Groupwise Bayesian Dimension Reduction

Bo Zhang*, Liwei Wang†, Song Yan‡ and Chul Sung§
*IBM, Durham, NC 27703, USA
Email: bozhang@us.ibm.com
†Pharmaceutical Product Development Inc., Morrisville, NC 27560, USA
‡University of North Carolina Chapel Hill, NC 27599, USA
§IBM, Austin, TX 78758, USA

***Abstract*—Nearly all existing estimations of the central subspace in regression take the frequentist approach. However, when the predictors fall naturally into a number of groups, these frequentist methods treat all predictors indiscriminately and can result in loss of the group-specific relation between the response and the predictors. In this article, we propose a Bayesian solution for dimension reduction which incorporates such group knowledge. We place a prior whose variance is constrained to the form of a direct sum on the central subspace and directly model the response density in terms of the sufficient predictors using a finite mixture model. This approach is computationally efficient and offers a unified framework to handle categorical predictors, missing predictors, and Bayesian variable selection. We illustrate the method using both a simulation study and an analysis of a temperature data set.**

I. Introduction

Statistical modeling from big data has been driven by the advance of modern technologies. One of its challenge is high dimensional predictors. Sufficient dimension reduction (SDR) [1], which aims to reduce the dimension of predictors prior to any statistical modeling efforts, is pivotal to such big data analysis.

For regression or classification of a response Y given a p-dimensional predictors $\boldsymbol{X}$, SDR looks for a *dimension reduction subspace* $\mathcal{S}$ of $\mathbb{R}^p$ whose orthogonal basis $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ satisfies that

$$Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}, \tag{1}$$

where $d = \dim(\mathcal{S})$ and $\perp\!\!\!\perp$ denotes independence. In other words, $\boldsymbol{X}$ can be replaced by $\boldsymbol{B}^T\boldsymbol{X}$ without losing any regression or classification information of $Y|\boldsymbol{X}$. $\boldsymbol{B}$ and the corresponding dimension reduction subspace $\mathcal{S}$ are not unique. Under weak conditions, the intersection of all $\mathcal{S}$ is itself a dimension reduction subspace, in which case the intersection is called a *central subspace*.

Traditional SDR is achieved without any restriction on such central subspace. However, in a large variety of applications, predictors of interest may naturally be collected from several groups. For example, in Paleoclimatology, scientists studied high dimensional various proxies for temperature, which are available for both historical period and modern instrumental period. A common goal is to design a statistical model for the observed temperature anomalies as a function of the proxy predictors for modern instrumental period, and then apply this model to predict temperature anomalies for the historical period to determine how temperature has evolved. Traditional SDR could be applied prior to any modeling effort for such high dimensional regression. However, the fact that these various proxies come from different group conveys important information that should be leveraged in the SDR process. It is more reasonable to assume that proxies of the same type are responding the same underlying temperature effect, and thus should be grouped when estimating the central subspace.

In such applications it is more desirable to conduct dimension reduction incorporating the prior domain knowledge for the following reasons. First, the resulting sufficient predictors may lend themselves readily to interpretation consistent with prior structure information. For example, in the temperature study, it would be highly informative it would be more helpful to get reduced data separated in different groups and know that certain groups of proxies have significant associations with temperature records, rather than obtain linear combinations of all proxies as the reduced-dimensional predictors. Secondly, incorporating domain information into dimension reduction is intuitively expected to improve estimation accuracy compared to a reduction method that ignores such knowledge, in part because groupwise sufficient predictors have smaller number of parameters needed for dimension reduction. This has been verified by our simulation studies and the gain in estimation accuracy is substantial.

Although many SDR methods have been proposed, relatively little attention has been paid to incorporating the prior domain information. [2] proposed an inverse moment based groupwise SDR method subject to linear constraints which are induced by the prior group information. [3] developed another inverse moment based groupwise SDR strategy by assembling the usual SDR estimates from individual groups of predictors. [4] developed three inverse moment based folded SDR techniques (folded SIR, folded SAVE, and folded DR), conducting dimension reduction on the matrix- or array-valued predictors. However, like other inverse moment based SDR methods, all approaches rely

on the elliptical assumption of the data, which may not be fulfilled in practice. [5] investigated a kernel smoothing based groupwise SDR method, conducting dimension reduction on the predictors that fall into several group. However, the kernel smoothing method requires careful choice of bandwidth parameter, and it is usually difficult to apply if the dimensionality is very high. [6] introduced a general and scalable way to embed the prior group information hidden in the predictors to existing SDR estimation process via the direct sum envelope. [7] proposed structured Ordinary Least Squares (sOLS) for predictors with group information. And sOLS requires less computation than the above mentioned groupwise SDR methods. [8] proposed a positive definite kernel-based solution for SDR which preserves the matrix structure of the sufficient predictors. Note that all the estimators mentioned above take the frequentist approach. Thus, to our knowledge, no Bayesian method has hitherto been proposed that can simultaneously handle the SDR problem with groupwise predictors without requiring restrictive assumptions.

In this paper, we propose a *groupwise Bayesian dimension reduction* (SPDR) procedure using the Bayesian method. The goal of our method is to incorporate a domain information in the predictors, while other existing methods in their present form cannot handle. We achieve this goal by imposing a constraint on the variance of the prior putting on the central subspace.

The rest of the paper is organized as follows. In Section II, we first introduce groupwise central subspace and then describe the groupwise dimension reduction via Bayesian Mixture Modeling. The simulation studies are presented in Section III. In Section IV, we present the real application. Section V concludes with a discussion.

II. Method

A. Notation

Given two matrices $\boldsymbol{A} = [\boldsymbol{a}_1 \dots \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} = [\boldsymbol{b}_1 \dots \boldsymbol{b}_q] \in \mathbb{R}^{p\times q}$, the *direct sum* is the $(m+p)$-by-$(n+q)$ matrix

$$\boldsymbol{A} \oplus \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}.$$

B. Groupwise Central Subspace

A theoretical formulation of groupwise dimension reduction at the population level is introduced in [5] and [6]. Let $\mathcal{X}$ denote the support of transformed predictors $\mathrm{vec}(\boldsymbol{X})$. Suppose the prior domain information in the predictor can be represented by a type of decomposition of $\mathcal{X}$ as

$$\mathcal{X} = \mathcal{X}_1 \oplus \cdots \oplus \mathcal{X}_d, \tag{2}$$

where $d = \dim(\mathcal{S})$ and $\mathcal{X}_1, \dots, \mathcal{X}_d$ are subspaces of $\mathcal{X}$. The symbol $\oplus$ denotes direct sum, which is just a mathematical abstraction of dividing the predictor space into several subspaces. Then a *groupwise dimension reduction subspace* is defined as follows:

Definition 1: For a given type of decomposition $\{\mathcal{X}_1, \dots, \mathcal{X}_d\}$, if there are subspaces $\mathcal{S}_l \subseteq \mathcal{X}_l$ whose orthogonal basis $\boldsymbol{B}_l$ satisfies that

$$Y \perp\!\!\!\perp \boldsymbol{X} | (\boldsymbol{B}_1 \oplus \cdots \oplus \boldsymbol{B}_d)^T \boldsymbol{X}, \tag{3}$$

then $\mathcal{S}_l$ is called the l-th groupwise dimension reduction subspace for $Y|\boldsymbol{X}$, $l = 1, \dots, d$. Also, $\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_d$ is called a groupwise dimension reduction subspace with respect to $\{\mathcal{X}_1, \dots, \mathcal{X}_d\}$.

Then the theorem which is used to define the smallest groupwise dimension reduction subspace is described as follows:

Theorem 1: Under mild regularity conditions, if $\mathcal{S}_l^{(1)}$ and $\mathcal{S}_l^{(2)}$ are two l-th groupwise dimension reduction subspaces for $Y|\boldsymbol{X}$, then $\mathcal{S}_l^{(1)} \cap \mathcal{S}_l^{(2)}$ is itself a l-th groupwise dimension reduction subspace for $Y|\boldsymbol{X}$.

The situation here is similar to that in the classical setting of dimension reduction spaces itself a dimension reduction space. This closure under intersection makes it possible to achieve maximal groupwise dimension reduction subspace because of the intersection all l-th groupwise dimension reduction subspaces is itself a l-th groupwise dimension reduction subspace, $l = 1, \dots, d$. Then a *groupwise central subspace* is defined as follows:

Definition 2: Let $\mathcal{S}_l$ be the intersection of all l-th groupwise dimension reduction subspaces for $Y|\boldsymbol{X}$. The subspace $\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_d$ is called the groupwise central subspace and is written as $\mathcal{S}_{Y|X}(\mathcal{S}_1, \cdots, \mathcal{S}_d)$.

It is then easy to see that

$$\begin{aligned} \mathcal{S}_{Y|\boldsymbol{X}}(\mathcal{S}_1, \cdots, \mathcal{S}_d) &= \mathrm{span}(\boldsymbol{B}_1) \oplus \cdots \oplus \mathrm{span}(\boldsymbol{B}_d) \\ &= \mathrm{span}(\boldsymbol{B}_1 \oplus \cdots \oplus \boldsymbol{B}_d). \end{aligned} \tag{4}$$

Hence, $\mathrm{span}(\boldsymbol{B}_1 \oplus \cdots \oplus \boldsymbol{B}_d)$ is an equivalent definition of $\mathcal{S}_{Y|\boldsymbol{X}}(\mathcal{S}_1, \cdots, \mathcal{S}_d)$. This equivalence is the key to our method in this paper in the sense that we can incorporate the predictor information by imposing the corresponding structure on the orthogonal basis $\boldsymbol{B}$. Concretely, the statement in (1) can be further expressed in a more concise form:

$$Y \perp\!\!\!\perp \boldsymbol{X} | (\oplus_l \boldsymbol{B}_l)^T \boldsymbol{X}, \quad \text{subject to} \quad (\oplus_l \boldsymbol{B}_l)^T (\oplus_l \boldsymbol{B}_l) = I_d, \tag{5}$$

where $l = 1, \dots, d$.

C. Bayesian Method for Estimating Central Subspace

It has been revealed that the Bayesian method can be applied to estimate the central subspace [9], which we briefly review below. The basic idea of this method is to capture higher moments information by modeling the conditional distribution as a finite mixture:

$$p(Y_i|\lambda_i) = \sum_{k_d=1}^{M} \cdots \sum_{k_1=1}^{M} p_{k_1,\dots,k_d}(\lambda_i) N(\mu_{k_1,\dots,k_d}, \sigma_y^2), \tag{6}$$

where $i = 1, \ldots, n$, and the mixture weights $p_{k_1,\ldots,k_d}$ satisfy

$$\sum_{k_d=1}^{M} \cdots \sum_{k_1=1}^{M} p_{k_1,\ldots,k_d}(\lambda_i) = 1.$$

Here $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{id})^T = \boldsymbol{B}^T X_i$ is a $d \times 1$ sufficient predictor and M is some specified constant, usually between 5 and 10.

For computational and conceptual convenience, [9] modeled the mixture weights with probit form, i.e.

$$p_{k_1,\ldots,k_d}(\lambda_i) = \prod_{j=1}^{d} [\Phi(\frac{\phi_{k_j+1} - \lambda_{ij}}{\sigma_{zj}}) - \Phi(\frac{\phi_{k_j} - \lambda_{ij}}{\sigma_{zj}})], \quad (7)$$

where Φ is the standard normal distribution function and $\phi_1 < \phi_2 < \ldots < \phi_{M+1}$ are cutpoints with $\phi_1 = -\infty$ and $\phi_{M+1} = \infty$. By introducing a separate latent probit parameter for each dimension, $z_{ij}, i = 1, ..., n$ and $j = 1, ..., d$, the Model (6) can be re-written as

$$\begin{aligned} y_i &\sim N(\mu_{g_{i1},\ldots,g_{id}}, \sigma_y^2), \\ g_{ij} = k &\text{ if } \phi_k < z_{ij} < \phi_{k+1}, \\ z_{ij} &\sim N(\lambda_{ij}, \sigma_{zj}^2), \\ \mu_{k_1,\ldots,k_d} &\sim N(0, \sigma_\mu^2), \end{aligned} \quad (8)$$

where $k = 1, \ldots, M$. [9] suggested to use M euqally spaced quantiles of the $N(0, 9p)$ distribution as the cutpoints ϕ.

To implement the model, $\boldsymbol{B} = (b_{lj})_{l=1,\ldots,p;j=1,\ldots,d}$ is assumed to have prior distribution such that $b_{lj} \sim N(0, \sigma_{blj}^2)$. The central subspace is then estimated as the span of first d eigenvectors of the component-wise posterior mean of the projection matrix $P = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}$.

D. Groupwise Bayesian Dimension Reduction

Now we consider the predictors $\boldsymbol{X}$ fall natually into several groups. The traditional way to conduct dimension reduction is to treat all groups of $\boldsymbol{X}$ indiscrinimately and then seek for a dimension reduction subspace $\mathcal{S}$ whose orthogonal basis $\boldsymbol{B}$ satisfies that $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}$. However, if the information provided by the predictors $\boldsymbol{X}$ can be attained by decomposing the vector space $\mathcal{X}$ into several subspaces using the operator $\oplus$, i.e. $\mathcal{X} = \mathcal{X}_1 \oplus \cdots \oplus \mathcal{X}_d$, then we can also incorporate this information by imposing the corresponding structure on matrix $\boldsymbol{B}$ using $\oplus$, where the symbol $\oplus$ denotes direct sum. In this case, the groupwise central subspace will be $\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_d$, where $\mathcal{S}_l$ are the intersection of all l-th groupwise dimension reduction subspaces for $Y|\boldsymbol{X}$. The corresponding structure imposed on the orthogonal basis $\boldsymbol{B}$ will be $\boldsymbol{B} = \underset{l}{\oplus}\boldsymbol{B}_l$. We incorporate the special structure of $\boldsymbol{B}$ by incorporating a corresponding structure of the variance of the prior of $\boldsymbol{B}$.

To incorporate group information in the predictors $\boldsymbol{X}$, our groupwise Baysian sufficient dimension reduction (GBDR) performs Bayesian dimension reduction within each group. Suppose $b_j = (b_{1j}, b_{2j}, \ldots, b_{pj})^T$ is a sufficient predictor for group h, then we can design the model such that

$$\begin{aligned} b_{lj} &\sim N(0, \sigma_{blj}^2), \\ \sigma_{blj}^2 &= a[1 - \mathbb{1}_A(j)] + \mathbb{1}_A(j) \end{aligned} \quad (9)$$

where

$$\mathbb{1}_A(j) = \begin{cases} 1, & \text{if } j \in \{j : X_j \text{ is a group h predictor}\}, \\ 0, & \text{otherwise,} \end{cases}$$

and a is a very small positive constant, e.g. $1\mathrm{e}{-6}$. We then further extend the two-component mixture model proposed in [9] to estimate the central space via GBDR. Now, the complete model is defined as following:

$$\begin{aligned} Y_i &\sim N(\mu_{g_{i1},\ldots,g_{id}}, \sigma_y^2), \\ \mu_{g_{i1},\ldots,g_{id}} &= \sum_{j=1}^{d} \alpha_{jk}^{(j)} + \sum_{l<j} \gamma_{k_l,k_j}^{(l,j)}, \\ \alpha^{(j)} &\sim N(0, \sigma_{\alpha j}^2), \\ \gamma^{(l,j)} &\sim N(0, \sigma_{\gamma lj}^2), \\ g_{ij} = k &\quad if \quad \phi_k < z_{ij} < \phi_{k+1}, \\ z_{ij} &\sim N(\lambda_{ij}, \sigma_{zj}^2), \\ \lambda_i &= (\lambda_{i1}, \ldots, \lambda_{id})^T = \boldsymbol{B}^T X_i, \\ b_{lj} &\sim N(0, \sigma_{blj}^2), \\ \sigma_{blj}^2 &= a[1 - \mathbb{1}_A(j)] + \mathbb{1}_A(j). \end{aligned} \quad (10)$$

All variance parameters, i.e. $\sigma_y^2, \sigma_{\alpha j}^2, \sigma_{\gamma lj}^2, \ldots, \sigma_{zj}^2$, are assumed to have inverse gamma priors. Also, with such setting, the whole model remains a fully conjugate model, as [9] demonstrated, which is computationally beneficial.

III. Simulation

In this section, we conduct a simulation study to compare our proposed groupwise Bayesian dimension reduction, denoted by GBDR below, with the regular groupwise dimension reduction, denoted by BDR below. We replicate the simulation designs of [9] as follows.

- Design 1: $Y = 0.4V_1^2 + 3\text{sin}(V_2/4) + \epsilon$,
- Design 2: $Y = 3\text{sin}(V_1/4) + 3\text{sin}(V_2/4) + \epsilon$,
- Design 3: $Y = 0.4V_1^2 + \sqrt{|V_2|} + \epsilon$,
- Design 4: $Y = 3\text{sin}(V_2/4) + [1 + V_1^2]\epsilon$.

Here, the two linear predictors V_1 and V_2 are defined as $V_1 = \boldsymbol{\beta}_1^T\boldsymbol{X}$ and $V_2 = \boldsymbol{\beta}_2^T\boldsymbol{X}$, where $\boldsymbol{\beta}_1 = (1, 1, 1, 0, 0, 0)^T$ and $\boldsymbol{\beta}_2 = (1, 0, 0, 0, 1, 3)^T$. The predictors $\boldsymbol{X}$ are generated independently from the standard normal distribution. The error ϵ is normal with mean 0 and standard deviation 0.2. To evaluate the estimation accuracy of the two estimators, we use the metric in [11], which measures the distance between $\mathcal{S}_{Y|\boldsymbol{X}}(\hat{\boldsymbol{B}})$ and $\mathcal{S}_{Y|\boldsymbol{X}}(\boldsymbol{B})$ using $||\text{span}(\underset{l}{\oplus}\hat{\boldsymbol{B}}_l) - \text{span}(\underset{l}{\oplus}\boldsymbol{B}_l)||$, where $||\cdot||$ is a Frobenius norm of a matrix. This is a measure of discrepancy between the subspaces $\text{span}(\underset{l}{\oplus}\hat{\boldsymbol{B}}_l)$

and span($\oplus_l \boldsymbol{B}_l$) and hence smaller values yield more accurate estimates.

We generate 100 data sets from each model with $n = 100$. We report the mean and standard error (in parentheses) of this matrix distance over 100 data replications in the Table I.

Table I: Mean and standard error (in parentheses) of the matrix distance in the simulations

Design	GBDR	BDR
1	0.16 (0.03)	0.05 (0.01)
2	0.28 (0.06)	0.59 (0.06)
3	0.19 (0.05)	0.26 (0.04)
4	0.49 (0.02)	0.51 (0.05)

As expected, GBDR outperforms regular BDR in terms of matrix distance in all except Design 1 since we added very strong prior on those zero parameters.

IV. Application

The temperature dataset was originally given in [12], then analyzed in [13], which is by far the most comprehensive publicly database of temperature and proxies collected to date. The dataset contains 1,209 proxies, local temperatures, and global average temperatures over time. For our analysis, only Northern Hemisphere temperature data are used. Our analysis starts with the dataset given in [13]. Proxy data are available for both historical period (999-1849) and modern instrumental period (1850-1998), while observed temperature data are only available for modern instrumental period. Proxies with more than 10% missing data for years 999-1998 are excluded. The missing data were imputed by inserting the subsequent year's value. The resulting dataset contains 116 proxies. These 116 proxies belong to six different groups, which are listed in the Table II.

Table II: Proxies information for temperature data

Group	Data type	Count
1	tree composites/reconstructions	4
2	lake sediments	13
3	various composites	5
4	cave deposits	9
5	ice cores	12
6	tree rings	73
Total		116

We first reduce the dimensionality of the proxies using the proposed GBDR method for data from year 1850 to 1998. The working dimension is chosen as 1 for each group. In other words, we extract information from each group with a single linear combination and consider them as constructed covariates. Then a linear regression is fit using these 6 constructed covariates. Compared to traditional dimension reduction method such as PCA, the proposed GBDR method enables one to draw conclusions regarding each individual group. The results show that proxies from lake sediments, cave deposits, ice cores and tree rings have significant effects in predicting temperatures. From the Table III, we could see that the adjusted R-square for this linear regression using all 116 proxies (NAIVE) and GBDR sufficient covariates are respectively 0.6865 and 0.6753. As we know, if we add more useful variables, adjusted r-squared will increase. But here little information is lost by reducing the proxy dimension from 116 to 6 using the proposed GBDR method.

Table III: Linear model fit to temperature data

Method	Adjusted R-squared
NAIVE	0.6865
GBDR	0.6753

In order to assess the predictive performance of the proposed method, we performed 100 replications of five-fold cross-validation using the above described two step procedure: dimension reduction and linear regression. For the dimension reduction step, groupwise sliced inverse regression (g-SIR), assembled sliced inverse regression (a-SIR), assembled PCA (a-PCA) and our proposed groupwise Baysian dimension reduction (GBDR) are compared, using 1 as the working dimension for each group. Therefore, all dimension reduction methods provide 6 sufficient predictors for the linear regression step.

The average residual mean squared errors (RMSE) on the testing data set is shown in the Table IV. It is suggested that our groupwise Bayesian dimension reduction clearly outperforms the other methods.

Table IV: Average Residual Mean Squared Errors (RMSE) of the linear model

Method	RMSE
g-SIR	0.051
a-SIR	0.062
a-PCA	0.059
GBDR	0.042

V. Conclusion

We proposed a Bayesian groupwise SDR method subject to linear constraints which are induced by the prior group information. More complicated prior information given by the predictor could be handled by structural dimension reduction, such as, predictor contains both the matrix prior information and the group prior information; predictor includes both the matrix predictor (such as image data) and the vector predictor (such as clinical, genetic data); predictor includes some predictor needed to be shielded from dimension reduction.

ACKNOWLEDGMENT

The authors would like to thank Professor Lexin Li from University of California, Berkeley for his constructive and inspiring comments and suggestions on our research on groupwise dimension reduction using the Bayesian method.

REFERENCES

[1] R. D. Cook, "Regression graphics: Ideas for studying regressions through graphics," *New York: Wiley*, vol. 1, 1998.

[2] P. A. Naik and C.-L. Tsai, "Constrained inverse regression for incorporating prior information," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 204–211, 2005.

[3] L. Li, "Exploiting predictor domain information in sufficient dimension reduction," *Computational Statistics & Data Analysis*, vol. 53, no. 7, pp. 2665–2672, May 2009.

[4] B. Li, M. K. Kim, and N. Altman, "On dimension folding of matrix- or array-valued statistical objects," *Annals of Statistics*, vol. 38, pp. 1094–1121, 2010.

[5] L. Li, B. Li, and L.-X. Zhu, "Groupwise dimension reduction," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 1188–1201, 2010.

[6] Z. Guo, L. Li, W. Lu, and B. Li, "Groupwise dimension reduction via envelope method," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1515–1527, 2015.

[7] Y. Liu, F. Chiaromonte, and B. Li, "Structured ordinary least squares: A sufficient dimension reduction approach for regressions with partitioned predictors and heterogeneous units," *Biometrics*, 2016.

[8] B. Zhang and L. Wang, "Structure preserving dimension reduction with 2d images as predictors," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3619–3624.

[9] B. J. Reich, H. D. Bondell, and L. Li, "Sufficient dimension reduction via bayesian mixture modeling," *Biometrics*, vol. 67, no. 3, pp. 886–895, 2011.

[10] E. I. George and R. E. McCulloch, "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[11] B. Li, H. Zha, and F. Chiaromonte, "Contour regression: a general approach to dimension reduction," *The Annals of Statistics*, vol. 33, no. 4, pp. 1580–1616, 2005.

[12] M. E. Mann, Z. Zhang, M. K. Hughes, R. S. Bradley, S. K. Miller, S. Rutherford, and F. Ni, "Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia," *Proceedings of the National Academy of Sciences*, vol. 105, no. 36, pp. 13 252–13 257, 2008.

[13] B. McShane and A. Wyner, "A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1,000 years reliable?" *Annals of Applied Statistics*, vol. 5, pp. 5–44, 2011.