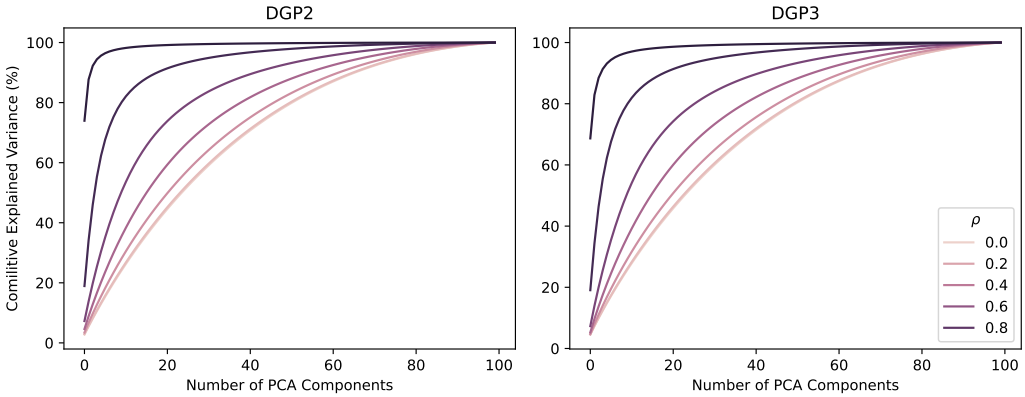


Figure 10: Number of PCA components for explained variance with different ρ

into a new set of uncorrelated variables, called principal components, into a lower-dimensional space. So, when constructing the first axis, the variance around it should be the greatest, and the second, orthogonal to the first, takes over the remaining large variance.

The principal components are ordered as the variance decreases. Each principal component is a vector, and its coefficients, that is, loadings, reflect the contribution or weight of each original variable to the formation of this principal component. The greater the magnitude of the coefficients, the greater the impact of the corresponding variable on the main component.

In this section, we first investigate the effect of correlation in our datasets on how the explained variance reacts as a result of an increase in the number of components. From figure 10, we can see that as the correlation increases for all DGPs, the number of principal components to explain the variance decreases, and therefore, we can explain a larger amount of variance with a fewer number of principal components. For DGP1 ($\rho = 0$) (figure 10), the number of principal components for the optimal level of explained variance is significantly larger than for DGP2 at a relatively large correlation. Thus, it is worth noting that as the correlation coefficient ρ increases, DGP2 and DGP3 behave similarly, reducing the number of principal components required to explain the variance in the data. Therefore, we can conclude that the PCA method is well-suited for a dataset with high correlation.

Now in order to get an estimation of the coefficients of our model, we construct a PCR (Principal Component Regression Model). In doing so, we consider the number of principal components (k) as the hyperparameter of our model. Then, we apply the following steps:

- Create a PCA model to converte $X_{T\times 100} \rightarrow Z_{T\times k}$

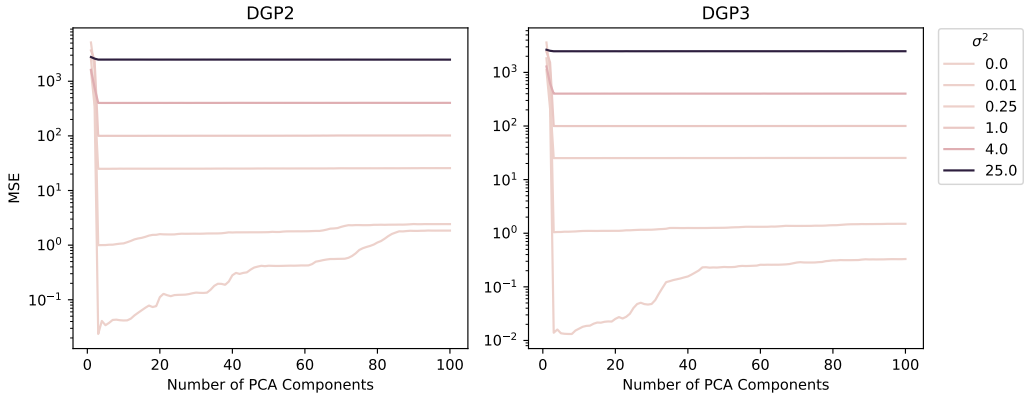


Figure 11: MSE in PCA for different error variance

- We assume that $Y_{T\times 1} = Z_{T\times k}\gamma_{k\times 1} + \epsilon_{T\times 1}$, and we run OLS to estimate $\hat{\gamma}_{k\times 1}$
- We use the inverse transform of PCA and convert $\hat{\gamma}'_{k\times 1} \to \hat{\beta}'_{100\times 1}$

Next, we looked at the MSE estimates for each of the DGPs. It can be seen from graph 11 that as the number of PCA components increases, MSE sharply decreases and remains at a low level. For DGP2, graph 11(a) shows that MSE is lower when error variance is low. Regarding DGP3, the influence of the error variance on our results is the same as that of DGP2; as the variance increases, the MSE value increases (graph 11(b)). However, correlation analysis showed that DGP3 is more sensitive to changes in the correlation coefficient than DGP2. Thus, figure 12 shows that when the number of principal components increases significantly, both DGP face an increase in MSE at a high level of correlation; however, regression based on DGP3 experiences a more significant increase.

It is shown in 13 that the implementation of Principal Component Analysis results in low bias. As the number of principal components increases, bias experiences a sharp decline. However, when the number of principal components increases significantly, it increases the bias. Obviously, for each of the DGPs, a high correlation value results in higher bias.

For extracting principal components from our generated series, highly correlated models such as DGP2 and DGP3 are the most appropriate due to their high explained variance with a small number of principal components.

When comparing the PCA, Lasso, and ITMA algorithms, it can be noticed that for each of the DGPs, Lasso and ITMA performed better. As for the fit of the model, it is evident from the 14(a) that R^2 is higher for Lasso and ITMA than for PCA3/PCA5 on every correlation level. Moreover, lower values of MSE are observed for Lasso and ITMA than for PCA3/PCA5 on every correlation level; however,

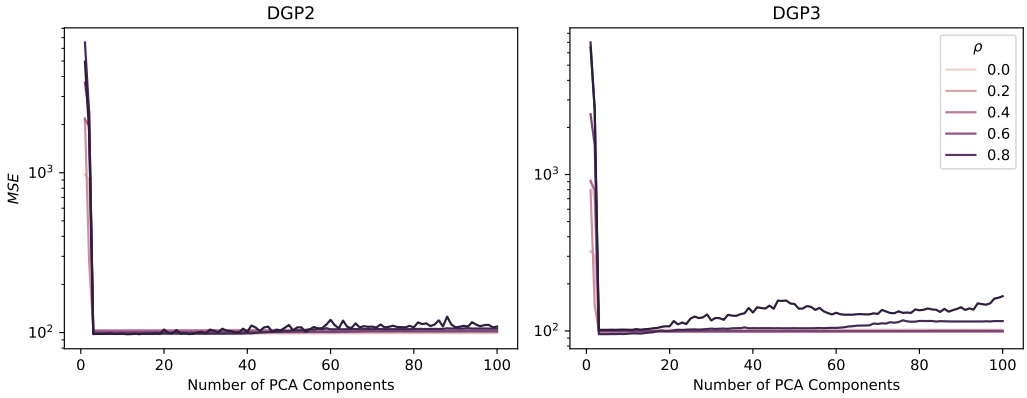


Figure 12: MSE in PCA for different correlation

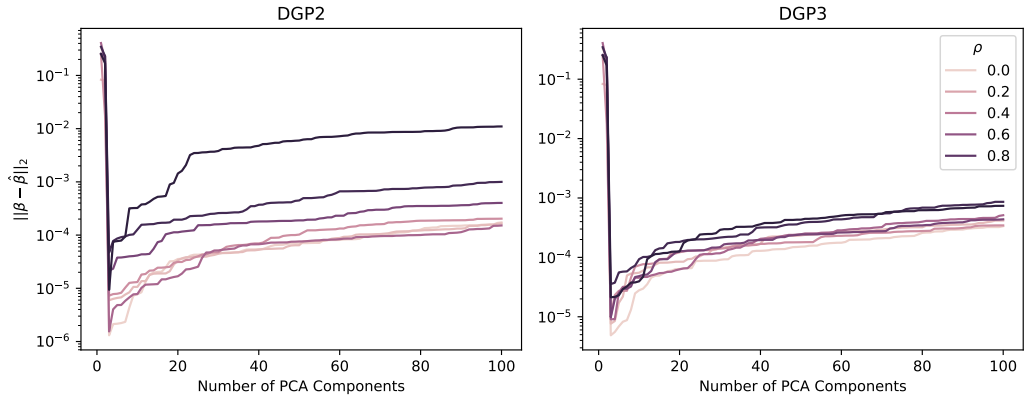


Figure 13: Bias in PCA for different correlation

with the growth of the correlation coefficient, the results for DGP2 become more accurate (figure 14). Regarding the bias, figure 14(c) illustrates that a lower bias is produced while performing Lasso and ITMA algorithms than while performing PCA. From the figure 15 with a smaller scale, it can be seen that the Lasso method produces the lowest bias. Thus, it can be concluded that Lasso implementation leads to more accurate estimations regarding our DGPs.

6 Comparing the Models

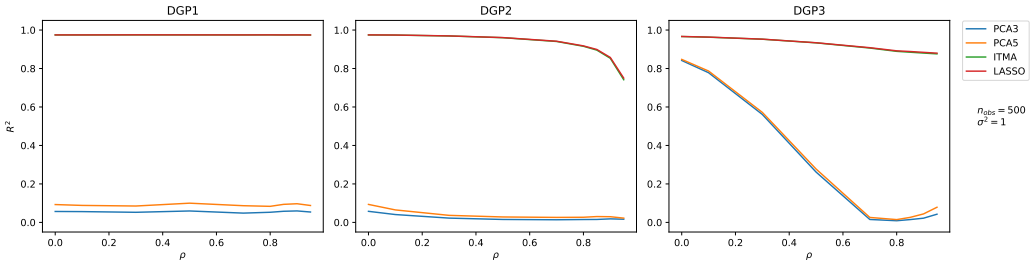
In this section, we aim to compare the power of each model introduced in sections 3, 4, and 5. To do so, we use 4 different models:

- PCA3: A model that uses the first three principal components, and runs an OLS model to estimate the responding coefficients. Then we use the reverse PCA transform to get the estimated values for our original estimations.
- PCA5: A model that uses the first five principal components, and runs an OLS model to estimate the responding coefficients. Then we use the reverse PCA transform to get the estimated values for our original estimations.
- ITMA: A model that uses the first 10 variables of the data and runs the information theoretical model averaging across 2^{10} models and calculates the probability of each model. Then the final estimated coefficient of the model is calculated using the weighted sum of each model combination probability.
- LACCO: A model that uses LassoCV to find the best regularization factor λ.

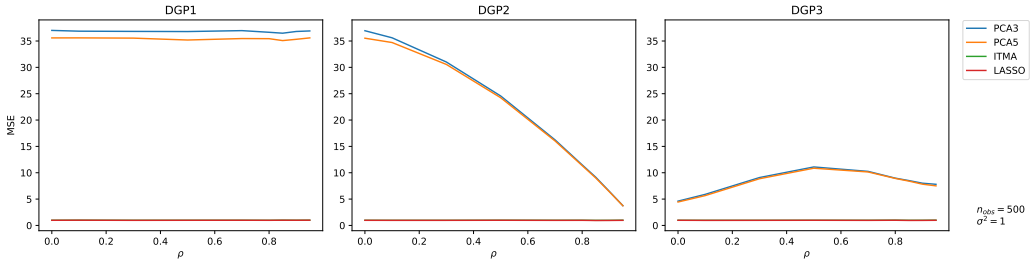
For the sake of comparing the models, we use R^2 as a score of the accuracy of the model and use the Euclidean distance between the real coefficient vector ($[-5, 3, 2, 0, 0, \ldots, 0]^t$) and the estimated coefficient vector. We run each model on a variety of DGP setups in order to check the response of each model to the DGP hyperparameters i.e. the correlation coefficient (ρ), the error variance (σ^2), and the number of observations in each data set (T). Then we use the Monte Carlo algorithm and run each model with 100 randomly generated DGPs and report the average results.

Figure 16 shows that the ITMA is unstable when faced with high error variance and high correlation and in that case, even PCA's are performing better. However, lasso is the most robust and unbiased estimator. This is the case for both DGP2 and 3 but ITMA is less robust in DGP3.

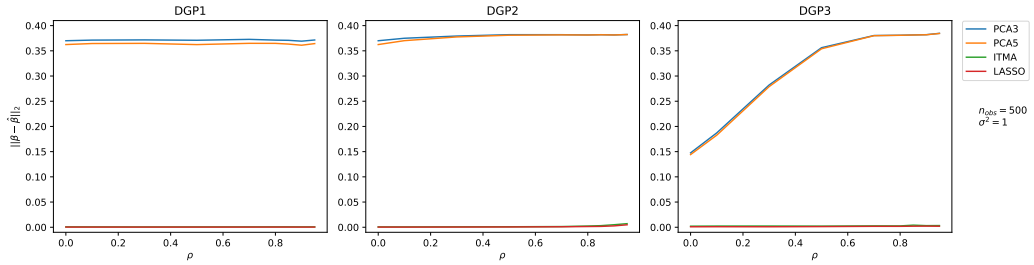
Figure 17(a) shows that for DGP1, the correlation has no effect (obviously as it is not involved in the data generation process of DGP1), and the number of



(a) R^2



(b) MSE



(c) Bias

Figure 14: Results for different correlations.

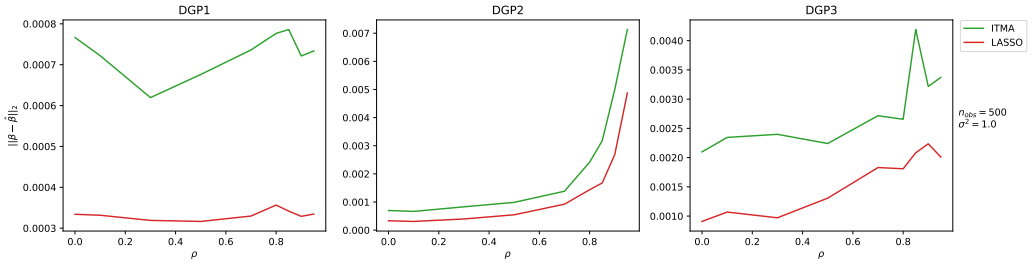
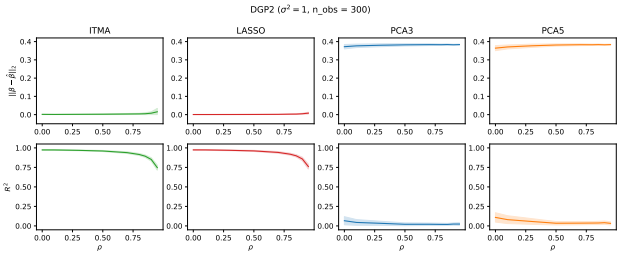
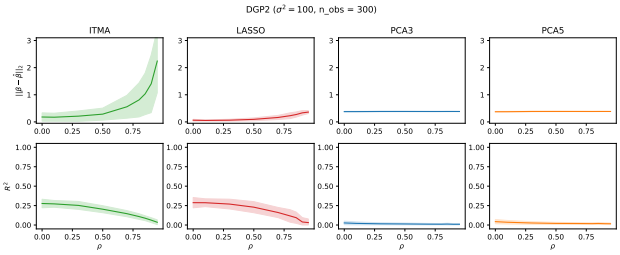


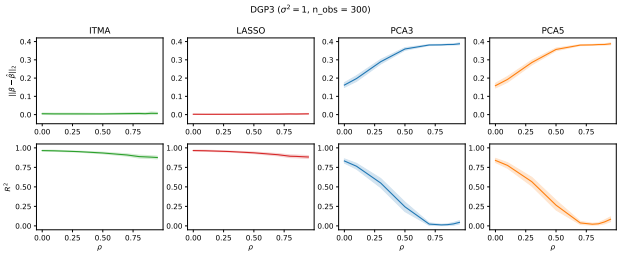
Figure 15: The bias for different correlation for ITMA and LASSO



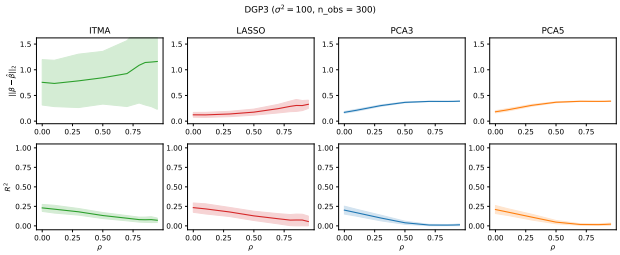
(a) DGP2 ($\sigma^2 = 1$)



(b) DGP2 ($\sigma^2 = 100$)



(c) DGP3 ($\sigma^2 = 1$)



(d) DGP3 ($\sigma^2 = 100$)

Figure 16: The bias of estimated coefficients and R^2 as a metric of in-sample fit across different models with respect to the correlation coefficient and the error variance. The results are drawn from a Monte Carlo algorithm with 100 iterations on DGPs with 300 observations.

observations has almost no influence on the fit of the model. However as the error variance increases, the results for PCA remain bad, and the results of ITMA and Lasso get worse (but with not a significant difference among them).

That is also the case for the bias of our estimator (figure 17(b)), with the difference that the result of ITMA gets much worse and diverges from the real parameters as the error variance increases. Therefore, for DGP1, when we have low error variance, PCAs are out of the picture, and ITMA and Lasso are performing well with no significant difference. However, for high error variance, ITMA diverges, and it is only reasonable to use Lasso.

As for DGP2, the number of observations almost has no effect on the models, and it is similar to the result of DGP1 (which is a special case of DGP2 with $\rho = 0$) with the difference that is when the correlation is high, the quality of ITMA and Lasso estimates decrease (figure 18(a)). Also, from figure 18(b), it can be seen that the results for DGP1 and 2 are close to each other.

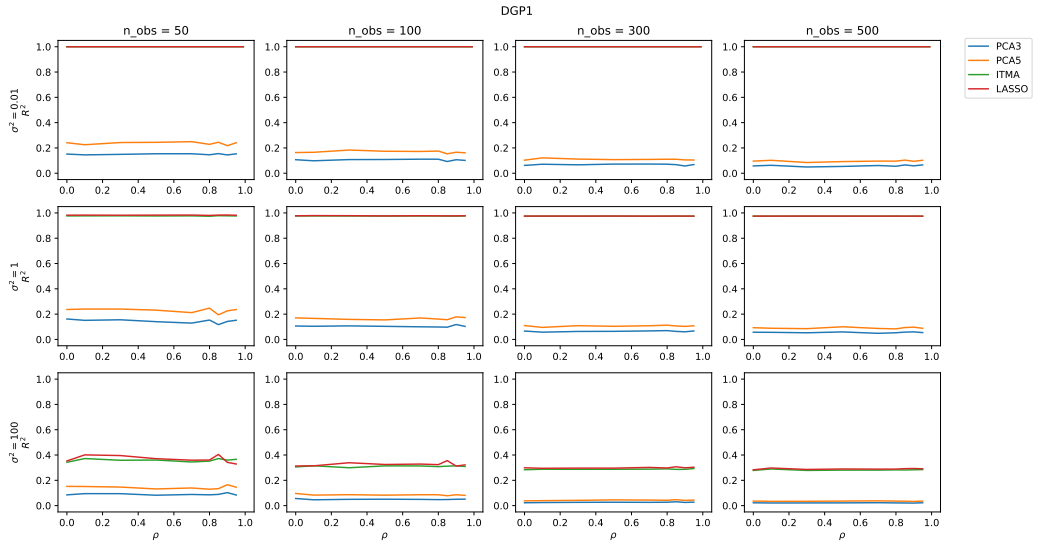
As for the previous two cases, the number of observations (T) has almost no effect on the quality of our model based on DGP3 (figure 19(a)). We can observe that when the correlation increases the quality of all our models decreases (PCA decreases much faster than ITMA and Lasso), and this decrease is much more significant when we are faced with high error variance. As for the bias of our estimators for DGP3, the case is similar to DGP1 and 2, and we can see that ITMA diverges from the real coefficients as the error variance increases (figure 19(b)).

In the process of analyzing these plots, it was observed that when we encounter high error variance, increasing the number of observations has a small effect on the bias of the PCA and Lasso estimators but reduces the bias in the ITMA estimator (figures 17(b), 18(b), and 19(b)). Therefore, it can be concluded that it is only reasonable to consider ITMA when the error variance is low, or the dataset is large.

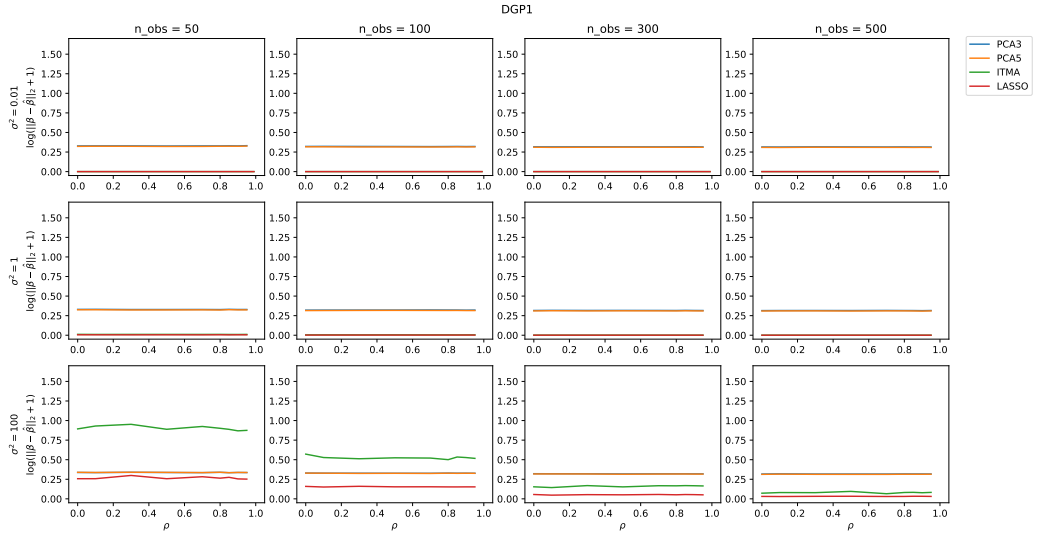
As the results have suggested, the number of observations does not have a significant effect on our model. That may be because of the Monte Carlo algorithm that is used to generate the results. As we aggregate all our results over 100 experiments each time, it means that when, for instance, the number of observations is 50, we are actually running our models with 5000 single data.

In order to solve this problem and investigate the effect of the number of observations on the quality of our models, we observed the variance of the Monte Carlo iteration. Figure 20 shows the result for a correlation coefficient equal to 0.7. Each line in the graphs indicates the average of the Monte Carlo iterations, and intervals show plus and minus one standard deviation ($\bar{x} \pm \sigma$).

Figure 20 suggests that an increase in the error variance would lead to higher variance in the bias of estimators for ITMA and Lasso and has a negligible effect on the result of PCA. Moreover, for all our models, the variance of the result would decrease as the number of observations increases and finally would lead to a more

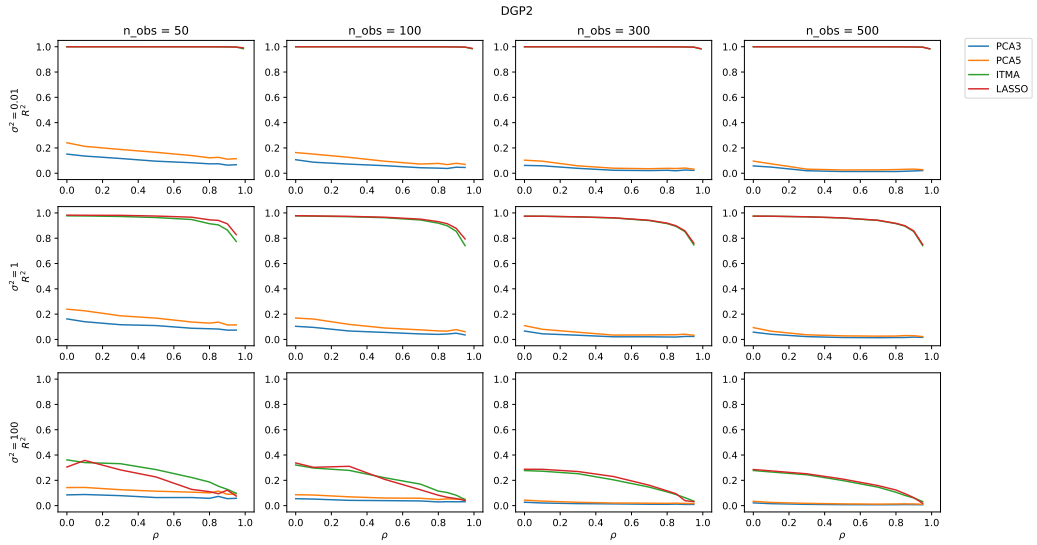


(a) R^2

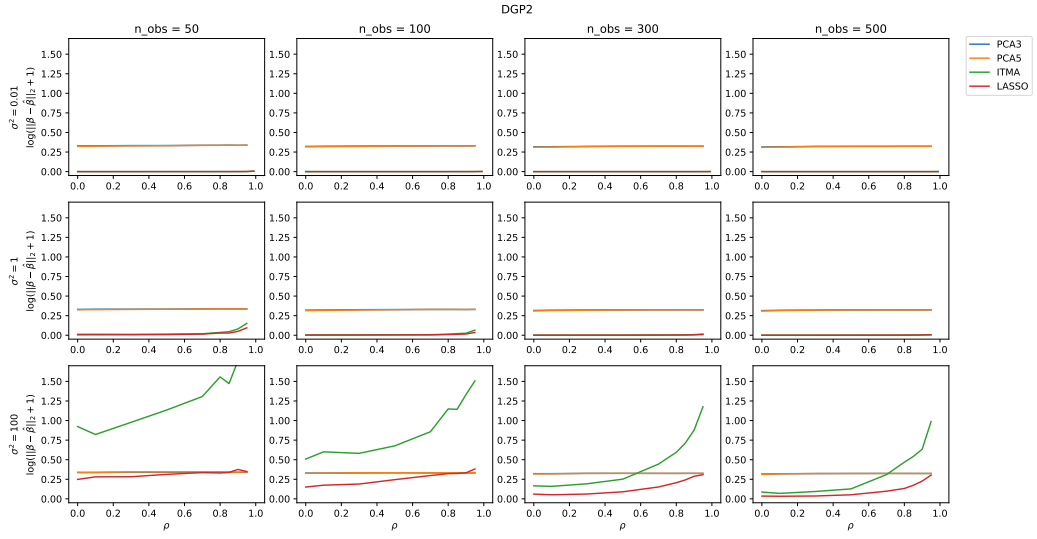


(b) Bias of estimated coefficients

Figure 17: The results of the model on DGP1 across different setups.

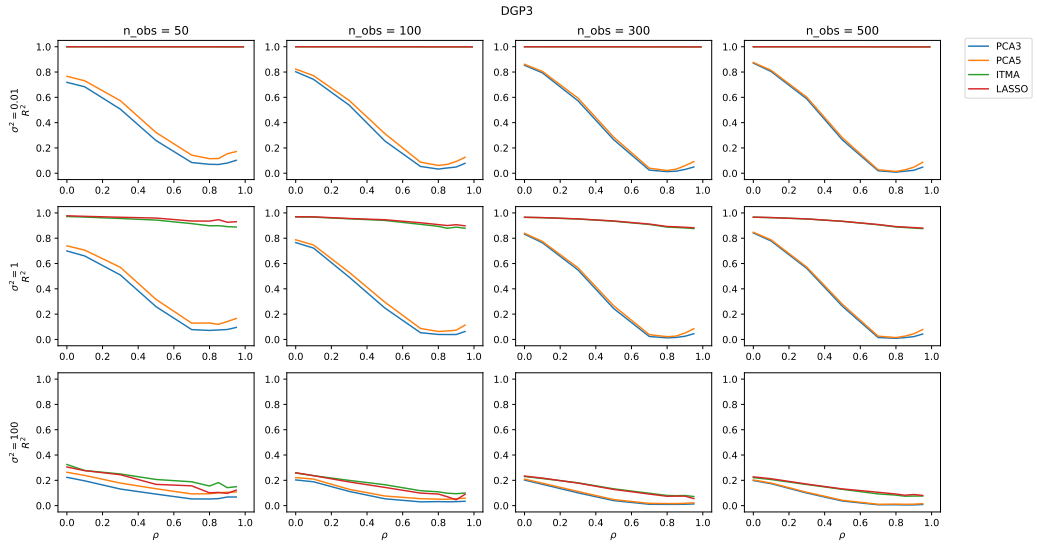


(a) R^2

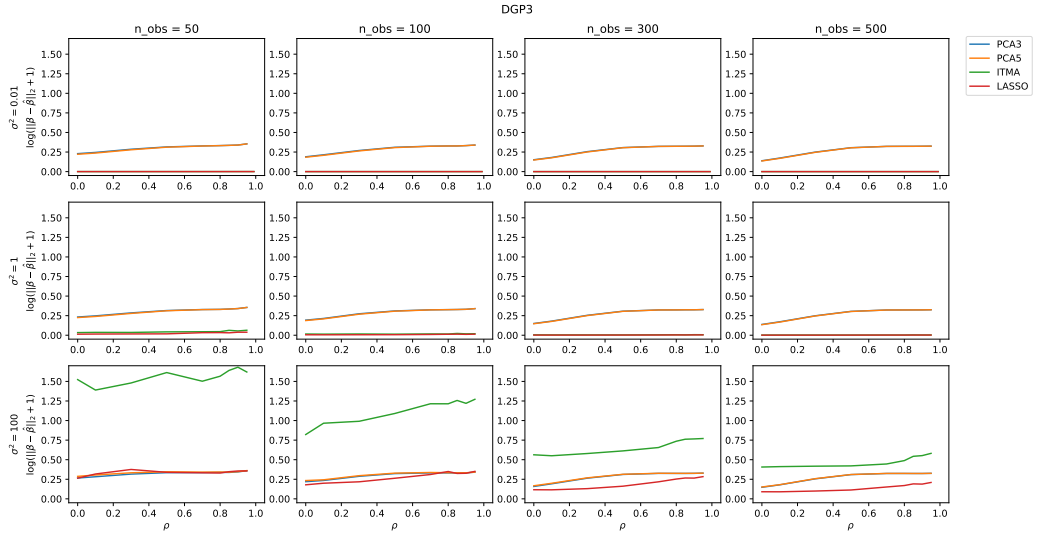


(b) Bias of estimated coefficients

Figure 18: The results of the model on DGP2 across different setups.



(a) R^2



(b) Bias of estimated coefficients

Figure 19: The results of the model on DGP3 across different setups.

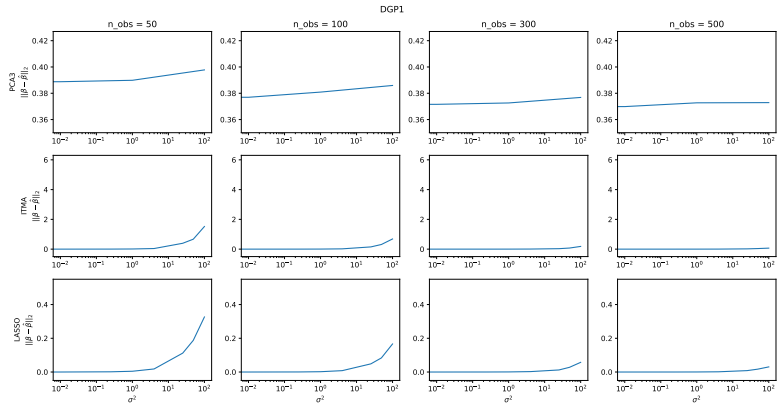
robust model.

In conclusion, results of our analysis suggest that ITMA model performs well for DGP1. When considering its relevance to DGP2, close attention must be given to factors such as error variance and correlation. ITMA proves effective for DGP2 when both error variance and correlation are low. Concerning DGP3, the ITMA algorithm is applicable under the condition of low error variance exclusively.

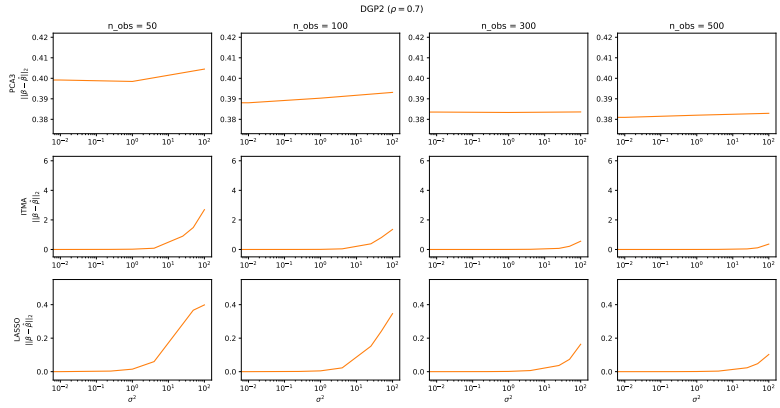
The Lasso algorithm is apt for fitting coefficients in models based on DGP1 and DGP3. Nevertheless, its accuracy diminishes when applied to DGP2, primarily due to the elevated multicollinearity within the data.

When extracting principal components from our generated series, models with high correlation, such as DGP2 and DGP3, appear to be the most suitable due to their high explained variance with a limited number of principal components. However, the PCA statistical procedure exhibits poorer performance when applied to DGP1.

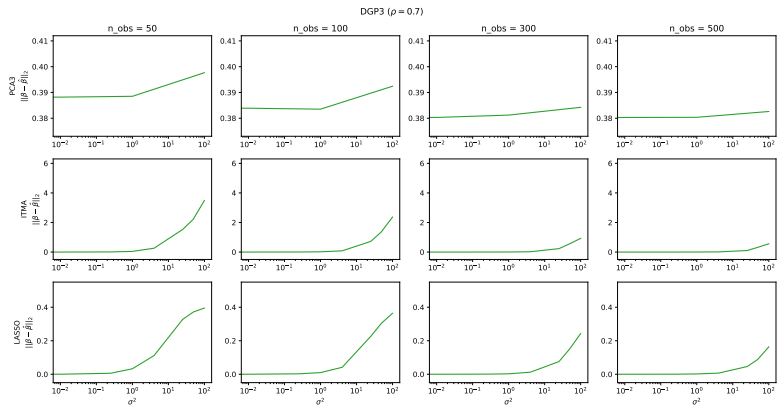
In advising a practitioner on forecasting with numerous predictors, our group recommends employing ITMA. The analysis demonstrates that the model well identifies relevant variables even in the presence of high error variance, with a very low probability of including irrelevant variables (figure 6). ITMA excels at variable identification rather than model accuracy. For instance, in a particular analysis, the true model had a 40% probability, yet the inclusion probability of two out of three relevant variables was almost 100%. However, in instances where ITMA exhibits suboptimal performance, we would suggest considering the use of LASSO.



(a) DGP1



(b) DGP2



(c) DGP3

Figure 20: The effect of the number of observations on variance of estimations.

References

Kapetanios, George, Vincent Labhard, and Simon Price (2008). "Forecasting using Bayesian and information-theoretic model averaging: An application to UK inflation". In: *Journal of Business & Economic Statistics* 26.1, pp. 33–41.

Kumar, Dinesh (n.d.). *A Complete understanding of LASSO Regression*. URL: `https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/`. (accessed: 15.11.2023).

Pesaran, M Hashem and Allan Timmermann (1995). "Predictability of stock returns: Robustness and economic significance". In: *The Journal of Finance* 50.4, pp. 1201–1228.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.