STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, August 2026, pp 1033–1044.
Published online in International Academic Press (www.IAPress.org)

The Factor-Augmented Regression Model (FARM) for Solving the Problem of High-Dimensional Data

Zainab Asaad Jasim *, Emad Hazim Aboudi

Department of Statistics, College of Administration and Economics, University of Baghdad Iraq, Baghdad

Abstract High-dimensional data analysis is a sub-topic of modern statistical applications that has become more significant when the number of explanatory variables grows bigger than the sample size. Such scenarios tend to make classic regression methods ineffective in the form of multicollinearity, high variance, and lack of the ability to estimate the parameters. This paper explores the effectiveness of the Factor-Augmented Regression Model (FARM) as an effective way of handling the issues of high-dimensional data. The given model splits the explanatory factors into latent variables and idiosyncratic elements that allow reducing the dimension but do not lose the necessary structural data. To maximize the prediction accuracy and the selection of the variables, the penalized estimation methods are used. The methodology is analyzed and approximated under the extensive simulation experiments in a high-dimensional data model, with varying sample sizes to assess the predictive performance. In addition, the model is used on real-world health-related data on body composition in order to determine its usefulness on enhancing the predictability of the model and to determine important predictors. The findings prove that Factor-Augmented Regression Model (FARM) is stable in its estimates and better predicting models than the conventional regression techniques in high dimensional situations. These results verify the appropriateness of the model in scientific applications that are based on data that are complex and large in scale.

Keywords FARM, High-Dimensional, Dimensionality reduction, Predictive performance

DOI: 10.19139/soic-2310-5070-3767

1. Introduction

Modern data is becoming bigger and more complex with the mushrooming of scientific and technological disciplines. As a result, the issue of data of high-dimension has become one of the most critical problems of interest among the researchers of statistics, economics, medicine, and other applied sciences. A high-dimensional data set is a scenario where the number of explanatory variables (p) is larger than the sample size (n), and this can be accompanied by severe complications in the statistical analysis, including multicollinearity, large estimation variance, and variable estimated parameters. In this case, it is no longer possible to use traditional regression methods to address these data structures.In recent years, the rapid growth of high-dimensional data, where the number of predictors exceeds the sample size $(p > n)$, has posed significant challenges for traditional regression models. Classical methods often suffer from multicollinearity, overfitting, and unstable estimates, which limit their applicability in modern data analysis. To address these issues, factor-based models have been widely adopted to capture latent dependence structures among predictors while reducing dimensionality.
Thus, the recent studies are directed to creating techniques that combine the dimensionality reduction with the selection of variables that should enhance the accuracy of the estimation and the prediction. The Factor-Augmented Regression Model (FARM) that combines factor analysis with regression analysis is among the most significant of these methods. The model identifies a number of latent factors which capture the prevalent variation across the explanatory variables and they thus scale down the dimensionality without compromising on the critical information

*Correspondence to: Zainab Asaad Jasim (Email: zainab.asaad1801b@coadec.uobaghdad.edu.iq. Department of Statistics, College of Administration and Economics, University of Baghdad Iraq, Baghdad.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

within the original data. The FARM model is applied in two stage process. During the first step, the latent factors are estimated through the use of statistical procedures like Principal Component Analysis (PCA). The second stage is regularization and variable selection which is used to estimate the regression coefficients and the most influential explanatory variables that influence the dependent variable. This is required to enhance the predictive capability of the model and its interpretability. The scientific value of the current research paper can be seen in the extension of the estimation model of the FARM model based on incorporating the current regularization models into a single framework. Specifically, the paper implements LASSO and Adaptive LASSO, and compares their performance, in the context of estimation accuracy, ability to select variables, and stability of the model when used in various high-dimensional conditions. The significance of the present study lies in incorporating the Adaptive LASSO method into the FARM model whereas the traditional LASSO technique is good in reducing the coefficients and choosing the variables, it might be biased in estimating the truly significant coefficients. Conversely, Adaptive LASSO owns Oracle Property, which allows it to discover the real nonzero variables all the time and forecast their coefficients much better. Thus, it is hoped that using Adaptive LASSO as part of the FARM system would offer a more effective and more robust model of high-dimensional data analysis. The proposed study will explore the performance of the FARM model on the LASSO and Adaptive LASSO estimation models on simulated and real high-dimensional data. The statistical measures applied to assess the methods are the Mean Squared Error and the coefficient of determination to determine the most suitable method. The results are anticipated to show that FARM-PCA + Adaptive LASSO model has a better predictive accuracy and explanatory power than the other methods used in this study, thus is an appropriate tool in practice where multidimensional and complex data are used.

The proposed model can be applied in several fields involving high-dimensional data, including healthcare, genomics, financial analysis, and biomedical studies, where the number of explanatory variables is typically larger than the sample size. The main contributions of this study can be summarized as follows:

1. This paper applies the Factor-Augmented Regression Model to effectively handle high-dimensional data by decomposing predictors into latent factors and idiosyncratic components.
2. The study integrates penalized estimation methods, including LASSO and Adaptive LASSO, to enhance variable selection and improve prediction accuracy.
3. Extensive simulation experiments are conducted to evaluate the performance of the proposed model under different sample sizes.
4. The proposed approach is applied to real-world health-related data to demonstrate its practical applicability.
5. The results show that the proposed FARM-based model outperforms traditional regression methods in terms of stability and predictive performance in high-dimensional settings.

The Factor-Augmented Regression Model has been used previously in a limited manner and with restricted estimation methods. In 2022, researchers [1] introduced a factor-augmented regression model that combines the idiosyncratic component and sparse regression as special cases. They also provided theoretical guarantees for model estimation under the presence of quasi-Gaussian and heavy-tailed noise. When comparing the model's results with latent regression and sparse regression, the model demonstrated effectiveness and robustness. In the same year, researcher [2] proposed a factor-augmented linear model to address the issue of rotation indeterminacy in panel data models. The researcher also proposed new estimation methods to improve the identification of latent factors, where the estimation techniques utilize internally generated instruments. The results from simulation studies and empirical applications using data from school and university students showed that the proposed methods provide accurate estimates compared to traditional approaches. In 2023, researchers [3] reviewed the dynamic factor-augmented regression model in their study, where it was used to handle high-dimensional data. This model incorporates dependent noise, and LASSO regularization techniques were applied to reduce the effect of inactive components. The model was tested on real economic data and demonstrated high predictive accuracy compared to traditional methods. As for previous studies that addressed the problem of high-dimensional data, they are as follows:

In 2024, authors [4] proposed the FARMR model to solve the data high-dimensional problem of matrices. The research aimed at coming up with effective algorithms used to estimate latent factors and regression coefficients to improve predictive accuracy and model stability. The model was applied in real-data testing, and in this case, it

had better performance, more interpretable, and gave more understandable relationships between variables. The model is believed to be a powerful and sophisticated tool of managing multivariate and high-dimensional data. In the context of regularization methods, recent work by Yi He (2024) investigated Ridge Regression under dense factor-augmented models and demonstrated that ridge-type estimators remain effective when predictors exhibit strong correlation structures. This supports the use of shrinkage methods alongside latent factor modeling for improving predictive performance [5]. Moreover, Faridoon Khan and Olayan Albalawi (2024) studied fat big data using factor models with penalization techniques through Monte Carlo simulation and real applications. Their results confirmed that combining factor decomposition with variable selection methods provides better prediction accuracy and model stability than traditional approaches [6]. More recently, Cai et al. (2025) proposed Matrix-Factor-Augmented Regression, extending the classical Factor-Augmented Regression Model to matrix-valued predictors. Their study showed that factor-augmented methodologies remain highly flexible and suitable for modern structured high-dimensional datasets [7]. The scientific contribution of the present research is in its expansion of the framework of Factor-Augmented Regression Model (FARM) to the implementation of contemporary regularization techniques, i.e. LASSO and Adaptive LASSO, in the one-dimensional framework of data analysis on the high-dimensional level. Besides, the study offers an extensive comparative analysis of such methods in various conditions when it comes to the sample size and the amount of explanatory variables. The methods are tested with respect to predictive accuracy and model stability. Another significance of the study is that the research fills a gap in the literature by introducing the Adaptive LASSO technique as part of the FARM model, which is a problem that was not given much consideration in earlier statistical research works. The hybrid FARM-PCA-based models were analyzed and tested both on simulation experiment and real health data applications as well as compared with the conventional methods. The findings indicated that the FARM-PCA + Adaptive LASSO model had the best overall performance as it generated the lowest Mean Squared Error (MSE) and highest explanatory power R^2, which qualifies it to be the best model to analyze high-dimensional data.
Recent studies have emphasized the importance of combining dimensionality reduction with regularization methods to improve prediction accuracy in high-dimensional data analysis.

1.1. Theoretical Comparison Between Methods Used in High-Dimensional Data

High-dimensional data problems have become one of the most important challenges in modern statistics and data analysis, particularly when the number of explanatory variables exceeds the sample size $p > n$. In such situations, classical regression methods become unstable and suffer from multicollinearity and weak predictive performance. Therefore, several regularization and dimension reduction techniques have been developed, including LASSO, Ridge Regression, Elastic Net, and the Factor-Augmented Regression Model [1] [5]. The LASSO method applies an $L_1 - penalty$, which shrinks some coefficients exactly to zero, thereby performing automatic variable selection. For this reason, it is particularly suitable for sparse models. However, LASSO may become unstable when explanatory variables are highly correlated [8]. On the other hand, Ridge Regression uses an $L_2 - penalty$, which shrinks coefficient values without completely removing them. This method is more stable in the presence of strong multicollinearity, although it does not perform variable selection directly [9]. Elastic Net combines both L_1 and $L_2 - penalties$, benefiting from the advantages of both LASSO and Ridge Regression. It is especially effective when groups of predictors are highly correlated [10]. In contrast, the Factor-Augmented Regression Model differs from the previous methods by decomposing predictors into latent common factors and idiosyncratic components. This approach helps reduce dimensionality and handle complex correlation structures before constructing the regression model. Therefore, FARM is more suitable for datasets containing strong latent structures and highly correlated variables [1]. In addition, Kernel Ridge Regression was considered as a benchmark comparison method due to its ability to model nonlinear relationships by combining kernel methods with Ridge Regression, making it appropriate for certain high-dimensional data applications [2].

1.2. Research problem

With increased number of explanatory variables, the model experiences the problem of parameter estimation and low results accuracy since the variance increases. Besides, when the sample size is smaller than the number of the

explanatory variables $(p > n)$ then the problem will be more intricate with respect to the choice of the important variables as well as having a dependable predictive capability.

1.3. Research Objective

The primary objective is to use the Factor-Augmented Regression Model (FARM) to address the problem of high-dimensional data and to compare it with commonly used methods for handling this issue through standard evaluation criteria, in order to assess its efficiency.

1.4. Research Importance

This study is significant because the given problem of high-dimensional data handling is treated with the use of Factor-Augmented Regression Model (FARM) since it allows reducing the dimensionality and improving the estimation accuracy of the traditional models. The study also aids in the support of applied research in other areas where data are high dimensional like in economics, medicine, and engineering.

1.5. Research Hypothesis

It is theorized in the study that the Factor-Augmented Regression Model (FARM) outperforms the regularized regression models and commonly applied dimensionality reduction techniques in estimating and predicting the outcome in high-dimensional data, particularly when the number of explanatory variables available is larger than the sample size.

1.6. Research Gap

The research gap is that there are few studies that have adopted a thorough methodological comparison of the FARM model with the widely utilized models to manipulate high dimensional data, that is, regularized regression model and dimensionality reduction methods. In addition, the majority of the studies have not analyzed the effectiveness of the models in various simulation conditions that demonstrate how sample size and the number of explanatory variables affect accuracy of the estimates and prediction.

2. Methodology

The proposed study will take a comparative quantitative study that has two main parts, simulation part and applied part. The simulation aspect produces high-dimensional data in various conditions with different sample sizes and variables, to test the behavior of the models in controlled conditions. High-dimensional health data are employed in the applied component, which implies a large number of explanatory variables with respect to the sample size. Factor-Augmented Regression Model (FARM) is thereafter implemented on these data and contrasted with regularized regression models and dimensionality reduction methods, and the objective of reviewing the performance of the models.The tuning parameter in penalized methods such as LASSO was selected using cross-validation (CV), which provides an optimal trade-off between model complexity and prediction error. Although penalization may introduce estimation bias due to coefficient shrinkage, it improves model stability and reduces variance in high-dimensional settings. Lastly, the performance of the models in the two components is compared through statistical analysis measures, that is, the Mean Squared Error (MSE) and the coefficient of determination R^2, with the aim of determining which model is most accurate and is the most efficient in terms of estimation and prediction.

2.1. Factor-Augmented Regression Model

Factor-Augmented Regression Model is a result of the latent factor regression model in conjunction with the sparse regression model. Latent factor regression model has the following form [1]:

$$x = \theta f_t + u_t \tag{1}$$

The sparse regression model takes the following form:

$$Y = x^T\beta + \varepsilon \tag{2}$$

By substituting (1) into equation (2):

$$Y = f_t\gamma^* + u_t\beta + \varepsilon \tag{3}$$

The main aim of this section is to use a regularized estimation procedure of our latent factor-augmented linear model and explore the same statistical characteristics.
Suppose that we observe independent and identically distributed (i.i.d.) random samples $\{(x_t, Y_t)\}_{t=1}^n$ of (x, y) that satisfy the following:

$$Y_t = f_t^T\gamma^* + u_t^T\beta + \varepsilon_t \tag{4}$$

where $f_1, ..., f_n \in R^k, \varepsilon_1, ..., \varepsilon_2 \in R$ are independent and identically distributed (i, i, d) and (p) represents the number of explanatory variables.
We rewrite (4) in a more precise and compact matrix form as follows:

$$\begin{aligned} X &= F\theta^T + U, \\ Y &= F\gamma^* + U\beta + \varepsilon \end{aligned} \tag{5}$$

where $X = (x_1, ..., x_n)^T$, $F = (f_1, ..., f_n)^T, U = (u_1, ..., u_n)^T, Y = (Y_1, ..., Y_n)^T, \varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ we suppose we can only get access to observations.$\{(x_t, Y_t)\}_{t=1}^n$, Both F and U are latent factors, and must be estimated using the observed predictors X. Therefore, we will start by showing how to estimate F and U and then follow through with the regularized estimation of model (5).

2.2. Estimation of the Parameters of the Factor-Augmented Regression Model (FARM)

The estimation of the parameters of the Factor-Augmented Regression Model (FARM) is conducted in two principal steps:

1. **Estimation of Factors in the Factor-Augmented Regression Model (FARM)**: The Principal Component Analysis (PCA) method is the most commonly used approach for estimating latent factors.
   Since only the vector x is observable, the latent factor f and the loading matrix θ cannot be identified under model (5). To make the representation more specific, for any non-singular matrix, we perform the following transformation [11] [12]:

   $$x = \theta f + u = (\theta S)(fS^{-1}) + u \tag{6}$$

   That is $F \to FS, \quad \theta \to \theta S^{-1}$ To solve this problem, the following selection criteria were imposed [11] [12] [13]:

   $$1 - Cov(F) = \frac{1}{n}F^TF = I_k \tag{7}$$

   Accordingly, the constrained least squares estimators of (F, θ) are obtained based on Xas follows:

   $$(\hat{F}, \hat{\theta}) = \arg\min ||X - F\theta^T||_F^2 \qquad , F \in R^{n\times k}, \theta \in R^{d\times k} \tag{8}$$

   Subject to $\frac{1}{n}F^TF = I_k$, $\theta^T\theta$ is a diagonal matrix
   As a result of this transformation, the columns of $\hat{F}/\sqrt{n}$ are the eigenvectors corresponding to the largest K eigenvalues of the matrix XX^T and θ is estimated in the following form:

   $$\hat{\theta} = (\hat{F}^T\hat{F})^{-1}\hat{F}^TX \tag{9}$$

   $$\hat{\theta} = n^{-1}\hat{F}^TX \tag{10}$$

   Then, the least squares estimates of U are obtained using the following formula:

   $$\hat{U} = X - \hat{F}\hat{\theta}^T \tag{11}$$

   $$\hat{U} = (I_n - n^{-1}\hat{F}\hat{F}^T)X \tag{12}$$

   Assume that:

$$\hat{P} = n^{-1}\hat{F}\hat{F}^T \quad (13) \qquad \hat{U} = (I_n - \hat{P})X \quad (14)$$

2. **Estimating parameters**: To estimate the parameters of the Factor-Augmented Regression Model (FARM), the following methods are used:

I. **LASSO Estimator** In 1996, Tibshirani introduced a penalty function of regression models called LASSO (and this is abbreviated (Least Absolute Shrinkage and Selection Operator)) with the objective of not only estimating regression models but also of selecting variables . The LASSO technique is founded on the reduction of the sum of squared errors, which contributes to the achievement of higher accuracy of the estimation and minimization of variables. Applying this approach to the factor augmented regression model one has as follows the LASSO estimator [14] [15]:

$$\hat{\beta}_{Lasso} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p) \quad (15) \qquad \hat{\gamma}^*_{Lasso} = (\hat{\gamma}^*_1, \hat{\gamma}^*_2, \ldots, \hat{\gamma}^*_k) \quad (16)$$

It is obtained according to the following formula [1]:

$$(\hat{\beta}_\lambda, \hat{\gamma}) = \arg\min \left\{ \frac{1}{2n}(Y - \hat{U}\ \beta - \hat{F}\gamma)'(Y - \hat{U}\beta - \hat{F}\gamma) + \lambda \sum_{j=1}^{p} |\beta_j| \ \right\} \quad (17)$$

Assuming that $\tilde{Y} = (I_n - \hat{P})Y$ Represents the residuals of the response vector Y after removing the effect of the latent factors, where $\hat{P} = n^{-1}\hat{F}\hat{F}^T$ is the projection matrix onto the space of latent factors, As previously mentioned, $\hat{U} = (I_n - \hat{P})X$, and therefore $\hat{F}^T\hat{U} = 0$
From which the above equation can be rewritten in the following form:

$$\hat{\beta}_\lambda = \arg\min \left\{ \frac{1}{2n}(\tilde{Y} - \hat{U}\ \beta)'(\tilde{Y} - \hat{U}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \ \right\} \quad (18)$$

where:

- $\hat{\beta}$: Vector of regression coefficients estimated using the LASSO method.
- $\arg\min$: Indicates the values of the parameters that minimize the objective function.
- (n): Sample size (number of observations).
- $\tilde{Y}$: Adjusted response vector after removing the effect of latent factors.
- $\tilde{U}$: Adjusted matrix of explanatory variables after removing the effect of latent factors.
- β: Vector of regression coefficients to be estimated.
- $\|\tilde{Y} - \tilde{U}\beta\|^2$: Residual Sum of Squares (RSS), which measures the difference between the observed and predicted values.
- λ: Regularization (penalty) parameter that controls the degree of shrinkage applied to the regression coefficients.
- $\sum_{j=1}^{p} |\beta_j|$: LASSO penalty term, representing the sum of the absolute values of the regression coefficients.
- (p): Number of explanatory variables.

And

$$\hat{\gamma} = (\hat{F}^T\hat{F})^{-1}\hat{F}^T Y = \frac{1}{n}\hat{F}^T Y \quad (19)$$

Where (λ) is the penalty parameter, also referred to as the regularization parameter [8]. Lasso method would be a better choice, because of its desirable properties that remain even after the selection of the variables. In particular, it pushes some of the regression coefficients to zero and derives the remaining coefficients by a specified degree of shrinkage all at minimum of the loss term. It means that, Lasso estimates are capable of providing a sparse estimation of regression coefficients, making the model easier to interpret [15].

II. **Adaptive Lasso Estimator** The Adaptive Lasso estimator was introduced by Hui Zou in 2006 [16] [17] in an attempt to create a Lasso model with what is known as the Oracle Properties. This is done through the introduction of adaptive weights that are applied in order to penalize the regression coefficients in the penalty function. One of the key aspects that this methodology is aimed to accomplish is the Oracle Properties, and their key features are as follows [18]:

- The estimator possesses asymptotic properties consistent with the normal distribution.
- Consistency in variable selection is given by

$$\lim_{n} P(\hat{\beta}^n = \hat{\beta}) = 1$$

The oracle property is the capacity of an estimator to do as well as would happen had the actual underlying model been known beforehand. It is composed of two chief properties: First, consistency in the selection of the variable, i.e. an estimator can not only identify the true set of nonzero coefficients in a probability measure which approaches one. Second, the asymptotic normality which means that the estimated coefficients of the nonzero coefficients are asymptotically normally distributed as though the right model were known in advance [19] [20]. As mentioned earlier, the estimator $\hat{\beta}_{AdapLasso}$ satisfies the Oracle properties but the common Lasso estimator does not. The point of the Adaptive Lasso is to weight the penalty term with various weights to each of the regression coefficients [18] [21]This can be summarized in two steps.
Step one: Estimate the adaptive weight vector $\hat{\omega}$ based on the data, where $\hat{\omega}$ takes only positive values and is calculated according to the following formula:

$$\hat{\omega}_j = \frac{1}{|\hat{\beta}_j|^{\alpha}} \tag{20}$$

Where (α) a positive constant$(\alpha > 0)$ that represents the strength of the adaptive weighting and is related to the model dimensions. $(\hat{\beta}_j)$ An initial consistent estimator, typically obtained using the ordinary least squares (OLS) method or ridge regression in the presence of multicollinearity.
Step two: Given the weight vector $\underline{\omega} = (\omega_1, \omega_2, \ldots, \omega_p)^T$ standard Lasso estimator is re-modeled based on the Weighted Lasso model to minimize the objective function as follows:

$$\hat{\beta}_{\lambda} = \arg\min \left\{ \frac{1}{2n}(Y - \hat{U}\beta - \hat{F}\hat{\gamma})'(Y - \hat{U}\beta - \hat{F}\hat{\gamma}) + \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j| \right\} \tag{21}$$

Both parameters α and λ are selected using cross-validation (CV) within a two-dimensional framework in order to properly tune the Adaptive Lasso estimator. The penalty term in equation (20) $\left\{\lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|\right\}$ is used to estimate the values of $\hat{\beta}_j$ [20], where p becomes small. Specifically, when $(p \ll n)$, $\hat{\beta}_j = \hat{\beta}_{ols}$, whereas in the case where p is very large $(p \gg n)$, $\hat{\beta}_j = \hat{\beta}_{RR}$, $\quad j = 1, 2, \ldots, p.$

The estimation procedure in the proposed Factor-Augmented Regression Model is based on an optimization framework that combines factor extraction with penalized regression techniques. The model parameters are estimated by minimizing a penalized loss function, where regularization terms are incorporated to control model complexity and improve prediction performance in high-dimensional settings. Despite the advantages of the proposed model, some interpretability considerations should be discussed. Although the use of latent factors in the Factor-Augmented Regression Model may reduce the direct interpretability of individual predictors, it substantially improves prediction accuracy and model stability in high-dimensional settings. By capturing the common dependence structure among predictors, the model effectively mitigates multicollinearity and reduces dimensional complexity. Therefore, the trade-off between interpretability and predictive performance is considered acceptable in many modern statistical applications involving complex data structures. For the purpose of comparing the results of the estimations and the performance of the models, two of evaluation criteria is adopted, including the Mean Squared Error (MSE) and the coefficient of determination (R^2) were adopted to evaluate and compare the performance of the estimation methods,

as follows [22]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{22}$$

$$R^2 = 1 - \left[\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 / \sum\nolimits_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2\right] \tag{23}$$

3. The empirical and applied aspect

3.1. Simulation study

Three sample sizes of $1,000$ replications were used to carry out the simulation experiments. The data were created to test the workability of the Factor-Augmented Regression Model (FARM) in a high-dimensional data structure. It was based on the latent factor model which defines the relationship $X = \theta F + U$.
Where X is the explanatory variables, F is the set of latent factors, and θ is the factor loading matrix, and (U) is set of random errors. Both the errors and the latent factors were drawn in the standard normal distribution with $mean0$ and$variance1$ to give the realistic simulation, which was random and stable over time i.e. $f_{ik} \sim N(0,1),\ U \sim N(0,1)$.
This framework is involved in offering a suitable simulation environment to assess the efficiency of the various estimation techniques in the FARM model and still retains the basic features of high-dimensional data.The analysis and interpretation of the results will be presented by comparing the methods in terms of the Mean Squared Error (MSE) measure of the performance of each method explained in the theoretical section, at varying sample sizes and varying values p of to evaluate how the methods can perform in various situations that one may come across when applying the methods in real life.
To evaluate the robustness of the proposed model, additional simulation settings were considered under different correlation structures and noise levels. These settings aim to assess the stability and performance of the Factor-Augmented Regression Model under varying data conditions. The results indicate that the proposed approach remains stable and maintains good predictive performance even in the presence of high correlation and increased noise. Further extensive robustness analysis is suggested as a direction for future research. The performance of the proposed methods was evaluated using MSE and R^2 which are widely used measures for assessing prediction accuracy and model fit. The selected sample sizes $(n = 20\,to\,60)$ were intentionally chosen to reflect the high-dimensional setting $(p > n)$, which represents one of the primary challenges addressed in this study. Such settings are commonly encountered in modern applications involving complex and large-scale datasets.

Table (1): presents the Mean Squared Error (MSE) and the coefficient of determination (R^2) were adopted to evaluate and compare the performance of the estimation methods for the estimated models, as follows:

Table 1. Performance Comparison of Estimation Methods Based on MSE and R^2

n, p	Criterion	Estimation Methods					
		PCA	**LASSO**	**Adaptive LASSO**	**FARM-PCA + LASSO**	**FARM-PCA + Adaptive LASSO**	**Kernel Ridge Regression**
$n = 20,\ p = 30$	MSE	1.005	1.782	1.195	0.318	0.317	1.475
	R^2	0.883	0.802	0.865	0.965	0.965	0.859
$n = 30,\ p = 40$	MSE	0.777	0.788	0.589	0.328	0.327	1.509
	R^2	0.912	0.907	0.933	0.965	0.965	0.851
$n = 50,\ p = 65$	MSE	0.547	0.404	0.449	0.284	0.283	1.587
	R^2	0.945	0.961	0.957	0.971	0.972	0.853
$n = 60,\ p = 80$	MSE	0.532	0.449	0.506	0.296	0.293	1.726
	R^2	0.947	0.957	0.952	0.971	0.972	0.842

The results presented in Table (1) show that the hybrid methods based on FARM-PCA outperform all methods across different sample size conditions and values of p. Such approaches achieved the smallest Mean Squared Error (MSE) values, where the FARM-PCA + Adaptive LASSO approach recorded the lowest values followed by FARM-PCA + LASSO. This indicates that dimensionality reduction is highly effective when combined with regularization methods to enhance estimation accuracy. Conversely, classical approaches such as PCA and regularization methods (LASSO, Adaptive LASSO) show relatively poorer performance, resulting in higherMSE values; however, their performance improves as the sample size increases. It is also observed that increasing the sample size from $(n = 20\, to\, n = 60)$ leads to an overall reduction in MSE values and an improvement in performance for all methods. Nevertheless, the hybrid methods maintain their clear superiority even in high-dimensional settings $(p > n)$.
The results also indicate that Kernel Ridge Regression exhibits performance that is highly dependent on the underlying data structure. Its performance tends to deteriorate in linear settings and under small sample sizes with high dimensionality, while it performs better in the presence of nonlinear relationships. However, it generally produces higher MSE values compared to regularization-based and hybrid methods, which is attributed to its lack of an explicit variable selection mechanism and its tendency toward over-smoothing, particularly in high-dimensional scenarios.
Regarding the coefficient of determination (R^2), the FARM-PCA methods achieved very high values ranging between $(0.965 - 0.972)$, indicating strong explanatory power and stable performance. In contrast, the other methods recorded relatively lower values, despite showing improvement with increasing sample size. These findings confirm that integrating dimensionality reduction with regularization techniques, particularly using the Adaptive LASSO penalty, provides a more efficient framework for achieving a balance between predictive accuracy and explanatory power in high-dimensional data contexts.

3.2. Practical aspect

3.2.1. Data: In this section, the models will be applied to real health data acquired by the Iraqi Ministry of Health during the period (2024-2025) in which the data had been gathered on the individuals with obesity. The data were recorded according to the health indicators that were obtained by the device (MSLCA07 Body Building Weight Test System / Human Body Fat Health Analyzer).This device is considered an advanced body composition analyzer in the body composition study since it gives the precise measurements that are based on a computerized control unit that is supported by an Automatic Voltage Regulator (AVR). The machine is a multi-frequency bioelectrical impedance analysis device that measures the fat weight index and Body Mass Index (BMI) of different parts of the body, including fat and non-fat among other health indicators of the body. It also offers health interpretation through the measured indicators, which offers a scientific foundation in determining the health status and forming proper programs of enhancing the health of individuals. The size of the sample achieved $(n = 30)$ of the people who were obese and (39) of the health indicators were taken on individual basis.

3.2.2. Application of the Proposed Models: In the practical section, analysis is done on the best-performing hybrid FARM-based methods that were found during the simulation study. This choice is taken to be able to guarantee practical efficiency and to prevent unwarranted model complexity.

3.2.3. Regularization Parameter Estimation: The estimation methods applied in the practical aspect, such as LASSO and Adaptive LASSO, involve regularization parameters that constitute a fundamental component of the estimation process. To determine the optimal values of these parameters, the Cross-Validation (CV) approach was employed, as this method has proven effective in achieving a balance between model complexity and estimation accuracy. Table (2) presents the optimal values of the regularization parameters for the applied models. It is observed that the models exhibit noticeable differences in their optimal parameter values, reflecting the variation in the way each approach handles high-dimensional data.

Table 2. presents the optimal Regularization parameterfor the models.

Models	Regularization Parameter
FARM–PCA + LASSO	0.1871
FARM–PCA + Adaptive LASSO	0.4010

Several models were estimated within the Factor-Augmented Regression Model framework combined with PCA, namely LASSO and Adaptive LASSO. In addition, Kernel Ridge Regression was included as a benchmark nonlinear comparison method to evaluate the robustness and predictive performance of the proposed approach.

Table (3) presents the comparative results of these models based on the Mean Squared Error (MSE) and the Coefficient of Determination (R^2), which are commonly used measures to assess prediction accuracy and model fit in high-dimensional settings

Table 3. Comparative Performance of FARM-PCA-Based Models Using MSE and R^2

Methods	FARM-PCA + LASSO	FARM-PCA + Adaptive LASSO	LASSO	Kernel Ridge Regression
MSE	0.295683	0.295774	0.482803	5.18106
R^2	0.971053	0.971045	0.9622	0.8702

The results of the comparison of model performance are provided in Table (3) in both the Mean Squared Error (MSE) and the coefficient of determination (R^2). The findings have shown that all the methods could explain over 97 percent of the total variance in the dependent variable which is a good sign of efficacy of such models in solving the linkage amid variables. The findings also indicate that (FARM-PCA + Adaptive LASSO) method had the lowest value of MSE than all the other methods meaning that the method has the best predictive accuracy. This excellence demonstrates the ability of the model to better approximate the association among variables of health, including fat percentage and other body composition measures. Health wise, this implies that the model can be trusted to give more precise measurements of the health status of individuals hence making their therapeutic and preventive decisions to be more accurate. Moreover, the findings show that the (FARM-PCA + Adaptive LASSO) approach achieved the maximum value of the coefficient of determination (R^2), which implies that it has a high ability to provide the variability of the dependent variable. This shows how the model is effective in the real sense in the underlying factors that influence health status. Medically, this helps in understanding the most significant factors in the composition of the body hence promoting better diagnostic and health monitoring practices.
The results show that traditional methods such as Lasso and Kernel Ridge Regression (KRR) suffer from a clear decline in performance when dealing with high-dimensional data and small sample sizes $(n < p)$. Lasso suffers from high bias due to excessive shrinkage, while the KRR model fails to achieve good generalization because of its high sensitivity to the data structure and its inability to perform variable selection. In contrast, the FARM-PCA methods achieved superior performance, especially when combined with the Adaptive LASSO penalty function. This is due to their ability to exploit the factor structure, reduce dimensionality, and achieve a better balance between bias and variance.

4. Conclusions

1. The research findings indicated that the classic regression models have inherent limitations in the context of high-dimensional data especially where the dimensions (p) exceed the sample size (n), since they can experience multicollinearity and high variance problems resulting in poor accuracy of estimation and unstable outcomes. Conversely, Factor-Augmented Regression Model (FARM) was highly efficient because it combined the dimensionality reduction approach through Principal Component Analysis (PCA) with regularization-based variable selection methods.

2. Simulation experiments and empirical applications showed that hybrid approaches using (FARM-PCA) were much better than the traditional models, with lower Mean Squared Error (MSE) and higher coefficient of determination (R^2), which is an unequivocal indication of a superior predictive accuracy and a superior explanatory power.

3. The comparative findings showed that the best performance among all the reviewed methods was the (FARM-PCA + Adaptive LASSO) method because it is highly effective in eliminating the estimation bias and effectively select the important variables and is therefore the most appropriate method in analyzing the high-dimensional data, particularly in health-related applications.

4. On the whole, these results support the idea that dimensionality reduction tools, accompanied by regularization-based selection, may be a powerful and reliable methodological framework to solve high-dimensional data issues and enhance the quality of the statistical model.

5. Despite the fact that all the methods were found to be highly predictive, the Adaptive LASSO method was found to be superior with a better balance between the estimation accuracy and the interpretability.

Future research may focus on extending the proposed model to larger and more complex datasets, as well as investigating alternative regularization and factor extraction techniques. In addition, future studies could explore nonlinear extensions and advanced validation approaches to further improve prediction accuracy and model interpretability in high-dimensional data analysis.

5. Recommendations

The research suggests using the (FARM-PCA + Adaptive LASSO) methodology in health related studies, especially in the analysis of obesity and body composition data and in high dimensional models of data because it has been shown that it is effective in estimation and prediction. It also suggests that further research in terms of sample size and comparisons with non-obese populations should be multiplied to generalize the findings. Moreover, comparative research with nonlinear models and machine learning methods is necessary in order to compare performance in more complex environments. Lastly, the estimated parameters could be applied in creating medical decision-support models and creating customized health monitoring programs that will help in enhancing the standard of healthcare.

REFERENCES

1. J. Fan, Z. Lou, and M. Yu, "Are latent factor regression and sparse regression adequate?" *Journal of the American Statistical Association*, vol. 119, no. 546, pp. 1076–1088, 2024.
2. M. Harding, C. Lamarche, and C. Muris, "Estimation of a factor-augmented linear model with applications using student achievement data," *arXiv preprint arXiv:2203.03051*, 2022.
3. W. E. Zou, "Parameter estimation on dynamic factor augmented regression model," Ph.D. dissertation, University of California San Diego, 2023.
4. E. Chen, J. Fan, and X. Zhu, "Factor augmented matrix regression," *Journal of the American Statistical Association*, pp. 1–14, 2026.
5. Y. He, "Ridge regression under dense factor augmented models," *Journal of the American Statistical Association*, vol. 119, no. 546, pp. 1566–1578, 2024.
6. Y. Shi, L. Cai, X. Guo, and S. Zheng, "Adaptive adequacy testing of high-dimensional factor-augmented regression model," *arXiv preprint arXiv:2504.01432*, 2025.
7. J. Fan, Y. Yan, and Y. Zheng, "When can weak latent factors be statistically inferred?" *arXiv preprint arXiv:2407.03616*, 2024.
8. R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.
9. A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
10. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B*, vol. 67, no. 2, pp. 301–320, 2005.
11. J. Bai, "Inferential theory for factor models of large dimensions," *Econometrica*, vol. 71, no. 1, pp. 135–171, 2003.
12. J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 75, no. 4, pp. 603–680, 2013.
13. Q. Li, G. Cheng, J. Fan, and Y. Wang, "Embracing the blessing of dimensionality in factor models," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 380–389, 2018.

14. P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
15. A. Alfons, C. Croux, and S. Gelper, "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *The Annals of Applied Statistics*, pp. 226–248, 2013.
16. F. Audrino and L. Camponovo, "Oracle properties and finite sample inference of the adaptive lasso for time series regression models," *arXiv preprint arXiv:1312.1473*, 2013.
17. X. Cui, R. Xiao, X. Liu, H. Qiao, X. Zheng, Y. Zhang, and J. Du, "Adaptive lasso logistic regression based on particle swarm optimization for alzheimer's disease early diagnosis," *Chemometrics and Intelligent Laboratory Systems*, vol. 215, p. 104316, 2021.
18. S. Muhammadullah, A. Urooj, F. Khan, M. N. Alshahrani, M. Alqawba, and S. Al-Marzouki, "Comparison of weighted lag adaptive lasso with autometrics for covariate selection and forecasting using time-series data," *Complexity*, vol. 2022, no. 1, p. 2649205, 2022.
19. L. Wang, J. Shen, and P. F. Thall, "A modified adaptive lasso for identifying interactions in the cox model with the heredity constraint," *Statistics & probability letters*, vol. 93, pp. 126–133, 2014.
20. K. Darwish and A. BÜYÜKLÜ, "Robust linear regression using l1 penalized mm estimation for high dimensional data," *American Journal of Theoretical and Applied Statistics*, 2016.
21. H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.
22. D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.