Seminar Paper

The Fragility of Sparsity

Department of Statistics
Ludwig-Maximilians-Universität München

Anjali Sarawgi

Munich, August 31, 2024

Submitted in partial fulfillment of the requirements for the degree of M. Sc Statistics and Data Science.
Supervised by Prof. Vincent Starck

Abstract

This seminar paper critically examines the robustness and reliability of Sparsity-Based Estimators (SBEs), such as Lasso, as discussed in the original paper, *"The Fragility of Sparsity"*. While SBEs have become widely used for handling high-dimensional datasets by selecting a subset of relevant variables, their performance is also sensitive to various factors and choices of normalization strategies. In addition to summarising and clarifying the main findings of the original paper, this study evaluates the performance of SBEs for different conditions, such as varying dataset sizes, dimensionality settings, and normalization techniques. The findings reveal that SBEs exhibit significant sensitivity to these factors, often resulting in high variability in the estimates. This paper also explores how these estimators perform in machine-learning contexts, highlighting the challenges they face in maintaining predictive accuracy and performance. The code is available at https://github.com/anjalisarawgi/fragility_of_sparsity_review

Contents

1 Introduction

Sparsity-based estimators (SBEs), such as Lasso have gained significant attention in statistics and econometrics due to their ability to handle high dimensional data, where the number of potential predictors often exceeds the number of observations. Traditional estimation techniques like Ordinary Lease Squares (OLS) become impractical in such scenarios. SBEs address this by assuming that only a few of the predictors affect the outcome, allowing for more efficient variable selection and model simplification.

However, the assumptions underlying these Sparsity-Based Estimators (SBEs), especially the sparsity assumption have been questioned. The sparisty assumption suggests that only a small subset of the many potential variables affect the outcome, which may not always hold. When this assumption is violated, the reliability of SBEs is compromised. Furthermore, the performance of SBEs is sensitive to various methodological choices, such as the construction of the control matrix, the choice of tuning parameters, and the method of normalization, all of which can introduce bias or reduce efficiency.

This seminar paper is centered around the findings of the paper "The Fragility of Sparsity", which critically examines the robustness of SBEs in various empirical contexts. Specifically, "The Fragility of Sparsity" highlights the sensitivity of SBEs to the assumptions and methodological choices made during model specification, raising concerns about their reliability. In this seminar paper, I begin with the interpretation of the original paper, clarifying its main findings along with some examples for better understanding.

Building on this, I investigated these concerns in detail, focusing on the empirical validity of the sparsity assumption and the robustness of SBEs to different normalization choices. To extend the discussion, I conduct experiments applying SBEs to a new dataset to test whether the issues highlighted in "The Fragility of Sparsity" are also true in different contexts. Additionally, I explored the impact of varying data sizes on prediction accuracy and loss in machine learning contexts and analyzed how changes in the number of covariates and the regularization parameter for Lasso influence these estimates. This empirical work aims to validate and explore the generalizability of existing findings and to explore how SBEs perform in real-world data scenarios.

2 Literature Review

In the recent years, the rapid surge in the availability of large datasets, combined with advances in statistics, machine learning, and econometrics, has generated significant interest in predictive models with many possible predictors (Domenico Giannone (2021)). This demand has led to the widespread adoption of Sparsity-based Estimators (SBEs), which are particularly well-suited for high-dimensional. Sparsity-based estimators such as the Lasso (Least Absolute Shrinkage and Selection Operator) estimator has revolutionalized statistical methods in high-dimensional data analysis, whereas traditional estimators like Ordinary Least Squares (OLS) struggle.

The introduction of Lasso by Tibshirani (1996) performs variable selection by shrinking coefficients towards zero, effectively performing variable selection and regularization simultaneously. Following this, many alternative methods of variable selection were additionally proposed. A few of these include Elastic Net which combines the penalties of Lasso and Ridge Regression (Hui Zou (2005)); SCAD (Smoothly Clipped Absolute Deviation) which aims to reduce the bias introduced by Lasso's penalty(Jianqing Fan (2001)); and various boosting approaches that incrementally build models to enhance predictive performance (Peter Bühlmann (2003)).

Despite this success, sparsity-based estimators have not been without criticism. The key assumption underlying these methods is the assumption of sparsity which implies that the true model is sparse. However, recent studies have highlighted the fragility of this assumption. For example, Giannone et al. (2021) and Wüthrich and Zhu (2023) pointed out that SBEs can be highly sensitive to the choice of the regressor matrix, leading to possibly misleading results. The issue of non-invariance to linear reparametrization, as discussed in the paper "The Fragility of Sparsity", further complicates the use of SBEs. It implies that the results can be highly dependent on the specific model setup, thereby undermining the robustness of these estimators.

"The Fragility of Sparsity" paper examines the robustness of SBEs through three empirical applications: the effect of abortion on crime (Donohue III and Levitt (2001)), occupational upgrading by Black Southerners (Ferrara (2022)), and the impact of moral values on voting behavior (Enke (2020)). The findings reveal that the SBEs are highly sensitive to normalization choices and often fail the sparsity assumption, which raises concerns about the reliability of these estimators in empirical research. The development of tests to assess the validity of the sparsity assumption, as proposed in the paper, represents an important contribution to this field. These tests provide researchers with important tools to verify the robustness of their results when using SBEs under the assumption of sparsity.

3 Methodology and Assumptions

To replicate and extend the findings from the original paper "The Fragility of Sparsity", we followed the methodological approach. We tested how robust sparsity-based estimates (SBEs) like the lasso are in different situations with different models. We followed the original paper's methodology and added additional steps to explore how its findings may apply in different empirical scenarios and contexts.

3.1 Methodological Setup

1. **Choice of Datasets:** The original paper analysed the results on three empirical applications to examine the robustness of Sparsity-Based Estimators (SBEs). They are:

    - Effect of Abortion on Crime (Donohue III and Levitt (2001))

- Occupational Upgrading by Black Southerners (Ferrara (2022))
- Impact of Moral Values on Voting Behaviour (Enke (2020))

To extend the analysis, our study implemented two additional datasets:

- Lalonde Dataset (Sharma and Kiciman (2020))
- Communities and Crime dataset (Redmond (2011))

2. **Baseline Comparison:**
   As in the original paper, we conducted a baseline comparison using regression models with both Ordinary Least Squares (OLS) and SBEs, including the post-double selection method. The authors used this comparison to see how for the differences in the estimates arising from the sparsity assumption. Our study also used this comparison to assess how well SBEs work compared to OLS in different situations.

3. **Choice of predictors (p)**:
   The original paper carefully chooses the number of predictors (p) relative to the number of observations (n)to ensure that p is close to but less than n. This allowed the authors to use OLS as a benchmark to evaluate the performance. The original paper mainly focused on a high-dimensional case where p is close to but less than n.

   In our study, we extend this approach by experimenting with three different cases:

   (a) **Case 1:** Where p has its original dimensions which is much smaller than n
   (b) **Case 2:** Where p is close to n
   (c) **Case 3:** Where p is more than n

   This approach allowed us to test the robustness of the SBEs across different dimensional settings.

4. **Model:** Both, the original paper and our experiments use the post-double lasso as the primary model. The post-double lasso is designed to enhance variable selection and address potential biases in high-dimensional settings. The post-double lasso is constructed as follows:

   (a) First Lasso Regression: Regressing treatment (D) on the control variables (X)
   (b) Second Lasso Regression: Regression outcome (Y) on the control variables (X)
   (c) Taking the union of selected features from both lasso regressions
   (d) OLS regression with only the selected features from the previous steps

5. **Normalization Variations:**
   The original paper evaluated the impact of various normalization strategies on the control matrix. In our study, we followed these methods too. They include:

   (a) Categorical variables: Dropping different columns as the reference categories to resolve multicollinearity among the control variables

(b) Numerical variables: Normalizing the baseline categories with different offset values which include demeaning, subtracting with median, min-max scaling or substracting with random offsets.

6. **Application of Sparsity Tests:**
   The authors presented two tests to evaluate the validity of the sparsity assumption. Our study also follows the same tests to evaluate the assumptions of sparsity. These tests are the 'Hausman Test' and the 'Residual Test'.

3.2 Assumptions

3.2.1 Choice of p

The authors assume that p is less than n i.e. $p < n$, but still relatively large. This assumption is important here because:

1. It allows us to use the OLS benchmark for comparisons. That is, OLS being a traditional estimator restricts us for p to be less than n in all cases. Using this restriction of $p < n$ allows us to use the OLS benchmark to understand the efficiency and robustness of our sparsity estimators.

2. By ensuring that p is stil high but less than n, we are able to create a high-dimensional scenario where traditional OLS may struggle, yet remain computationally possible.

This setup allows us to directly compare the performance of OLS against Sparsity-Based Estimators (SBEs) in a setting where p approaches n, testing the limits of OLS while also exploring the potential advantages of SBEs.

Furthermore, we also evaluate for the case of p larger than n in our experiments.

3.2.2 Approximate Sparisty

For the tests and analysis in this study, the assumption of sparsity is relaxed from exact sparsity to approximate sparsity. Approximate sparsity ensures that even if our coefficients γ are not exactly sparse, most of the predictive power comes only from a few non-zero coefficients. This means that for any given transformation, we accept a scenario where the model coefficients are not exactly sparse but are close to sparse.

To understand this mathematically, a sparsity index s is introduced, indicating the maximum number of non-zero coefficients allowed. The sparsity index is given by:

$$s = o\left(\sqrt{p}/\log p\right) \tag{1}$$

This indicates that the number of non-zero coefficients (s) should grow slower than $\sqrt{p}/\log p$. This criterion ensures the model remains sparse enough to allow for accurate inference.

3.2.3 Homoscedasticity and Heteroscedasticity

In regression analysis, the behavior of the error terms (residuals) is crucial for the validity and efficiency of the estimates. The paper primarily focuses on the assumption of homoscedasticity, where the variance of errors is constant across all observations. This assumption forms the basis for the theoretical results and comparisions between Ordinary Least Square (OLS) estimators and Sparsity-Based Estimators (SBEs).

Under the assumption of homoscedasticity, OLS estimators are considered the Best Linear Unbiased Estimators (BLUE), according to the Gauss-Markov theorem, meaning they achieve the smallest variance among all linear unbiased estimators. This allows for straightforward calculations of standard errors, confidence intervals, and hypothesis tests which are essential for statistical inference.

The authors also extend the analysis to cases where heteroscedasticity is present i.e. when the variance of error terms varies across observations. In high-dimensional scenarios where p is large relative to n, heteroscedasticity could reduce the potential efficiency gains of SBEs. The authors thereby also present the theoretical insights for heteroscedastic cases though the assumption of homoschedasticity is the baseline assumption. Despite these challenges caused by heteroschedasticity, SBEs remain a viable option because they leverage the sparsity assumption to mitigate some of the problems caused by heteroscedasticity.

4 Review of "The Fragility of Sparsity": Analysis and Findings

4.1 Linear Regression for Causal Effects

In causal inference, the primary objective is to estimate the causal effect of a treatment or an intervention on an outcome variable. Here, linear regression is useful because it allows us to control for confounding variables to isolate the effect of the treatment. By including these con founders as control variables, we can accurately isolate the effect of treatments.

The key assumption here is **the assumption of conditional randomness**. This assumption states that after controlling for a set of variables, the remaining variation in the treatment variable is as good as random. When this assumption holds, the coefficients derived from this linear regression model can be interpreted causally. This assumption gets stronger as we include more relevant variables.

Model Specification:

$$Y_i = D_i\beta + W_i'\gamma + U_i, \quad E\left[U_i \mid D_i, W_i\right] = 0 \tag{2}$$

where, Y_i is the outcome variable, D_i is the treatment variable, β is the causal effect of D_i on Y_i, W_i is a vector of control variables, γ is the vector of coefficients for W_i, and U_i

is the error term.

Note: $E[U_i \mid D_i, W_i] = 0$ ensures no omitted variable bias and $E[U_i \mid D_i, W_i] = 0$ is the control function.

In practice, researchers often seek to increase the number of predictors (p) in their studies. This is because the higher the value of p, more stronger our assumption of conditional randomness becomes. However, OLS is a traditional estimate that limits p to be smaller than n and is the most effective when $p \ll n$. Even if we consider p to be large but less than n, OLS estimates can be unreliable and noisy giving us less reliable estimates. These limitations of OLS have led to the development of Sparsity-based estimates (SBEs).

4.2 Sparsity and Sparsity Based Estimates (SBEs)

Sparsity refers to the idea that a model can be well-approximated by a small number of non-zero coefficients. This is particularly useful in a high dimensional setting where we want to choose only the relevant variables. In this study, when we have a large set of predictors, only a small subset of these predictors (s) is assumed to significantly influence the outcome. That is, in our control matrix $W_i'\gamma$ most of the coefficients γ are zero.

The **sparsity assumption** implies that the true underlying model involves only a few active predictors, even when the number of potential predictors is large. Formally, let β be the vector of coefficients. The sparsity assumption can be written as:

$$\beta_j = 0 \text{ for most } j = 1, 2, ...p$$

This assumption allows for a more efficient estimation on identifying and estimating only the non-zero coefficients β.

Sparsity Based Estimators
Sparsity-based estimators are built upon the assumption of sparsity. The most commonly used sparsity-based estimator is the Lasso estimator along with some of its variations, such as Debaised Lasso, Double Lasso, and Grouped Lasso have also been developed. The lasso estimator is a statistical regularization technique that adds a penalty to the OLS estimator to enforce sparsity in the model. The lasso estimator can be defined as:

$$\hat{\beta}_\lambda = \arg\min_\beta \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

where:

- λ is the regularization parameter which encourages sparsity my shrinking the coefficients to 0
- y is the $n \times 1$ response vector
- X is the $n \times p$ design matrix

- β is the $p \times 1$ vector of coefficients
- $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L1 norm, imposing sparsity on the model

Credibility concerns of SBEs
Though sparsity-based estimators have gained popularity due to their ability to handle high-dimensional data, the assumption of sparsity can be fragile. While they are useful, they also come with problems. The authors highlight these concerns regarding to the credibility of sparsity:

- Sensitivity: SBEs are highly sensitive to how we normalize our control matrix. Even small changes in how we handle collinear columns or change baseline controls can lead to significant variations in the estimates. For example, changing the way we handle categorical education variables or how we center age as a cont variable can change the SBE results. This sensitivity can make SBEs less reliable in practice.
- Robustness: OLS is usually robust to different normalizations. However, SBEs may not be robust and reliable. For instance, if the decision on how to group education levels or center age is arbitrary, the results from SBEs may not be trustworthy!
- Lack of theoretical reasoning: In social sciences, there isn't a strong theoretical reason to believe that only a few control variables can capture most of the confounding effects. Additionally, its unclear what level of sparsity is even needed to ensure unbiased estimates. This lack of clarity and theoretical backing further complicates the use of SBEs in these fields.

4.3 Normalization of Control Matrix

In regression analysis, 'normalization' just means the transformations applied to the variables before they're used in models. In this case of estimating causal effects, these normalizations are about transforming the control variables. In these cases, we create a control matrix (W_i), which includes all of our control variables. The paper looks at how setting up the control matrix with different normalization methods affects the results for the coefficients of estimates when we use SBEs.

For example, consider a study to analyze the effect of education on income. A decision has to be made on how to handle control variables like age (continuous variable) and education level (categorical variable). As a pre-processing step, we may handle variables differently such as centering the age variable (by subtracting its mean vs median), or choosing which education category should be treated as a baseline variable (eg "High School" vs "PhD"). These choices which one may assume are not that important, might be very crucial as they may affect our results and estimates significantly.

The paper explores how different normalization methods affect the performance of sparsity-based estimators (SBEs). Unlike traditional OLS where the choice of normalization doesn't affect the final estimates, SBEs are highly sensitive to normalization and how

the control matrix is constructed. This highlights the importance of thoughtful normalizations in obtaining reliable results with SBEs.

The normalization methods examined in this paper include:

1. **Grouping columns to resolve multicollinearity**, where different baseline categories are compared
2. **Normalizing baseline controls**, such as subtracting the mean or median
3. **Grouping categories together as subsets**, where multiple categories like "High School" and "Bachelors" are combined into one subset.
4. **Varying offset values** for continuous variables, such as age, by subtracting different values of λ (eg. 40, 41, 42, 43).

This lack of variance to linear reparameterization means that researchers need to be very careful when applying or specifying the control matrix when using sparsity-based estimators. The study also shows that the likelihood of getting a sparse representation when we randomly choose a normalization matrix decreases as the number of observations increases. This means that relying on the default choices for making our control matrix (W) in statistical software is probably not going to help us achieve sparsity. Instead, in practice these choices should be carefully considered and a well-justified normalization method is essential for ensuring the validity of the analysis.

The authors present this sensitivity of SBEs to normalizations theoretically and experimentally (on the three empirical datasets). The following sections explain the theoretical findings of the paper.

4.3.1 Rotation

We study rotation to understand how likely it is to maintain sparsity when the control variables are re-expressed in different ways.

Consider a rotation matrix R (a square matrix) which when multiplied by a vector, rotates the vector in space without altering the magnitude of angles between the vectors. The rotation matrix comes with two key properties:

1. Orthogonality: $R'R = I_p$ where I_p is the identity matrix. This property ensures that the rotation preserves the lengths of vectors and angles between them, thereby not distorting the data.
2. Determinant of 1: $det(R) = 1$. This property implies that the transformation is a pure rotation or reflection without any scaling.

Effect of Rotation on Sparsity:
Consider a regression model where the coefficients γ are sparse, i.e. most of the coefficients are zero. When the control variables are transformed using a rotation matrix R, the new

predictors W_i are given by RW_i. Correspondingly, the coefficients are also transformed as $\tilde{\gamma} = R\gamma$. This transformed coefficient vector $\tilde{\gamma}$ tends to lose its sparsity because the values get mixed up and spread out, making more coefficients non-zero.
This effect occurs because rotation redistributes the values across the entire vector space, which dilutes the sparsity present in the original coefficient vector. As a result, the post-rotation coefficient vector $\tilde{\gamma}$ is less likely to remain sparse.

Theorem 1 *Suppose that the eigenvalues of $E[W_i W_i\prime]$ are bounded away from zero and infinity, and that $||\gamma||_2 \approx 1$. Then the logarithm of the probability that the model with regressor $W_i = RW_i$ satisfies the sparsity conditions (sufficiently sparse) is of the order $\frac{-p}{n}\log(p)$.*

Interpretation: This theorem gives a theoretical analysis of the impact of rotations on sparsity. It says that as p increases, the probability that the randomly rotated model will remain sparse also decreases rapidly. Specifically, this probability is given by a bound of the order $p^{-4}log(p)$, which becomes exceedingly small as p grows larger. For instance, when $p >= 50$, the probability of maintaining sparsity is less than 10^{-21}. Therefore, achieving sparsity through random rotation is highly impractical, particularly in high-dimensional settings where the number of predictors is large.

This theorem is based on two conditions:

1. The eigenvalues of the matrix $E[W_i W_i\prime]$ must be bounded away from zero and infinity. This condition ensures that variability is appropriate ensuring that there is numerical stability in the model.

2. The norm of the coefficient vector γ should be approximately 1 to ensure the coefficients are of reasonable scale.

Moreover, this theoretical theoretical analysis presents that the robustness of sparsity-based estimators is heavily influenced by how control variables are transformed.

4.3.2 Categorical Data

In regression models, dealing with categorical variables is common. Categorical variables, such as education levels are typically represented as groups of binary variables (also known as dummy variables). Further, to avoid the problem of perfect collinearity, we drop one of these dummy variables as a 'reference category' to avoid perfect collinearity.

For example, consider an education variable with four levels: High School, Bachelor, Master, and PhD. We could create three binary variables "High School", "Bachelor" and "Master" - while using "PhD" as a reference category. Sometimes, the categories can also be grouped as subsets to understand a combined effect on the outcome. For instance, High School and Bachelor's can be combined into one category.

Effect of choice of reference category on Sparsity: The study shows that the choice of the baseline (reference) category can significantly influence the sparsity pattern.

If a different category is chosen as a reference, the coefficients of the dummy variables created for these categorical variables will change, leading to a different set of variables being selected. Thus, the sparsity of the model can be heavily dependent on how the categorical variables are represented.

Theorem 2 *Suppose a single coefficient on $W_i = A_0 Z_i$ is constant and non-zero, and the number of zeros K in the corresponding row of A_0 satisfies $0 < \lim_{n\to\infty} K/p < 1$. If all baseline categories have population fractions of the same order, then the probability that the model with on $W_i = A_0 Z_i$ satisfies the sparsity assumption is no larger than $(1 - q + \varepsilon)^K$ for all $\varepsilon > 0$ and large enough p.*

Interpretation: Theorem 2 theoretically explains how different representations of the categorical variables in a regression model affect the model's sparsity.

Consider,

1. A categorical variable with p categories
2. $\mathbf{A}$ as the set of all possible full-rank $pxp - 1$s matrices
3. $A_0 \in \mathbf{A}$ as the standard way of encoding this categorical variable. Under this standard coding, let's assume the model is sparse, i.e. the coefficient vector γ_0 has few non-zero coefficients.
4. An alternative encoding matrix $\mathcal{A} \in \mathbf{A}$ where, $\mathcal{A}$ represents different ways of representing this categorical variable.

Theorem 2 investigates the probability that a randomly chosen encoding matrix $\mathcal{A}$ will result in a sparse representation. It shows that if a model is sparse under one encoding scheme (like A_0), then there is little chance it will remain sparse if the categorical variables are re-encoded using a different randomly chosen scheme (like $\mathcal{A}$). This result shows how easily the sparsity assumption can be affected and suggests that the results of sparsity-based estimators can vary a lot depending on how the categorical variables are encoded. Additionally, as the number of categories p increases, the probability of achieving a sparse representation approaches zero, further highlighting the sensitivity of sparsity to a randomly chosen encoding scheme.

4.3.3 Offsets for Normalization

For modeling complex relationships in regression, **Hermite polynomials** are useful because they help us capture non-linear effects. By introducing nonparametric regression, the paper extends the discussion to more complex scenarios involving continuous variables.

One of the most important things to consider when using Hermite polynomials is how we center or shift the variables of interest before applying the polynomial transformation. The authors indicate that this offset or shift can have a big impact on the results of regression, especially to maintain sparsity.

Theorem 3 Suppose $\lambda = L/\log p, L > 0$. If L is fixed, then for $1 \leq j \leq L\sqrt{p}/\log p$ and $p \geq \max\{(2L)^2, 6\}, \tilde{\gamma}^2_{p-j} \geq Ce^{j/2}$, where C is an absolute constant, and sparsity assumption fails. In contrast, if $L \to 0$, sparsity assumption holds.

Interpretation: Theorem 3 explains the impact of choosing an offset λ when using Hermite polynomials in a regression model. Specifically, it presents how this choice of an offset can affect the sparsity of the model.

Consider:

1. **Scalar Variable z_i:** This represents the continuous variable of interest which is to be transformed using Hermite polynomials
2. **Hermite Polynomials $H_j(x)$:** to capture non linear relationships between the continuous variable z_i and the dependent variable. They are particularly effective when z_i follows a normal distribution.
3. **Offset λ:** The offset λ is a sift applied to the variable z_i before it is transformed by the Hermite polynomial
4. $\boldsymbol{p}$: p is the number of polynomial terms in the model
5. $\boldsymbol{L}$: where L is a positive number that does not change in some cases but can decrease in others
6. $\boldsymbol{\lambda = \frac{L}{\log p}}$: This is the formula that determines the offset on L based on p

Fixed L: When L is fixed, as you add more terms to the model (i.e. as p gets bigger) the importance of many of these terms (represented as coefficients) starts to grow quickly. This is because $\log p$ grows much more slowly than p itself causing the λ to not decrease quickly enough. This slow decrease in λ means that the coefficients in the model start to grow too quickly as more terms are added (as p increases). Consequently, many terms in the model become important, causing the model to become more complex and less sparse.

L approaches 0: If L decreases as p increases (i.e.$L \to 0$), the offset λ decreases more effectively, adjusting for the slow growth of $\log p$. This adjustment ensures that the coefficients remain small and helps to keep the model sparse.

Key point: The way we adjust our variables with the offset λ either makes our model very complex with a lot of important terms, or it can keep it very simple with only a few important terms. If we pick the wrong shift λ like when L is fixed, your model might end up being too complicated with too many important terms. Whereas, if we carefully choose λ where L decreases as more terms are added, the model remains sparse retaining only a few terms that matter.

4.4 Efficiency gains under sparsity

In statistical modeling, efficiency gains refer to the reduction in variance of an estimator leading to more precise estimates. Here, this concept is useful to compare the Ordinary Least Squares (OLS) estimators to alternative estimators like Sparsity Based Estimators (SBEs). This section quantifies the potential efficiency gains of alternative estimators like SBEs relative to OLS, showing that these efficiency gains are limited unless p is large enough and comparable to n.

Theoretical Framework

To quantify the potential gains from SBEs over using OLS estimators, the study compares the standard errors. It presents that the proportional variance reduction of SBEs relative to OLS is bounded by $1 - \frac{p}{n}$. This implies that significant efficiency gains can only be achieved when p is large but still less than n.[1]

This efficiency gain can be expressed as the ratio of standard errors under SBE and standard errors under OLS. That is,

$$\frac{s^*}{s_{\text{OLS}}} = \sqrt{1 - \frac{p}{n}} \tag{3}$$

where:

- s^* is the standard error under sparsity-based estimator
- s_{OLS} is the standard error under OLS estimator
- p is the number of variables (controls)
- n is the number of observations

This formula indicates that the efficiency gain depends on the ratio of $\frac{p}{n}$.[2]As p increases relative to n, the efficiency gains also increase but the gains are bounded by the ratio $\frac{p}{n}$. This means that there is a limit on how much more efficient SBE can be over OLS.

Practical Examples of Efficiency Gains

In practice, the efficiency gains under sparsity depend on the size of p relative to. Consider the following two scenarios for n = 100 :

[1]Here, p is less than n is considered to be able to use the OLS benchmark for comparison.

[2]This formula assumes that the errors (residuals) in the model have constant variance i.e. is homoscedastic. If the variance of the errors varies across observations (heteroskedasticity), the efficiency gains from SBEs may be smaller than expected. This happens because the estimation of variance might be biased, reducing the effectiveness of SBEs in lowering standard errors compared to OLS. So, while SBEs can still offer efficiency gains, these gains might be less pronounced in the presence of heteroskedasticity. The efficiency gain formula for heteroscedastic scenarios is given as: $\frac{s^*}{s_{\text{OLS}}} = \sqrt{(1 - \frac{p}{n}) * k}$, where k is a factor that adjusts for the heteroscedasticity present in the data. This factor k represents the extent to which heteroscedastic inflates the variance of the errors, thereby diminishing the relative efficiency gains of SBEs.

1. **Case 1:** $p = 80$
   The efficiency gain formula becomes:

   $$\frac{s^*}{s_{\text{OLS}}} = \sqrt{1 - \frac{80}{100}} \approx 0.45$$

   - OLS is efficient and satisfies the Best Linear Unbiased Estimators (BLUE) property making it a better choice in this case
   - Here, the standard error of the SBE is approximately 45% of the standard error of the OLS. This indicates that the SBE is more efficient in this case.

2. **Case 2:** $p = 20$
   The efficiency gain formula becomes:

   $$\frac{s^*}{s_{\text{OLS}}} = \sqrt{1 - \frac{20}{100}} \approx 0.89$$

   - OLS becomes noisy, exhibits multicollinearity, and tends to overfit, making SBE the ideal choice
   - Here, the standard error of the SBE is approximately 89% of the standard error of the OLS. This indicates that the SBE is more efficient in this case.

These examples illustrate the limitation that efficiency gains are capped and are minimal when the ratio $\frac{p}{n}$ is small. Thus, the potential benefit of using SBEs over OLS is more pronounced when p is large relative to n.

Theoretical Foundations: Understanding Lemma 1
Lemma 1 is a foundational result that asserts the reliability of OLS estimators under certain conditions, particularly when the ratio $\frac{p}{n} < 1$ and the model's errors are well-behaved (i.e. homoscedastic).

The lemma states that:

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}(\beta, \sigma^2)$$

This implies that under the given conditions, the OLS estimator is asymptotically normal, meaning that as the sample size n increases, the distribution of the OLS estimator approaches the normal distribution, with mean β and variance σ^2.

4.5 Tests for Sparsity

The authors introduce two statistical tests - **Hausman Test** and the **Residual Test** to evaluate the robustness of sparsity-based estimators (SBEs) under varying normalization methods. These tests help us understand the reliability of these estimators in high-dimensional settings.

4.5.1 Hausman Test

The Hausmann Test acts as a relevance check in econometric models. The Hausman tests is a statistical method used to study the consistency of different estimators by comparing the results of a sparisty-based estimator to the Ordinary Least Square Estimators. This test aims to determine whether the difference between these two estimates is statistically significant. If a significant difference is found, it may indicate that the sparsity-based estimators (SBEs) are unreliable or biased as compared to the OLS estimators.

Steps involved:

1. Define the Hypothesis:
   H_0 : The difference between OLS and SBE estimates is not statistically significant (assumptions hold)
   H_1 : The difference between OLS and SBE estimates is statistically significant

2. **Calculate the Estimates:** Calculate the coefficient estimates $\hat{\beta}_{OLS}$ using the OLS estimator and $\hat{\beta}_{SBE}$ using SBE estimators.

3. **Calculate the Variances:** Next, determine the variance of of these estimates i.e. s^2_{OLS} and s^2_{SBE}.

4. **Determine the difference and its standard error:** Compute the difference between the two estimates: $d = \hat{\beta}_{OLS} - \hat{\beta}_{SBE}$

   Additionally, calculate the standard error of this difference as $SE(d) = s_H = \sqrt{s^2_{OLS} - s^2_{SBE}}$

   The authors present a Lemma (Lemma 2) which provides the theoretical foundations of the Hausman Test, ensuring that under the right conditions, this difference can be interpreted reliably in statistical inference. [3]

5. **Compute the test statistic:** The test statistic z is computed as: $z = \frac{d}{SE(d)}$
   This z-statistic follows a normal distribution and it is used to make inferences about whether the difference between OLS and SBE is statistically significant. A large value of z indicates a significant difference leading to the rejection of the null hypothesis.

Interpretation: If H_0 is rejected, the Hausman test finds that the difference between the OLS and SBE estimates is statistically significant. This significant difference suggests the assumptions under which SBEs are applied might not hold, or that the SBE estimators might be producing biased or inconsistent estimates.

[3] Lemma 2: It presents that under certain conditions, the difference between the OLS estimator ($\hat{\beta}_{OLS}$) and another estimator ($\hat{\beta}$) - denoted as $d = \hat{\beta}_{OLS}$ - $\hat{\beta}^*$ - follows a standard normal distribution when normalized with its standard error s_H. Specifically, $\frac{d}{s_H} \sim \mathcal{N}(0,1)$, which allows the use of the Z-test to determine if the difference d is statistically significant.

4.5.2 Residual Test

The residual test evaluates the validity of the sparsity assumption by comparing the residual sum of squares (RSS) from the sparsity-based estimators (SBEs) to the Ordinary Least Squares (OLS) regressions. Here, the test uses the F-statistic to verify if the difference is significant. F-statistic is used here because it is a standard test to compare variances.

Steps Involved:

1. **Linear Regression Setup:** Consider a linear regression model $Y_i = X_i\prime\alpha + \epsilon_i$ where Y_i is the dependent variable, X_i represents the independent variables and ϵ_i is the error term. Additionally, consider the expectation $E[\epsilon_i|x_i] = 0$ and ϵ_i is independent across observations conditional on the regressors.

2. **Define the Hypothesis:**
   H_0 : The difference in RSS between OLS and SBE is not statistically significant
   H_1 : The difference in RSS between OLS and SBE is statistically significant

3. **Assumption of Sparsity:** Introduce a subset $S^* \subset \{1, \ldots, p\}$ such that the sparse approximation error $||(I - P_S^*)X_\alpha||^2$ is small. Here, P_S is a projection matrix associated with X_{S^*} and X_α represents the fitted values when using the full set of predictors and their associated coefficients.[4]

4. **Compare RSS:** Calculate and compare the RSS from the SBE model with that of the OLS model. The Residual Sum of Squares (RSS) is calculated as the sum of the squared differences between the observed values Y_i and the predicted values $\hat{Y}_i$ produced by the model.

5. **Use F-statistic to check for the significance:** Compute the F-statistic to test whether the difference in RSS between the models is statistically significant. If the F-statistic is large, the null hypothesis is rejected which thereby indicates that the sparsity assumption is not valid since the differences in RSS between the two estimators are significantly different.

   Additionally, the authors present a Lemma (Lemma 3) which provides the theoretical foundations for a residual test, ensuring that under the right conditions, this test can be applied reliably in statistical inference.[5]

[4]The sparse approximation error $||(I - P_S^*)X_\alpha||^2$ measures how well the model can be approximated by a sparse set of predictors. Here, S represents the subset of the variables which provides the best sparse approximation. If the projection of X_α onto the subspace spanned by X_S captures most of the variability in the data, the approximation error is small. Essentially, it checks how much of the information is lost when only a sparse set S is used instead of a full set X. A small error indicates that the sparse approximation is a good representation of the full model.

[5]Lemma 3: This lemma validates the assumption of sparsity in a high dimension setting. Showing the test statistic follows a standard normal distribution $N(0, 1)$, allows researchers to apply standard hypothesis procedures. Specifically, if the test statistic aligns with $N(0, 1)$, p values can be calculated for statistical inference. This convergence does not confirm that sparsity holds, rather it provides a statistical basis for testing the assumption. Thereby, by comparing the residuals from SBEs and OLS, and observing whether they align with the normal distribution, researchers can determine if the sparsity assumption is reasonable for a particular scenario.

Interpretation: If the H_0 is rejected, the Residual Test finds that the difference between the residual sum of squares of the OLS and SBE estimates is statistically significant. That is, a lower difference suggests that the sparsity assumption holds whereas, a higher difference indicates significant differences in the model indicating the sparsity assumptions may not hold.

4.6 Empirical Results

The empirical analysis presented in this paper demonstrates the sensitivity of Sparsity-based estimators (SBEs) when utilized with high-dimensional econometric models. The authors test their findings in three empirical studies and demonstrate that even small changes in the control matrix can result in significant fluctuations in SBE estimates. In particular, the findings suggest that even small modifications by using different choices of normalizations may shift the results by more than two standard errors in the estimated values.

Additionally, the authors also applied the Hausman and Residual test to these three empirical cases. They frequently observed the rejection of sparsity assumption across the applications in high dimensional scenarios (i.e. when $p \approx n$). This empirical evidence also suggests that while SBEs may offer some efficiency gains when the number of controls approaches the sample size, these gains are often minimal and come at the cost of increased sensitivity and potential bias.

This paper shows that SBEs should be used with caution in research. These estimators are popular because they can handle large numbers of controls, but they are not robust in practice because they rely on the sparsity assumption. The authors also advise researchers to practice to justify why sparsity is plausible in their specific applications and consider the potential trade-offs between efficiency and robustness. This paper adds to the discussion on the use of modern econometric techniques, showing that it is important to think carefully about the assumptions behind such methods.

5 Experiments

The goal of my experiments is to test the findings from the original paper "The Fragility of Sparsity", in a different context. I focus on evaluating the robustness of sparsity-based estimators (SBEs) against Ordinary Least Squares (OLS) regression across different dimensions, also considering the case for $p > n$. For our analysis, we use two different datasets: **Communities and Crime Dataset** available from the UCI repository (Redmond (2011)) and **Lalonde dataset** from the doWhy package (Sharma and Kiciman (2020) and Blöbaum et al. (2024)). Additionally, I also aim to evaluate if similar concerns about the validity of sparsity assumptions arise in machine learning contexts, as they do in social science applications.

5.1 Overview

I utilized these two datasets to provide a comprehensive evaluation of the sparsity assumption under different empirical conditions. The Lalonde dataset is a well- known benchmark in causal inference offering a stable environment to validate the results of our experiments. In contrast, the Communities and Crime Dataset presents a high-dimensional data scenario with a larger number of observations, allowing us to examine the robustness of SBEs under various conditions. This includes testing fragility for different sample sizes by subsetting the data; and evaluating the predictive performance and model explainability using metrics like Mean Squared Error and R-squared.

(A) Lalonde Dataset: The Lalonde dataset looks at how a job training program affects people's earnings. The dataset has information on 445 people. It has 12 variables, including data from the treatment and control groups. This allows us to test causal inference methods. Our variables are:

1. **Treatment variable:** The "treat" variable in this dataset treatment variable which shows if someone took part in the training program (1 = yes, 0 = no). It allows us to assess if there is an impact of job training on the outcome variable.

2. **Outcome variable:** "re78" is the outcome variable for this dataset which represents real earnings in 1989. It measures post-treatment earnings to evaluate the impact of the job training program.

3. **Control variables:** The control variables for our model from this dataset include: age, educ, black, hisp, married, nodegr, re74, re75, u74, u75. These variables account for demographic factors (age, educ, black, hisp, married, nodegr), pre-treatment earnings (re74, re75), and employment status before the intervention (u74, u75), allowing us to control for the differences that could impact the outcome.

(B) Communities and Crime Dataset: The Communities and Crime dataset provides data on various socio-economic, law enforcement, and crime-related metrics for different communities in the United States. It consists of 2215 observations and 125 features where each row represents individual communities and municipalities in the United States. The dataset also consists of both categorical and numerical variables, which applies well for evaluating the results of the original paper.

1. **Treatment variable:** The "pop" variable in this dataset treatment variable represents the population of each community. It allows us to assess if there is an impact of the population levels on the crime.

2. **Outcome variable:** "violentPerPop" is the outcome variable which represents the number of crimes per 1,000 people in each community.

3. **Control variables:** The control variables in this dataset consist of a wide range of community-level indicators which include demographic variables (pop, pctBlack, pctWhite, pctAsian), socio-economic factors (medIncome, pctUnemploy, pctPoverty), variables representing police presence and composition, and other community characteristics.

Preprocessing: The Communities and Crime Dataset had to be preprocessed before we performed experiments to ensure the validity of our results. We begin with handling missing values. We dropped all columns with missing values more than 50%. Additionally, we removed all rows with these state codes ['MN', 'MI', 'IL', 'AL', 'NY', 'IA'] due to a high number of missing values in these states. We filled the rest of the missing values with the mean values of the respective columns. Additionally, we checked for outliers and solved for multicollinearity to refine our data for further analysis. These pre-processing steps significantly reduced the dataset size; the initial shape of the dataset was 2215 rows and 143 columns, which was reduced to 1892 rows and 32 columns.

5.2 Results and Findings

This section consists of the results of my experiments with different hyperparameters and methods to evaluate the performance and robustness of sparsity-based estimators(SBEs).

5.2.1 Sparsity Based Estimators with different Normalizations

In the original paper, the authors tested for the case where the number of predictors (p) is smaller than but very close to the number of observations (n). To understand the robustness of sparsity-based estimators (SBEs) across different dimensions, in these experiments, I considered three cases. Each of these cases represented a different relationship between p and n:

1. **Case 1: Original number of predictors ($p \ll n$)**
   In this scenario, the OLS estimator is expected to perform well, as it is typically robust and efficient when p is much smaller than n. The sparsity-based estimator (SBE) should also perform reasonably well.

2. **Case 2: Number of predictors close to the number of observations ($p \approx n$)**
   In this setting, the OLS estimator may become unstable. In theory, SBEs should provide more stable and efficient estimates than OLS. However, if the sparsity assumption is fragile, we may notice variability in the estimates for different types of normalizations.

3. **Case 3: Number of predictors is more than the number of observations ($p > n$)**
   In such a high-dimensional setting, OLS can no longer be used and SBE becomes a preferred choice. This case allows us to rigorously test the robustness of the sparsity assumption when p is higher than n.

Just like the original paper, we applied various feature transformations to the dataset to increase the dimensions of the data to the appropriate sizes for the second and third cases. The transformations included:

1. **Polynomial Features:** Creating higher-degree terms to capture non-linear relationships

2. **Interaction Terms:** Generating interaction features between existing predictors for interaction effects
3. **Statistical Transformations:** Applying logarithmic and square root transformations to the original features
4. **Noise Additions:** Introducing random noise to assess the robustness of SBEs when redundant features are also present

Case	Drop1	Drop2	Demean	Median	OLS
(i) Treatment Coefficient Estimate					
Original p ($p = 10$)	1670.71	1670.71	1691.39	1670.71	1670.71
p close to n ($p = 366$)	2584.34	2701.57	2425.79	2608.61	2642.21
p higher than n ($p = 499$)	2474.55	2396.02	2431.1	2505.07	-
(ii)Treatment Coefficient Standard Error					
Original p ($p = 10$)	641.13	641.13	638.67	641.13	641.13
p close to n (p = 366)	829.40	821.09	809.31	830.15	824.34
p higher than n ($p = 499$)	824.44	832.19	818.07	833.46	-
(iii) Number of Variables Selected by lasso					
Original p ($p = 10$)	10	10	8	10	-
p close to n ($p = 366$)	268	271	229	268	-
p higher than n ($p = 499$)	353	363	306	353	-

Table 1: *Estimated Treatment Coefficients and Standard Errors using different specifications for the Lalonde Dataset where* $n = 445$. The table reports treatment coefficient estimates, their standard errors, and the number of variables selected by Lasso under the three scenarios. The lasso alpha parameter is set to 1.0, the default in scikit-learn, to maintain consistency with the software defaults mentioned in the original paper.

Interpretation: The table shows how the estimated treatment coefficients, standard errors, and the number of variables selected by Lasso change with different models and normalization. Specifically:

(A) Treatment Coefficients and Standard Errors:

- When the predictors are small (p = 10), the coefficients and standard errors are stable across all normalizations and their standard errors are relatively low (around 641).

- When p gets closer to n (p = 336), both coefficients and standard errors increase, showing greater sensitivity and potential fragility
- When the number of predictors exceeds the number of observations (p=499), the treatment coefficients and standard errors continue to show variability across different normalizations. The standard errors remain high, indicating less confidence in the estimates as the dimensionality increases.

(B) Number of variables selected by Lasso:

- For a small number of predictors, lasso consistently selects all available variables, regardless of the normalizations
- As p approaches n, the number of selected variables selected varies significantly for different normalizations i.e. from 229 to as high as 271. This inconsistency reflects that different normalizations also change how many variables the lasso selects. This could highly affect our estimates. A similar behavior is seen in the case when $p > n$. This highlights the instability of the model in high-dimensional settings.

The results in Table 1 are consistent with the findings of the original paper "The Fragility of Sparsity" as it shows the sensitivity of sparsity-based estimators in high-dimensional settings. Specifically, as the number of predictors increases relative to the number of observations, both the treatment coefficient estimates and the number of variables selected by Lasso estimator demonstrate high variability. This reflects the fragility of SBEs when applied to high-dimensional data. However, the standard errors of the SBEs don't decrease as p gets larger. This contradicts the findings of the original paper on efficiency gains, which states that the standard errors of SBEs are supposed to decrease as the number of dimensions increases. Instead, we observe that the standard errors remain high, suggesting less confidence in the estimates.

However, when these tests were conducted with different Lasso regularization parameter values, the results significantly changed. In this experiment, if we were to change the alpha from 1.0 to use LassoCV, which automatically tunes the hyperparameters for Lasso, we observe very different results

When we replace the default Lasso regularization parameter (alpha = 1.0) with LassoCV, our results change significantly. The use of LassoCV leads to more stable results across all three cases, yielding similar treatment coefficient estimates and standard errors. Additionally, the number of variables selected by Lasso also reduces to a more consistent set as seen in Table 2.

Critical Note: Moreover, these results also indicate that once we use a properly tuned Lasso regularization parameter (alpha), the fragility of the SBEs for different normalization methods significantly reduces. The estimators become more stable even in high-dimensional scenarios. This means that Lasso is effectively able to handle high-dimensional scenarios with similar standard errors throughout. Thereby we can consider the fact that an appropriate regularization parameter may make a significant difference in our results.

Case	Drop1	Drop2	Demean	Median
(i) Treatment Coefficient Estimate				
Original p ($p = 10$)	1795.55	1795.55	1772.60	1795.55
p close to n ($p = 366$)	1791.82	1791.82	1794.34	1791.82
p more than n (p = 366)	1791.82	1791.82	1794.34	1791.82
(ii)Treatment Coefficient Standard Error				
Original p ($p = 10$)	631.20	631.20	632.60	631.20
p close to n (p = 366)	633.66	633.66	632.85	633.663
p more than n (p = 366)	633.66	633.66	632.85	633.663
(iii) Number of Variables Selected by Lasso				
Original p ($p = 10$)	1	1	2	1
p close to n ($p = 366$)	1	1	0	1
p more than n ($p = 366$)	1	1	0	1

Table 2: *Estimated Treatment Coefficients and Standard Errors using different specifications for the Lalonde Dataset where* $n = 445$. The table reports treatment coefficient estimates, their standard errors, and the number of variables selected by Lasso under the three scenarios. LassoCV package from a python package scikit-learn is used.

5.2.2 Testing for Sparsity

We evaluate the assumption of sparsity using the Hausman and Residual tests introduced in the original paper. The Hausman test looks at the difference in estimates between the SBE and OLS. If this difference is significant, it means the sparsity assumption does not hold. Meanwhile, the Residual test checks if the sum of squared residuals (RSS) from the lasso regression is similar to that from OLS. If this is not true, the assumption of sparsity does not hold.

Interpretation: When we test for original predictors p, the p-values for both Hausman and Residual tests are relatively high (greater than 0.1 at 10% level of significance). This means that we do not reject the null hypothesis and suggests that the difference between OLS and SBE estimates is not statistically significant. This thereby implies that the model is consistent with the sparsity assumption in this case.

When p increases in dimensions and approaches n, the Hausman test yields lower p-values which range from 0.04443 to 0.0973. Under a 10% level of significance, we would reject the null hypothesis. This suggests that the sparsity does not hold well when p increases in dimensions and approaches n.

Test	Drop1	Drop2	Demean	Median
(i) Case 1: Original p				
Hausman Test	0.133	0.133	0.177	0.177
Residual Test	0.263	0.263	0.216	0.216
(ii) Case 2: p close to n				
Hausman Test	0.0533	0.0858	0.0973	0.0444
Residual Test:	0.9981	0.9967	0.9975	0.9982

Table 3: *P-values from Hausmann and Residual Tests for Sparsity Assumption on the Lalonde Dataset.* The table presents the p-values for two tests (Hausman and Residual) under different normalizations (Drop1, Drop2, Demean, Median) across two cases: (i) the original number of predictors (p) and (ii) when p is close to the number of observations (n).The tests assess the validity of the sparsity assumption with varying dimensions. LassoCV from scikit-learn is used to determine the optimal alpha value.

However, the Residual test p-values are very high (close to 1), indicating that the test does not find evidence against the sparsity assumption in this setting and the sparsity assumption may hold. This means that our residual sum of squares (RSS) for both OLS and SBE methods are not very different. This indicates that SBE does not give a higher error even after variable selection, supporting the assumption of sparsity.

Moreover, with a small number of predictors, both tests support the sparsity assumption. However, as $p \approx n$, the Hausman test suggests rejecting the sparsity assumption, while the Residual test does not, leading to unclear interpretations. Additionally, for cases where $p > n$, these tests cannot be evaluated because the OLS benchmark is not valid.

Critical Note: The results in Table 3 introduce some ambiguity, making the conclusion less clear. While the results in Table 3 may be valid, it is also possible that this ambiguity stems from the artificial addition of features. These transformations could change the true sparse structure and make the sparsity assumption less reliable in higher dimensions. This is important to consider because even the original paper uses similar methods to increase the features of the data. The relevance of such transformations has to be analyzed more carefully, and these tests should be applied to datasets that are naturally high-dimensional from the start to better assess the validity of the sparsity assumption.

5.2.3 Evaluating the Robustness of SBEs in Machine Learning

To understand the behavior of sparsity-based estimators in a machine-learning context, I conducted two experiments. These tests aimed to test how (i) standard errors behave for varying number of observations (n) and (ii) to evaluate whether the different normaliza-

tion methods affect the predictive power of the models measured by mean squared error (MSE).

For these tests, we use the **Communities and Crime Dataset**. This dataset provides us with a large number of observations which is useful to evaluate how the results change for different sample sizes.

Note: While the communities and crime dataset consists of approximately 1,900 observations, which is smaller than typical datasets used in machine learning evaluations, it was chosen intentionally to test the sparsity assumption in high-dimensional cases. Testing the findings of the main paper, often requires the number of predictors to be extremely high, approaching the number of observations itself. Using a much larger dataset, such as one with 20,000 observations would necessitate increasing p to be around 10,000 or more. This could result in highly redundant datasets due to the excessive number of artificially added features. The Communities and Crime dataset thus serve as an ideal compromise, providing sufficient observations to evaluate the prediction performance without introducing unneeded redundancy.

(A) Testing for varying number of observations (n)
To examine the effect of varying sample sizes on the fragility of SBEs, we performed experiments using different sample sizes from the Communities and Crime dataset: the full dataset with 1,892 observations, and subsets of 150, 500, 800, 1000, and 1500. This allows us to evaluate whether the results were consistent across different sample sizes and provided insights into the robustness of SBEs for varying dataset sizes.

Number of Observations (n)	150	500	800	1000	1500	Full
Standard Error (SBE)	0.015	0.014	0.013	0.012	0.011	0.010

Table 4: *Treatment Standard Error for Communities and Crime Dataset.* The table presents the standard errors for sparsity-based estimators across different sample sizes (150, 500, 800, 1000, 1500, and full dataset). Each standard error value has been multiplied by 100% for better interpretability. The results are based on the standard case of dropping the first category along with no offset normalization.

Interpretation: Table 4 shows that as we decrease the number of observations n, the standard error also continues to increase. This pattern indicates that larger sample sizes help stabilize the estimates from sparsity-based estimators (SBEs), reducing their fragility. Smaller sample sizes on the other hand are associated with larger standard errors which indicates greater variability and less reliable estimates.

More intuitively, this result supports the fact that as the ratio of p to n reduces, the model performs better. This reinforces the importance of adequate sample size for the stability of sparsity-based estimates.

Additionally, I also tried to maintain a consistent ratio of 80% of p to n for each subset to analyze the effect of the dataset sizes on high dimensional scenarios. However, this approach made the results difficult to interpret and less comparable because it essentially created different datasets with different structures. This also highlights the challenge of drawing proper conclusions when comparing models with different predictors to outcomes ratios.

Note: The communities and crime dataset is not a standard causal inference dataset. When testing with "pop" (population) as the treatment variable, we observed that the coefficients for the treatment were very small. This outcome suggests that there is a minimal causal effect of the treatment on the outcome variable. However, this should not affect our findings on the robustness of SBEs because our focus is to study the behavior of the standard errors rather than the causal effect size itself.

(B) Predictive Performance Analysis
This experiment is aimed to evaluate the impact of SBEs within a machine learning context, where the primary focus is the prediction performance of models. For classification tasks, prediction performance can be assessed using metrics such as accuracy, while for regression tasks, metrics such as Mean Squared Error (MSE) or R-squared are usually used. Our experiments focused on these predictive measures to evaluate if fragile estimates also impact the overall MSE and R-squared of the model.

Case: Original p ($p \ll n$)	Drop1	Drop2	Drop3	Demean	Median
Mean Squared Error (MSE)	47791.44	47839.11	48066.58	46633.51	46376.03
R-Squared (R2)	0.8701	0.8699	0.8693	0.8732	0.8739

Table 5: *Mean squared error and R squared measures for SBE method predictions for Communities and Crime Dataset.* This table shows the results for the original case where p is much lower than n.

Interpretation: The Mean Squared Error (MSE) varies across different normalizations for the sparsity-based estimates. However, the magnitude of these variations is relatively small which suggests that these differences may not be very significant. Further analysis in different scenarios might be required to check if these differences are meaningful.

Further, the R squared remains consistent across the different normalization methods. This suggests that the explainability of the model as indicated by R-squared does not vary much for different normalization choices.

However, the results in Table 6 show that if we choose the high dimensional case where p is close to n, the MSE starts to fluctuate a lot. This variability is observed even for what appear to be minor decisions, such as the choice of the reference category - decisions that are often overlooked. Additionally, the MSE of OLS in such high-dimensional scenarios

only slightly fluctuates, indicating greater consistency is less fragile.

Case: p close to n ($p \approx n$)	Drop1	Drop2	Drop3	Drop4
Mean Squared Error (SBE)	1989043.14	2797171.52	1345545374.66	962625.71
Mean Squared Error (OLS)	1120277.30	1245272.190	1413348.61	1240219.09

Table 6: *Mean squared error and R squared measures for SBE and OLS predictions on the Communities and Crime Dataset.* This table shows results for the case there p is close to n.

These results may raise serious concerns about the fragility of SBEs even in the context of machine learning when we have very high dimensional data. While SBEs may perform adequately in low-dimensional settings, their fragility in higher dimensions suggests that their use in such contexts should be approached with caution. To make a more confident statement about the limitations of SBEs, it is important to conduct more experiments on other datasets, especially those with naturally high dimensions, to validate these findings and to evaluate the generalisability of the observed fragility.

6 Discussion

Building on the findings presented in *"The Fragility of Sparsity"*, our experiments focused on how different normalization methods and different dimensional settings affect the performance and stability of sparsity-based estimators (SBEs).

Our experiments confirmed the significant sensitivity of SBEs to normalization techniques, especially as the dimensionality of the data increases. As the number of predictors p approaches or exceeds the number of observations n, we observed that the estimates of SBEs become unstable, as reflected in increased standard errors and inconsistent variable selection. These findings agree with the conclusions of the original paper and highlight the fragility of sparsity assumption under certain conditions.

The results of the Hausman and Residual tests further emphasized this fragility. The Hausman test in our experiments indicated significant differences between OLS and SBEs when $p \approx n$. This suggested that the sparsity assumption may not be applicable in these scenarios. However, the residual test supported the sparsity assumption, as it found no significant difference between SBEs and OLS. It suggested that the sparsity assumption could still be valid under specific circumstances. However, these opposite results from the Hausman and the Residual test are unclear, and further analysis may be required for a better explanation of why this is true.

Our experiments for evaluating the effect of normalization choices on the predictive performance of SBEs revealed interesting patterns. It suggested that in a low dimensional scenario the prediction accuracy, measured by MSE may not be drastically affected by these normalization choices. Additionally, R-squared values remained consistent across

different normalization methods, indicating that the overall explainability power of the models did not change significantly with different preprocessing choices. However, in higher-dimensional cases, we observed a different behavior. The MSE fluctuated drastically due to seemingly minor choices, such as the selection of the reference category. This sensitivity raises concerns about the potential fragility of SBEs even in the context of machine learning. Nevertheless, these findings need to be validated across different datasets and contexts.

We also found that preprocessing the dataset is crucial for ensuring the stability of the estimators. For instance, in the Communities and Crime dataset, preprocessing steps such as outlier detection and multi-collinearity checks were essential for obtaining valid results. Without these steps, both SBEs and OLS estimators produced highly unstable coefficients. This highlights the importance of careful prepossessing in real-world datasets, to avoid misleading results.

Critical Discussion

(A) Choice of Regularization parameter for Lasso: In these experiments, the choice of hyperparameters, particularly the regularization parameter (alpha) in Lasso, proved to be critical. We evaluated two approaches: using a constant alpha and using a properly tuned alpha through a Python library.

For example, when we applied different alpha values for the same study to evaluate the fragility of sparsity (refer to Table 1 and Table 2) we observed significantly different outcomes. We observed that a properly tuned alpha significantly stabilized our estimates. This indicated that varying alpha values can substantially impact both variable selection and the stability of our estimates.

In another experiment with the Communities and Crime dataset, different values of alpha also led to significantly different results in the sparsity tests, sometimes even giving opposite conclusions. For more details on this experiment, please refer to Appendix A.

Thereby, the choice of the regularization parameter for lasso should be carefully considered and more analysis has to be done to check if the right choice of the regularization parameter can significantly reduce the problems of sparsity.

(B) Problems with artificially increasing the feature dimensions: In our experiments, as discussed in the results of Table 3, we noticed that artificially increasing the number of features could contribute to the fragility of the estimates. This is because such transformations would make the dataset highly redundant possibly causing other complexities. This artificial increase in the number of features is similar to the methodology implemented in the original paper, which we followed and attempted to replicate.

However, it is possible that other factors may have influenced the results. To evaluate whether the artificially increased features add to the fragility of the estimates, further

experiments need to be conducted, particularly on naturally high-dimensional datasets.

Limitations

A limitation of these experiments is their focus on specific datasets, normalization choices, and models. While the choices of these datasets were aligned with the original paper's methodology, they may limit the generalizability of the findings. Further validation of these results is required through additional experiments across diverse datasets and contexts, as each dataset may have unique challenges that could affect the applicability of SBEs.

One particular challenge in our study was interpreting the results of SBE estimates and coefficients for the Communities and Crime dataset. This dataset was not originally designed for causal inference. Thereby, while checking for the causal effect between the treatment and the outcome, the coefficients were particularly small, making the interpretations challenging.

7 Conclusion

Based on the analysis and findings of this seminar paper, it is evident that though sparsity-based estimators (SBEs) like Lasso regression are good for high-dimensional data, they are also very sensitive to normalization choices. Our experiments agree with the original paper's conclusion that SBEs can be unreliable, especially in high-dimensional settings, where normalization decisions can significantly impact stability of the estimates. However, our tests also suggest that these normalization choices may not drastically affect model explainability but may impact the prediction performance.

Our results show that it is important to think carefully about the choice of hyperparameters, normalization methods, and preprocessing methods. To be more confident about these results, further experiments across different datasets and diverse contexts are necessary to evaluate the impacts of these choices and methods on the estimates. Additionally, there is still a need for the development of more robust estimators that are less sensitive to normalization choices and hyperparameter settings.

In conclusion, the multitude of choices involved in the application of SBEs-ranging from normalization methods to hyperparameter settings-can significantly impact our estimates, and consequently, the overall results. Each dataset and each choice has the potential to impact the stability and reliability of the estimators. This underlines the need for meticulous considerations in practical applications and highlights the ongoing challenges for ensuring the robustness of SBEs across different contexts.

A Appendix

A.1 Evaluating for different Lasso regularization parameter specifications (cont.)

Test	Drop1	Drop2	Demean	Median
(i) alpha = 10				
Hausman Test	0.22052	0.47046	0.3609	0.2580
Residual Test	1.0	1.0	1.0	1.0
(ii) LassoCV (Cross-validated alpha)				
Hausman Test	5.84e-13	5.84e-13	3.66e-13	1.23e-13
Residual Test:	1.11e-16	1.11e-16	1.11e-16	1.11e-16

Table 7: *P-values from Hausman and Residual Tests for different normalization strategies and Lasso regularization parameters. Dataset used: Communities and Crime Dataset*

In Table 7, for the first case ($\alpha = 10$), the results indicate that we accept the null hypothesis for both tests at a 10% level of significance. However, for an alpha value that is chosen by the LassoCV method in the Scikit-learn package, the tests indicate entirely different results. The p-values suggest that we reject the null hypothesis for both tests, across all normalizations indicating the sparsity assumption does not hold.

This result highlights how different choices of lasso values can significantly affect our results, potentially leading to entirely different results.

Note: The lassoCV method chooses an extremely high alpha value than the alpha value 10. This is likely the reason for the stark contrast in the results.

A.2 Evaluating OLS behaviour

Table 8 shows the standard error of the Ordinary Least Squares (OLS) estimator as the number of predictors (p) increases. The results show that as p gets larger, the standard error increases, indicating that the OLS estimator is less reliable and stable in high-dimensional settings. This result confirms the statement that OLS gets noisy as the dimensionality of the data increases.

Number of Predictors (p)	10	54	99	188	321
Standard Error (OLS)	641.1320	694.8235	751.5491	792.3967	827.02682

Table 8: *Standard Errors of the OLS Estimator for the Lalonde Dataset across a different number of predictors.*

A.3 Evaluating the number of variables selected by lasso (cont.)

Table 9 shows the number of variables selected by the Lasso method across different normalization methods for the Communities and Crime Dataset. These results are consistent with those observed in the Lalonde dataset. For the original case, where the number of predictors (p) is smaller than the number of observations (n), the number of variables selected remains consistent across different normalization methods.

However, as the dimensionality(p) increases, there is greater fluctuation in the number of variables selected across the different normalization methods. This variation highlights the sensitivity of Lasso to the choice of normalization in high-dimensional settings, even in this case of the Communities and Crime Dataset.

Case	Drop1	Drop2	Drop3	Demean	Median
Original p ($p = 159$)	65	67	66	65	65
p close to n ($p = 1583$)	1283	1269	1270	1272	1284
p higher than n ($p = 2240$)	1441	1439	1430	1424	1449

Table 9: *Number of variables selected by Lasso for Different Normalization Methods in the Communities and Crimes Dataset. Here, n = 1,892*

Note: The number of predictors (p) was increased through feature transformation techniques which include, polynomial expansions, interaction terms, and noise additions. We did not include the coefficient values as they were very small, making them challenging to interpret meaningfully.

B Electronic appendix

All scripts and raw data used in this study are available in the following GitHub repository: https://github.com/anjalisarawgi/fragility_of_sparsity_review

The repository is structured as follows:

- **'/Data/'** contains both raw and processed data for the datasets used in the paper.
- **'/results/'** stores the results of our experiments in CSV files. It is further subdivided into folders for each dataset, with additional subfolders for specific cases.
- **'/src/'** includes has the main codebase for the experiments.
- **'main.py'** is thee primary script needed to reproduce the experimental results, once the processed data is saved in the /Data/ directory. This script calls the necessary functions from the `/src/` directory.

For detailed instructions on reproducing the results, including setting up the environment and running the code, please refer to the `README.md` file in the GitHub repository.

References

Blöbaum, P., Götz, P., Budhathoki, K., Mastakouri, A. A. and Janzing, D. (2024). Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models, *Journal of Machine Learning Research* **25**(147): 1–7.
URL: *http://jmlr.org/papers/v25/22-1258.html*

Domenico Giannone, Michele Lenza, G. E. P. (2021). Economic predictions with big data: The illusion of sparsity, *Econometrica* **89**: 2409–2437.

Donohue III, J. J. and Levitt, S. D. (2001). The impact of legalized abortion on crime, *The Quarterly Journal of Economics* **116**(2): 379–420.

Enke, B. (2020). Moral values and voting, *Journal of Political Economy* **128**(10): 3679–3729.

Ferrara, A. (2022). World war ii and black economic progress, *Journal of Labor Economics* **40**(4): 1053–1091.

Giannone, D., Lenza, M. and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity, *Econometrica* **89**(5): 2409–2437.

Hui Zou, T. H. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**: 301–320.

Jianqing Fan, R. L. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association* **96**: 1348–1360.

Peter Bühlmann, B. Y. (2003). Boosting with the l2 loss: Regression and classification, *Journal of the American Statistical Association* **98**: 324–339.

Redmond, M. (2011). Communities and Crime Unnormalized, UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC8X.

Sharma, A. and Kiciman, E. (2020). Dowhy: An end-to-end library for causal inference, *arXiv preprint arXiv:2011.04216* .

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**: 267–28.

Wüthrich, K. and Zhu, Y. (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis, *Review of Economics and Statistics* **105**(4): 982–997.