A New Sufficient Dimension Reduction Predictive Model using Maximum Entropy Covariance Estimator with Information Complexity

Kabir Opeyemi Olorede[1] Waheed Babatunde Yahya[2]

[1]Department of Statistics and Mathematical Sciences
Kwara State University, Malete Nigeria
kabir.olorede@kwasu.edu.ng

[2]Department of Statistics
University of Ilorin, Ilorin Nigeria
wbyahya@unilorin.edu.ng

Preamble

- As resultant effects of increasing digitization of our society and progress in the development of new measurement devices emanating from technological advances, scientific data have grown in both size n and complexity p.
- This "curse of dimensionality", (Bellman, 1961) hinders data visualization, reduces prediction accuracy on future observations and makes it difficult to detect the dependence between a response variable and the collection of the covariates.
- To circumvent these challenging statistical problems, three different approaches exist in statistical literature- *Variable Selection (VS)*, *Dimension Reduction (DR)* and *Sufficient Dimension Reduction (SDR)*

Preamble

- SDR is akin to VS in that they both try to reduce the number of variables that predict the response.
- VS tries to reduce the number of covariates in the vector of predictors; whereas SDR tries to reduce the predictor to a few linear combinations, or a few nonlinear functions, of all the the predictor variables.
- VS reduces the data to achieve sparsity, whereas SDR reduces the data to achieve low rank.
- DR lacks information about the response, Y, while SDR retains all regression information about Y. (Li, 2018)

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

Introduction: Sufficient Dimension Reduction (SDR)

- Consider a regression or classification problem with univariate response variable, Y, and a $p \times 1$ vector, $\mathbf{X}$ of continuous p predictors.

- SDR (Cook1994a; Cook1996; Li, 1991) involves finding a reduction $R(\mathbf{X})$ of dimension $d < p$ that captures all regression information of Y on $\mathbf{X}$ without requiring a pre-specified parametric model for $Y|\mathbf{X}$.

- SDR is a powerful tool to extract the core information hidden in the high-dimensional data, for the purpose of classifying or predicting one or several response variables based on the notion of a statistical concept called *sufficiency* (Fisher, 1922) derived from conditional independence.

Introduction: Sufficient Dimension Reduction (SDR)

- Numerous SDR methods have been proposed in the statistics literature since the seminar paper of Li (1991) on *Sliced Inverse Regression* (SIR).

- Among these methods are the *Sliced Average Variance Estimation* (SAVE, Cook and Weisberg, 1991), *Principal Hessian Directions* (PHD, Li, 1992), *Directional Regression* (Li and Wang, 2007), and the *Inverse Regression Estimation* (IRE, Cook and Ni, 2005).

- Other recently developed methods include *Minimum Average Variance Estimation* (MAVE, Xia et al., 2002), *Covariance Reduction* (CORE, Cook and Forzani 2008a), *Likelihood Acquired Directions* (LAD, Cook and Forzani 2009), *Principal Fitted Components* (PFC, Cook 2007; Cook and Forzani 2008b), and the *Envelopes model* (Cook, Li, and Chiaromonte, 2010). A more detailed list is in Ma and Zhu (2013b).

The Sufficient Dimension Reduction Methodology

- Let $d < min(n, p)$ and let $\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X} \in \mathbb{R}^p$ define smallest number of first few linear combinations of the stochastic covariate vector $\mathbf{X}$ so that

$$Y \perp\!\!\!\perp \mathbf{X}|(\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X}), \tag{1}$$

- where $\perp\!\!\!\perp$ signifies statistical independence (Dawid, 1979) which implies that Y is independent of $\mathbf{X}$ given the d linear combinations $\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X}$ of $\mathbf{X}$ by placing no restrictions on the regression in equation (1).
- If (1) is true, the linear combinations $\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X}$ are called *sufficient dimension reduction directions* or *sufficient predictors* because they contain all the regression information that $\mathbf{X}$ has about Y.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4

Introduction: Sufficient Dimension Reduction (SDR) Problem

- The SDR problem relates to estimating a low-dimensional subspace $\mathbb{S}$ of the predictor space containing the fewest proxy variables called *sufficient predictors* that can serve as substitute for $\mathbf{X}$ in the regression without loss of information and without prespecifying a parametric model.

- Let there exist a $p \times d$ matrix $\mathbb{B}$ whose columns are the smallest d linear combinations $\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X}$ of of the predictors $\mathbf{X}$, such that
$$P_r(Y \leq y|\mathbf{X}) = P_r(Y \leq y|\mathbb{B}^T\mathbf{X}) \text{ for all } y \in \mathbb{R}, \qquad (2)$$

- The conditional distribution of Y on $\mathbf{X}$ is the same as that conditional on $\mathbb{B}^T\mathbf{X}$.

- Then, the regression relation between Y and $\mathbf{X}$ can be summarized without losing information by replacing $\mathbf{X}$ by $\mathbb{B}^T\mathbf{X}$.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

What is the problem and why?

- Unfortunately, involvement of covariance matrix inversion in the basic step of SDR methods has plagued their success in applications where data sets contain high-dimensional predictors, p and undersized samples, n ($n \ll p$) due to *covariance matrix ill-conditioning* or *eigenvalue degeneracy* which poses serious challenge to computational tools.
- Another draw back of some existing SDR methods in both low- and high-dimensional settings is that estimated directions are hard to interpret.
- The existing sparse solutions lack consistency in high dimension or require data-dependent weights to the shrinkage term to force consistency (Bach, 2008; Olorede and Yahya, 2019).

What has been done about it?

- Most of the existing methods either involve non-convex optimization or time-consuming covariance optimization, which are often based on arbitrary choice of tuning parameters.
- Some other existing methods involve computationally intensive sequential framework, which are heavily dependent on the choice of bootstrap samples.
- Some other methods involve step-wise estimation of a sparse solution for each SDR direction and this does not directly imply variable selection unless an entire row of the basis matrix $(\beta_1, \ldots, \beta_d)$ is set to zero (Tan et al., 2018).
- Some other existing methods involve preliminary gene selection based on time consuming cross-validation for optimal number of principal components.

What is the presenter doing (or has done) about it? I

This work addresses the computational challenges plaguing SDR applications in high-dimensions, using a three-way hybridization of *(SDR)*, *maximum entropy principle (MEP)* and *information-theoretic measures of complexity (ICOMP)*. The specific objectives are to:

1. circumvent limited-sample-size problems in SDR applications by addressing the "loss of covariance information" based on ***Maximum Entropy (ME) Principle***;

2. propose a new maximum entropy covariance estimator (MEC) (Olorede and Yahya, 2019) for efficient SDR estimation in undersized sample problems;

What is the presenter doing (or has done) about it? II

3. propose the hybridized and smoothed covariance estimators that generate ridge-type (ℓ_2-norm) shrinkage parameters data adaptively for use in the DR step;

4. propose the used of ***information theoretic maximal entropic covariance complexity measures (ICOMP)*** (Bozdogan, 1988 and 2010) to replace the popular time-consuming cross-validation step for generating lasso-type (ℓ_1-norm) shrinkage parameters in the shrinkage regression step; and

5. propose a hybrid regression-type formulation of sufficient dimension reduction methods and shrinkage estimation, to produce sparse and accurate solutions in the large p, small n data problems.

Loss of Covariance Information (LoCI)

- The **LoCI** paradigm describes the fact that, in $n \ll p$ data applications, covariance regularization by penalizing or shrinking the sample covariance estimate S_i or its mixture with the pooled covariance estimate S_p equally all over the feature space almost always loses covariance information.
- This is because as sample covariance estimates become ill-posed in high-dimensional applications, the estimates of the corresponding largest eigenvalues are larger than the eigenvalues of the true covariance and the smaller ones are biased towards lower values.
- consequently, the regularized covariance estimate loses covariance information and hence still almost always remains irregular or unstable. **(Olorede and Yahya, Proposition 3.1)**

The Maximum Entropy Principle (MEP) I

- The principle of maximum entropy (ME) principle states that:

The probability distribution which best represents the current state of knowledge is the one with largest entropy.

- **The implication:** when we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have **(Jaynes, 1982)**.
- Let a p-dimensional sample $\mathbf{X}_i$ of class probability π_i be normally distributed with true mean μ_i and true covariance matrix Σ_i, **i.e.**$\mathbf{X}_i \sim \mathbf{N}_p(\mu_i, \Sigma_i)$.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

- The entropy $\mathbf{h}\left(\mathbf{X}_i\right)$ of such multivariate distribution is defined as the negative expected value of the natural logarithm of the probability density function of $\mathbf{X}_i$, which as emphasized by Fukunaga (1990), can be written as:

$$\begin{aligned} h(X_i) &= -E\left\{\ln\left[p(x|\pi_i)\right]\right\} \\ &= -E\left\{\ln\left[\frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right]\right]\right\} \\ &= -E\left\{-\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\left|\Sigma_i\right|-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \\ &= -E\left\{-\frac{p}{2}\ln(2\pi)\right\}-E\{-\frac{1}{2}\ln\left|\Sigma_i\right|\}-E\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \\ &= \frac{p}{2}\ln 2\pi+\frac{1}{2}\ln\left|\Sigma_i\right|+\frac{p}{2}. \end{aligned} \tag{3}$$

Algorithm 1: Maximum Entropy Covariance (MEC) Estimator

1. Find covariance $\mathbf{S}_i$ for each sample group i (or slicing category h) and the pooled sample covariance matrix $\mathbf{S}_p$.
2. Find the maximum entropic sample group covariance estimate $\mathbf{S}_{\text{me}}$.
3. Find the eigenvectors Ψ_i^{me} of the convex covariance mixture given by $\mathbf{S}_{\text{me}} + \mathbf{S}_p$.
4. Calculate the variance contribution of both $\mathbf{S}_{\text{me}}$ and $\mathbf{S}_p$ on the Ψ^{me} basis, i.e.,

$$\begin{aligned} \text{diag}(\mathbf{Z}^{\text{me}}) &= \text{diag}\left[(\Psi^{\text{me}})^T \mathbf{S}_{\text{me}} \Psi^{\text{me}}\right] = \left[\varphi_1^{\text{me}}, \varphi_2^{\text{me}}, \ldots, \varphi_p^{\text{me}}\right] && (4) \\ \text{diag}(\mathbf{Z}^{p}) &= \text{diag}\left[(\Psi^{\text{me}})^T \mathbf{S}_{p} \Psi^{\text{me}}\right] = \left[\varphi_1^{p}, \varphi_2^{p}, \ldots, \varphi_p^{p}\right] && (5) \end{aligned}$$

5. Form a new variance matrix based on the largest values, that is,

$$\mathbf{Z}_i^{\text{me}} = \text{diag}\left[\max\left(\text{mean}(\varphi_1^{\text{me}}, \varphi_1^{p})\right), \ldots, \max(\text{mean}(\varphi_p^{\text{me}}, \varphi_p^{p}))\right] \quad (6)$$

6. Form the MEC estimator, $\mathbf{S}^{\text{mec}}$ as:

$$\mathbf{S}^{\text{mec}} = \Psi^{\text{me}} \mathbf{Z}^{\text{me}} (\Psi^{\text{me}})^T. \quad (7)$$

Proposed Unified Regularized Sufficient Dimension Reduction Regression Estimation Strategy I

- Following ideas in Li (2007), we formulated the sufficient dimension reduction methods as a generalized eigenvalue problem of the form:

$$\mathbb{M}\Phi_i = \lambda_i \mathbb{G}\Phi_i \qquad \text{for } i = 1, \ldots, p \tag{8}$$

Where $\mathbb{M}$ is a nonnegative definite symmetric kernel matrix; $\mathbb{G}$ is a symmetric and positive definite matrix, often taking the form of the covariance matrix Σ_x of $\mathbf{X}$; vectors $\Phi_1, \ldots, \Phi_p$ are eigenvectors satisfying $\Phi_j^T \mathbb{G}\Phi_j$ if $i = j$, and 0 if $i \neq j$; and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are corresponding eigenvalues.

Proposed Unified Regularized Sufficient Dimension Reduction Regression Estimation Strategy II

- We give a summary of $\mathbb{M}$ and $\mathbb{G}$ matrices for SIR estimator in the following generalized eigenvalue formulation of SDR methods.
- **Sliced inverse regression (SIR):**

$$\mathbb{M} = cov[E\{\mathbf{X} - E(\mathbf{X}) \mid Y\}] \qquad \mathbb{G} = \Sigma_x \tag{9}$$

Without loss of generality, the generalized eigenvalue problem in Equation (8) becomes principal component analysis when $\mathbb{M} = \Sigma_x$ and $\mathbb{G} = \mathbf{I}_p$.

Proposed Unified Regularized Sufficient Dimension Reduction Regression Estimation Strategy III

- Under appropriate assumptions on the marginal distribution of **X** for model (9), it can be shown that the eigenvectors $\{\Phi_1, \ldots, \Phi_p\}$ in Equation (8) that correspond to the nonzero eigenvalues $\{\lambda_1 \geq, \ldots, \geq \lambda_d > 0\}$ form a basis for the central subspace under inquiry.
- The structural dimension, d is treated as known in the following derivation of sparse dimension reduction for large p, small n problems.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

Algorithm 2: Proposed Modified Alternating Minimizing

Step 1. Choose the usual sufficient dimension reduction estimator with no lasso constraint as an initial value for α.

Step 2. Given fixed α, solve d independent lasso problems to obtain the estimate of $\beta = (\beta_1, \ldots, \beta_d)$

Step 3. For fixed β, carry out singular value decomposition of $\mathbb{G}^{-1/2}\mathbb{M}\beta = \mathbf{UDV}^T$, and update $\alpha = \mathbb{G}^{-1/2}\mathbf{UV}^T$.

Step 4. Repeat Steps 2 and 3 until β converges.

Step 5. Normalize β as $\beta_j = \beta/\|\beta_j\|$, $j = 1, \ldots, d$.

SYMPOSIUM ON DATA SCIENCE & STATISTICS
BEYOND BIG DATA: SHAPING THE FUTURE
VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

The Sparse Sufficient Dimension Reduction Estimator I

- In the SSIR-MEC procedure, the eigenvalue problem (8) was first transformed to regression-type optimization problem based on Proposition (I) in Li (2007). The lasso (Tibshirani, 1996) shrinkage parameters were obtained data-adaptively based on covariance complexity rather than through arbitrary time-consuming parameter tuning using cross-validation.
- As a consequence of the lasso constraint in Proposition 5 in Olorede and Yahya (2019, not shown here), the resulting estimator in Algorithm II is expected to have some coefficient shrunk to zero, which leads to easier interpretation.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

The Sparse Sufficient Dimension Reduction Estimator II

- Choice of the tuning parameters λ_1 and λ_2 were made data-adaptively from the smoothed and hybridized version of the MEC (HSMEC) and the Bozdogan's maximum covariance complexity measures (ICOMP, Bozdogan, 1988 and 2010), respectively.
- The major contributions in this work include replacement of the usual maximum likelihood covariance estimator $\mathbb{G} = \Sigma_x$ with the suitable smoothed and hybridized covariance estimators, to circumvent covariance ill-conditioning in high-dimensional data applications.

The Sparse Sufficient Dimension Reduction Estimator III

- In the regularization stage, we replaced the choice of Akaike-type model selection criterion for choosing tuning parameters λ_1 and λ_2 from among several arbitrary values in two novel statistically meaningful ways. Parameter λ_2 is obtained data-adaptively as the shrinkage value from hybridized smoothe covariance matrices without cross-validation.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

Bozdogan's Information Theoretic Measure of Covariance Complexity (ICOMP) I

- Bozdogan's ICOMP was defined as:

$$C_{1F}(\hat{\Sigma}) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{p}(\lambda_j - \bar{\lambda}_a)^2 \tag{10}$$

  Where $\bar{\lambda}_a$ is the arithmetic mean of the eigenvalues. $C_{1F}(\Sigma)$ is a second-order equivalent measure of complexity.

- It is scale-invariant and $C_{1F}(\Sigma) \geq 0$ with $C_{1F}(\Sigma) = 0$ only when all $\lambda_j = \bar{\lambda}_a$. $C_{1F}(\Sigma)$ measures the relative variation in the eigenvalues rather than absolute variation of the eigenvalues.

Bozdogan's Information Theoretic Measure of Covariance Complexity (ICOMP) II

- However, because $C_{1F}(\Sigma)$ values can be greater than 1 and λ_1 cannot exceed 1, relative weights:

$$\mathbf{W}_i = \frac{e^{-(C_{1F} - min(C_{1F_i}))/2}}{\sum_{i=1}^{p} e^{-(C_{1F_i} - min(C_{1F_i}))} / 2}, \tag{11}$$

were computed. Each weight is interpreted as probability that a given C_{1F} is most appropriate. Where i indexes p dimensions evaluated and C_{1F} denotes maximal complexity measure of covariance matrix.

Data Set: Quarter-Hourly Smart Meter Data

- This data set (Omitaomu2013, Omitaomuetal2012a, Omitaomuetal2012b) contains quarter-hourly (15 minute) electricity consumption records arising from a sample of $n = 56$ housing units in the city of Knoxville, Tennessee, USA, for a period of one month spanning a vector $X = (X_1, \ldots, X_p)^T$ with $p = 2975$ input variables.

- The response (outcome) variable vector $Y = (Y_1, \ldots, Y_n)^T$ representing average electricity consumption (as used for regression example. See results in Table 1) of the $n = 56$ clients (residents) from the input data was manually curated from the data by obtaining average of all the 2975 quarter-hourly (15 minutes) loads per client per day for the 31 days in the month of January, 2010.

Quarter-Hourly Electricity Load Profile Data

- Instead of the raw average consumption, the response variable Y here is the binary electricity profile/consumption status $(0 = \text{low}, 1 = \text{high})$ manually curated from the raw smart meter records, as used for classification example (see results in Table 2).
- Any housing unit whose average monthly electricity consumption is above the overall average electricity consumption for the $n = 56$ housing units is considered to have high quarter-hourly electricity load profile and low quarter-hourly load profile otherwise.

Results I

- In the regression example in Table 1, the use of the HSMEC estimator in dimension reduction step successfully prevented SIR from breaking down in the high-dimensional application.

	NUM1	NUM2	$\widehat{\beta}_1 COR$	$AdjR^2$	P.VALUE	RMSE
SDR-HSMEC	0	0	0.99	0.98	2.2e-16	0.04
SSDR-HSMEC	2679	2851	0.98	0.95	2.2e-16	0.03

Table 1: OLS Results with first SIR- and SSIR-HSMEC predictors

Results II

- The proposed estimation method achieved excellent regression performances including adjusted coefficient of determination ($AdjR^2$) of 98%, tiny root mean square error (RMSE) of prediction of just 4% with only the first SIR-HSMEC estimate.

- NUM1 and NUM2 represent number of shrunken covariates by the SIR-HSMEC and sparse SIR-HSMEC (SSIR-HSMEC) dimension reduction steps.

Results III

- The usefulness usefulness, prediction efficiency and flexibility of the SIR-HSMEC in limited-sample-size application is further evident in the estimated high absolute correlation of 99% between the estimated first SIR-HSMEC direction and the response Y.
- The tiny p.value 2×10^{-16} also reveals that SIR-HSMEC estimated linear term of the single index model is significant.
- The final model is

$$\hat{y}_i = 0.0074 - 0.0204\mathbb{B}_{1i} \tag{12}$$

Results IV

- Equation 12 utilized the entire $p = 2975$ covariates in estimating the SIR-MEC directions while the sparse version SSIR-MEC utilized only 296 and 124 covariates in estimating the first two SIR directions after shrinking $NUM1 = 2679$ and $NUM2 = 2851$ original covariates, respectively.
- This yielded improved prediction performance and enhanced model interpretation of the estimated directions for the final model.

Results V

- In addition to model interpretability, the SSIR-MEC model 13 achieved $AdjR^2$ and RMSE of 95% and 0.03, respectively. The estimated sparse SSIR-MEC first direction has a strong association $\hat{\mathbb{B}}_1 \, COR = 0.98$ with the response variable, Y as expected.

$$\hat{y}_i = 0.0449 - 0.0441\mathbb{B}_{1i} \tag{13}$$

- These inferences are self-revealing on Figure 1.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

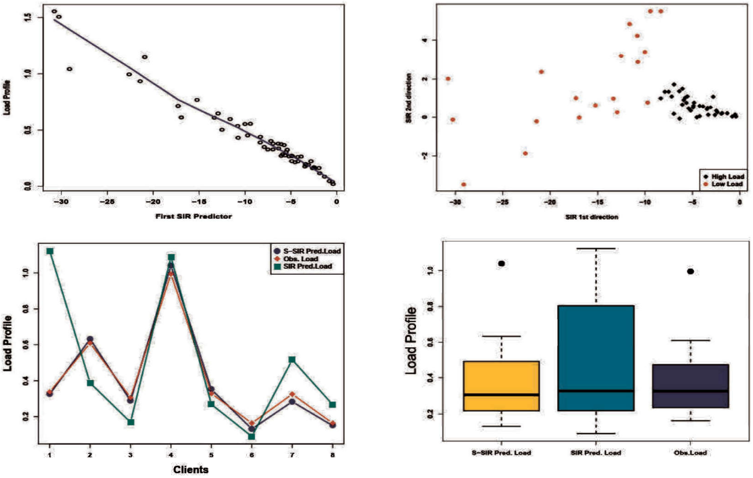

Figure 1: Summary plot for electricity smart meter data. (a) Shows line plot of Y versus SSIR-MEC 1st predictor (b) shows the 1st two sparse SIR-MEC estimates (c) shows overlayed prediction of Y, SIR-MEC and SSIR-MEC (d) shows boxplot of Y, SIR-MEC, and SSIR-MEC load prediction

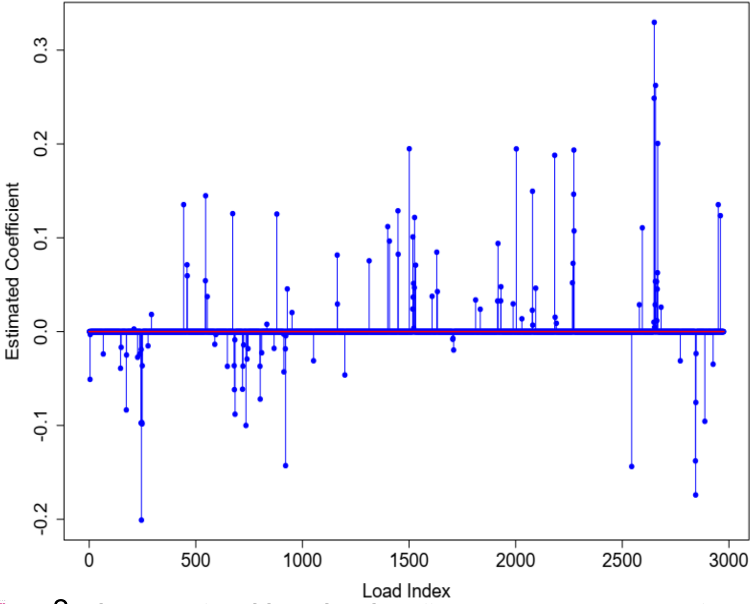

Figure 2: Stem plot of 1st SSIR-HSMEC coefficient showing sparseness of solution.

Classification Example

- In the classification example (See results in Table 2), SDR was performed using the proposed SIR-HSMEC and SSIR-HSMEC methods.
- Only the first estimated sufficient predictor was used as a single covariate in the reduced data to classify the 56 household according to their electricity consumption rates.
- Using 80:20 holdout data partition scheme, 80% of the reduced data were randomly sampled to train seven standard statistical classifiers for the classification task.
- The held out 20% samples were used to validate the classifiers models using several performance metrics.

Classification Results

Metrics	Logistic	Random Forest	QDA	LDA	NB	1-NN	Ctree
Accuracy	0.91	0.91	1.00	0.91	1.00	1.00	0.91
Sensitivity	0.75	0.75	1.00	0.75	1.00	1.00	0.75
Specificity	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Balance Accuracy	0.88	0.88	1.00	0.88	1.00	1.00	0.88
G-Mean	0.87	0.87	1.00	0.87	1.00	1.00	0.87
False Positive Rate	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Positve Predictive Vaue	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Negative Predictove Value	0.88	0.88	1.00	0.88	1.00	1.00	0.88
Area Under the Curve	93.75	93.75	100.00	93.75	100.00	100.00	93.75

QDA: Quadratic Discriminant Analysis, **LDA:** Linear Discriminant Analysis, **NB:** Naïve Bayes,
CTree: Classification Trees, **k-NN:** k-Nearest Neighbor with 1 optimal neighbor

Table 2: Binary classification performances with 1st sufficient predictor

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

Conclusion I

- The unified estimation methods developed in this work and the new proposals have contributed majorly to existing sufficient dimension reduction knowledge and application in terms of better prediction of future observations, mitigation of dimensionality issues and data visualization without loss of information.
- the proposed MEC estimator utilizes the most stable and informative convex mixture of covariance matrices to achieve highest classification and regression accuracy in statistical covariance based methods in limited-sample-size problems.
- The proposed MEC estimator fully addresses loss of covariance information in ultrahigh regression and classification problems.

Conclusion II

- The proposed MEC estimator can also be used in applications other than SDR to circumvent covariance ill-conditioning.
- This work has introduced the use of hybridized smoothed maximum enropy covariance estimator (HSMEC) for the first time (to the best of our knowledge) and proposed a new SDR-HSCE modeling approach to efficiently circumvent covariance singularity and eigenvalue degeneracy that plagues SDR applications with high-dimensional data.
- The eigenvalues of the HSMEC are well-conditioned and are positive definite thereby providing a positive and invertible plug-in covariance matrix needed in the SDR step.

Conclusion III

- it is believed that proposed unified estimation strategy is a viable means of classification, dimension reduction regression, and data visualization in undersized high-dimensional applications.
- It is a new approach that is extensible and applicable to other supervised statistical learning problems.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

References I

- Bach, F. R. (2008). Bolasso: Model Consistent Lasso Estimation through the Bootstrap. Paper presented at 25th International Conference on Machine Learning, Helsinki, Finland, July 5–9, pp. 33–40.
- Bellman, R. (1961). Adaptive Control Processes: A Guided Tour. *Princeton University Press*, Princeton, NJ.
- Bozdogan, H. (1988). ICOMP: A new model-selection criterion. In classification and related methods of data analysis. Elsevier Science Publishers, Amsterdam. H. H. Bock (Ed.) series, 599:608.
- Bozdogan, H. (2010). Shrinkage Covariance Estimators. Unpublished Lecture Notes

References II

- Cook, R.D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association* **89**:177-190.
- Cook R. D., Forzani L. (2008a). Covariance Reducing Models: An Alternative to Spectral Modelling of Covariance. Biometrika, 95(4), 799-812.
- Cook R. D., Forzani L. (2009). Likelihood-Based Sufficient Dimension Reduction. Journal of the American Statistical Association, 104(485), 197-208.
- Cook R. D., Forzani L. (2008b). Principal Fitted Components in Regression. Statistical Science, 23(4), 485-501.

References III

- Cook R. D., Li B., Chiaromonte F. (2010). Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. *Statistica Sinica*, 20(3), 927-1010.
- Cook, R.D. (1996). Graphics for Regression with a Binary Response. *Journal of the American Statistical Association* **91**:983-992.
- Dawid, A. P. (1979). Conditional Independence in Statistical Theory. Journal of the Royal Statistical Society, Series B (Methodological). Published by Royal Statistical Society, Wiley. 41(1), 1:31. ISSN:00359246, http://www.jstor.org/stable/2984718

References IV

Cook, R.D. & Ni, L. (2005). Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach. *Journal of the American Statistical Association*, **100**(470), 927-1010.

Cook, R.D. & Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association* **86**, 328-332.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Statistical Society A: Mathematical, Physical and Engineering Sciences. 222, 594-604.

References V

- Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition (2nd Ed.). Academic Press Professional Inc., USA, isbn:0122698517
- Jaynes, E.T. (1982). On the Rationale of Maximum-Entropy Methods. *Proc. IEEE*, **70**, pp. 939-952.
- Li, B. (2018). Sufficient dimension reduction: Methods and Applications with R. Monographs on Statistics and Applied Probability, *CRC Press Taylor & Francis Group* 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742.
- Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction (with Discussion). *Journal of the American Statistical Association* **86**:316-342.

References VI

- Li, L. (2007). Sparse sufficient dimension reduction. Biometrika, 94(3), 603:613.
- Li, K.-C. (1992). On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association,* **87**: 1025-1039
- Li, B. & Wang, S. (2007). On Directional Regression for Dimension Reduction. *Journal of American Statistical Association*, **102**(479), 997-1008.
- Ma, Y. and Zhu, L. (2013b). A Review on Dimension Reduction. *Int. Statist. Rev.* **81**, 134-50.

SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS BEYOND BIG DATA: SHAPING THE FUTURE VIRTUAL • JUNE 2–4, 2021

KWARA STATE UNIVERSITY KWASU Skills & Integrity

References VII

- Olorede K. O. and Yahya, W.B. (2019). A New Covariance Estimator for Sufficient Dimension Reduction in High-Dimensional and Undersized Sample Problems. Submitted to journal of Computational and Graphical Statistics,urlabs/1909.13017, https://arxiv.org/pdf/1909.13017.pdf

- Omitaomu, O. A., Budhendra, L. B., Christopher, S. M., Jeffrey, B. K., Amanda, M. N. (2012a). CoNNECT: Data Analytics for Energy Efficient CommunitiesI, American Society of Mechanical Engineers, 45226, 559-569.

- Omitaomu, O. A. (2013). Profiling Real-Time Electricity Consumption Data for Process Monitoring and Control. Institute of Industrial and Systems Engineers (IISE). In IIE Annual Conference Proceedings,

References VIII

- Omitaomu, O. A., Christopher, S. M., Ian, S. K., Jeffrey, B., Kodysh, B. L., Bhaduri, C. A. and ... Aaron, T. M. (2012b). An Integrated Geovisual Analytics Framework for Analysis of Energy Consumption Data and Renewable Energy Potentials. Proceedings of the Seventh International Conference on Geographic Information Science (GIScience)
- Tan, K. M., Wang, Z., Zhang, T., Liu, H. & Cook, R.D. (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *BiometriKa*, Oxford University Press, **105**(4): 769-782
- Xia Y., Tong H., Li W. K., and Zhu L.X. (2002). An adaptive estimation of dimension reduction space (with discussion). J. R. Stat. Soc. Ser. B.; **64**:363–410.

References IX

THANK YOU FOR LISTENING !!!