Expected Gradient Estimator With Many Environments

Navid Ardeshir
Department of Statistics
Columbia University
navid.ardeshir@columbia.edu

Abstract

In this paper we deal with the problem of finding an invariant subspace where the regression function only varies in that subspace across different environments. This problem can also be thought of as a representation learning with nonlinear link function. Inspired by ideas in the sufficient dimension reduction literature [Li, 2018] and recent developments in multi-task learning [Boursier et al., 2022] we propose a novel method which exploits data from all environments and provide numerical evidence for its tractability.

1 Introduction

Consider the setting where we have observed datasets from multiple environments $\mathcal{E}$. Each dataset $\mathcal{D}_e := \{(X_i^e, Y_i^e) : i \leq n_e\} \subset \mathcal{X} \times \mathcal{Y}$ contains n_e samples from the population distribution $\mathbb{P}^e \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ where $e \in \mathcal{E}$. In particular, we impose certain parametric assumptions over these distributions, that is, the response variable Y^e are only related to the covariates X^e through a linear subspace that is invariant among different environments. In other words, there exists a matrix $B \in \mathbb{R}^{d\times k}$ where $Y^e \perp\!\!\!\perp X^e \mid B^\top X^e$, and thus, the rgeression function satisfies $\mathbb{E}\left[Y^e \mid X^e\right] = \mathbb{E}\left[Y^e \;\middle|\; B^\top X^e\right]$. Although, the matrix $B \in \mathbb{R}^{d\times k}$ is not necessarily identifiable, but classical sufficient dimension reduction methods allows us to recover the column space of B denoted by $\mathscr{L} = \operatorname{span}\{B_i \in \mathbb{R}^d : 1 \leq i \leq k\} = \operatorname{col}(B)$ which we also call central subspace. Our goal in this article is to recover this central subspace by taking advantage of datasets from different environments, enabling out of distribution generalization in unseen environments with similar representations with very few samples.

Estimating the regression function under such structures are indeed more amenable and adaptive, especially when we are dealing with high dimensional datasets such that $k \ll d$, as the number of necessary samples only grows with the subpace dimension k as opposed to the ambient dimension d [Xia et al., 2002] and even parametric rates are achievable under further distributional assumptions [Yuan et al., 2023]. These approaches enjoy their adaptivity property owing to:

(i) learning the invariant subspace, enabling to project data onto a low dimensional subspace.

(ii) estimate the regression function in the projected subspace.

Note that generalization to unseen environments with only few samples is a simple consequence of reusing the invariant subspace in observed environments which can be learned by samples from the observed environments. In the following we focus on the first part and elaborate on a famously used method for learning the central subspace known as Expected Gradient Outerproduct [Li, 2018, Hristache et al., 2001]. Assuming the regression function $f_e : \mathbb{R}^k \to \mathbb{R} : z \mapsto \mathbb{E}\left[Y^e \;\middle|\; B^\top X^e = z\right]$ is differentiable and denoting $Z^e = B^\top X^e$ to be the latent variable then the main observation is that the gradient of the regression function is in the desired central subspace , i.e. $\nabla_x f_e(B^\top X^e) =$

Preprint. Under review.

$B\nabla_z f_e(Z^e) \in \mathscr{L}$. Taking the average over the poppulation distribution yields expected gradient outer product (EGOP),

$$M^{e,e} := \mathbb{E}\left[\nabla_x f_e(B^\top X^e)\nabla_x^\top f_e(B^\top X^e)\right] = B\mathbb{E}\left[\nabla_z f_e(Z^e)\nabla_z^\top f_e(Z^e)\right] B^\top \in \mathbb{R}^{d\times d}$$

Indeed the central subspace $\mathscr{L}$ is inclusive of columns space of M^e and equality only holds if the latent gradient average outerproduct $\mathbb{E}\left[\nabla_z f_e(Z^e)\nabla_z^\top f_e(Z^e)\right] \in \mathbb{R}^{k\times k}$ is invertible. Under this condition which can be shown to be equivalent to $\mathscr{L} = \mathrm{col}(B)$ being the smallest subspace for which the regression function only varies in directions of columns of B, then the central subspace $\mathscr{L}$ can be identified. Interestingly one can take advantage of other datasets and take average gradient outer product with respect to other measures to recover the central subspace more accurately so long as the resulting latent gradient outer product is inveritble. This naturally leads to defining,

$$M^{e,e'} := \mathbb{E}\left[\nabla_x f_e(B^\top X^{e'})\nabla_x^\top f_e(B^\top X^{e'})\right] = B\mathbb{E}\left[\nabla_z f_e(Z^{e'})\nabla_z^\top f_e(Z^{e'})\right] B^\top \in \mathbb{R}^{d\times d}$$

where the gradient outer product is obtained from model in environment $e \in \mathcal{E}$ but evaluated at enviromnet $e' \in \mathcal{E}$. Indeed, each individual EGOP can be used for recovering the central subspace, but the remaining challenge is how to combine them.

Our Method: We propose an estimator for central subspace using singular value decomposition of the following matrix,

$$M^{\mathcal{E}\times\mathcal{E}} = \begin{bmatrix} M^{e_1,e_1} & \dots & M^{e_1,e_T} \\ M^{e_2,e_1} & \dots & M^{e_2,e_T} \\ & \vdots & \\ M^{e_T,e_1} & \dots & M^{e_T,e_T} \end{bmatrix} \in \mathbb{R}^{dT\times dT}$$

where $T = |\mathcal{E}|$ is the number of observed environments. This matrix, under the invariant subspace assumption has low rank (less than k) and captures the interactions across different environments. More formally, let $r = \mathrm{rank}\left(M^{\mathcal{E}\times\mathcal{E}}\right)$ be the rank of this matrix then we have the following decomposition,

$$M^{\mathcal{E}\times\mathcal{E}} = U^{\mathcal{E}}\Lambda V^{\mathcal{E}\top} = \begin{bmatrix} U_1^{e_1} & \dots & U_r^{e_1} \\ U_1^{e_2} & \dots & U_r^{e_2} \\ & \vdots & \\ U_1^{e_T} & \dots & U_r^{e_T} \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \end{bmatrix} \begin{bmatrix} V_1^{e_1} & \dots & V_r^{e_1} \\ V_1^{e_2} & \dots & U_r^{e_2} \\ & \vdots & \\ V_1^{e_T} & \dots & V_r^{e_T} \end{bmatrix}^\top$$

where $U_i^{\mathcal{E}} = [U_i^{e_1\top}, U_i^{e_2\top}, \dots, U_i^{e_T\top}]^\top \in \mathbb{S}^{dT}$ are orthonormal vectors (of length one) in column space of $M^{\mathcal{E}\times\mathcal{E}}$ with $U_i^{e_j} \in \mathbb{R}^d$, and similarly $V_i^{\mathcal{E}} = [V_i^{e_1\top}, V_i^{e_2\top}, \dots, V_i^{e_T\top}]^\top \in \mathbb{S}^{dT}$ are orthonormal vectors (of length one) in row space of $M^{\mathcal{E}\times\mathcal{E}}$ with $U_i^{e_j} \in \mathbb{R}^d$. Finally, we estimate the central subspace using the column space of the following matrix,

$$\mathrm{col}\left(\left[\tfrac{1}{T}\textstyle\sum_{e\in\mathcal{E}} \frac{U_1^e}{\|U_1^e\|_2}, \quad \dots \quad , \tfrac{1}{T}\sum_{e\in\mathcal{E}} \frac{U_k^e}{\|U_k^e\|_2}\right]\right)$$

The normalization is in fact crucial in our setting and we suspect it acts as some form of "debiasing" (see Section 3 for further detail).

2 Preliminaries

2.1 Sufficient Dimension Reduction

Our focus in this work is to estimate the shared central subspace among various environments. In the following we proceed to formally define this subspace and specify some of its properties.

Definition 1 (Central subspace)**.** *For a pair (X, Y) distributed according to $\mathbb{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ we define the central subspace as the smallest subspace for which the regression function $\mathbb{E}[Y \mid X]$ only varies in that subspace, i.e.*

$$\mathscr{L} := \bigcap \left\{\mathrm{col}\,(B) \subseteq \mathbb{R}^d : \mathbb{E}[Y \mid X] = \mathbb{E}\left[Y \;\middle|\; B^\top X\right] \;\; \mathbb{P}\text{-}\textit{almost-everywhere}\right\}$$

where $\mathrm{col}(\cdot)$ represent the column space of its argumnet.

The existence of such a subspace can be guaranteed under mild assumptions over the support of X (see [Li, 2018] for further details). As discussed earlier the minimality assumption on the central subpace is crucial and implies its identifiability.

Proposition 1 (identifiability of central subapce adopted from [Li, 2018])**.** *Suppose the regression function for $(X, Y) \sim \mathbb{P}$ denoted by $f : \mathcal{X} \to \mathcal{Y} : x \mapsto \mathbb{E}[Y \mid X = x]$ is differentiable. Then, the central subspace $\mathscr{L}$ associated with $\mathbb{P}$ defined in 2 satisfies the following:*

(i) The gradient of the regression function is in the central subspace at almost every point and,

$$\text{Support}(\nabla_x f(X)) = \mathscr{L}$$

(ii) The central subspace can be recovered from the column space of the average gradient outer product (EGOP) matrix,

$$\text{col}\left(\mathbb{E}\left[\nabla_x f(X) \nabla_x^\top f(X)\right]\right) = \mathscr{L}$$

We omit the proofs for Proposition 1 (see [Li, 2018] Theorem 11.1). We may now state our central assumption.

Assumption 1 (Invariant central subspace)**.** *A collection of probability distributions $(\mathbb{P}^e)_{e \in \mathcal{E}}$ have invariant central subspaces if their corresponding central subspaces are all equal, i.e.*

$$\mathscr{L}^e = \mathscr{L} \qquad \forall e \in \mathcal{E}$$

Although this assumption may seem to be strong, our numerical experiments demonstrates robustness to small deviations of this assumption where a few of the environments do not share the same central subspaces.

In the sequel we take $P_{\mathscr{L}} \in \mathbb{R}^{d \times d}$ to be the projection matrix onto the central subspace and let $k = \text{tr}(P_{\mathscr{L}}) = \dim(\mathscr{L})$ to be the dimension of the central subspace and assume its value is known to the statistian.[1] In order to evaluate the performance of an estimator for the central subspace we use the following distance.

Definition 2 (Grassmanian distance)**.** *For a pair of subspces $\hat{\mathscr{L}}, \mathscr{L} \subseteq \mathbb{R}^d$ of dimension k we define the grassmanian distance using their projection matrices $P_{\mathscr{L}}, P_{\hat{\mathscr{L}}} \in \mathbb{R}^{d \times d}$ as,*

$$d(\hat{\mathscr{L}}, \mathscr{L}) := \left\| P_{\hat{\mathscr{L}}} - P_{\mathscr{L}} \right\|_{\text{Fr}} = \sqrt{\text{tr}\left(\left(P_{\hat{\mathscr{L}}} - P_{\mathscr{L}}\right)^2\right)}$$

Remark 1. *We remark that this distance is closely related to $\sin\theta(\cdot, \cdot)$ distance commonly used in the literature where $\theta(\cdot, \cdot)$ represents the principal angles between the two subspaces. More formally, for $i \leq k$ the i'th principal angle corresponds to*

$$\sin^2 \theta_i(\hat{\mathscr{L}}, \mathscr{L}) = \lambda_i\left(\left(P_{\hat{\mathscr{L}}} - P_{\mathscr{L}}\right)^2\right)$$

where $\lambda_i(\cdot)$ denotes the i'th eigenvalue of its argument. Thus, the distance defined in 2 corresponds to $\sqrt{\sum_{i=1}^k \sin^2 \theta_i}$ and accounts for all principal angles.

2.2 Non-parametric Estimation of Gradient Outer Product

In this section we demonstrate how we can estimate the central subspace using non-parametric kernel regression methods.

Suppose $\mathcal{H}_h$ is a reproducing kernel hilbert space induced from kernel $K_h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ indexed by a tunable bandwidth parameter $h > 0$. In particular we use the radial basis kernel $K_h : (x, x') \mapsto h^{-d/2} \exp(-\frac{\|x - x'\|_2}{2h^2})$. Given a dataset $\mathcal{D} = \{(X_i, Y_i) : i \leq n\}$ where $(X_i, Y_i) \sim \mathbb{P}$ we define the regularized kernel estimator as,

$$\hat{f}_{\eta,h} = \arg\min_{f \in \mathcal{H}_h} \sum_{i=1}^n (Y_i - f(X_i))^2 + \eta \|f\|_{\mathcal{H}_h}^2$$

[1] We later discuss further how can one estimate the value of $k = \mathscr{L}$ using data.

which can be explicitly written using Mercer's theorem as,

$$\hat{f}_{\eta,h}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i K(\cdot, X_i), \qquad \hat{\alpha} = \left(K_h(X,X) + \eta \mathbb{I}_{n\times n}\right)^{-1} Y \in \mathbb{R}^n$$

where $K_h(X,X) = (K_h(X_i,X_j))_{i,j} \in \mathbb{R}^{n\times n}$ is the kernel covariance matrix. Our EGOP estimate can be obtained via,

$$\hat{M}_{\eta,h} := \frac{1}{n}\sum_{i=1}^{n} \nabla_x \hat{f}_{\eta,h}(X_i) \nabla_x^\top \hat{f}_{\eta,h}(X_i)$$

Although, non-asymptotic rates for the regularized kernel estimator exists (e.g. [Bach, 2017] for a specific ReLU kernel), we are unaware of sample complexity guarantees for the EGOP estimate defined above, that is how many samples are needed so that $\hat{M}_{\eta,h}$ to be sufficiently close to the true expected gradient outer product. Existing results mostly deal with Nadaraya-Watson type estimates (e.g. [Yuan et al., 2023, Trivedi et al., 2014]) or local linear regression approaches (e.g. [Wu et al., 2010]) as opposed to inverse methods where the regression coefficients are obtained from the inverse of a matrix. However we use inverse methods due to their practical superiority. It is worth reiterating that we are merely concerned about consistency of the column space of the estimated EGOP matrix, and consistently estimating the regression function is not of essence in our setting.

We use kernel regression described above to obtain an estimate regression function for each environment and then form its corresponding EGOP matrices. Using sample splitting schemes we tune the hyperparameters, namely, regularization parameter η_e, and kernel bandwidth h_e, for each individual environment $e \in \mathcal{E}$ and denote the regression function estimates as $\hat{f}_e$ (see Appendix for further details).

Definition 3 (EGOP estimation per environment)**.** *Given datasets $\mathcal{D}_e = \{(X_i^e, Y_i^e) : i \le n_e\}$ where $(X_i^e, Y_i^e) \sim \mathbb{P}^e$ and it's esimate regression function $\hat{f}_e$ described above we obtain the expected gradient outer product estimates using,*

$$\hat{M}^{e,e'} := \frac{1}{n_{e'}}\sum_{i=1}^{n_{e'}} \nabla_x \hat{f}_e(X_i^{e'}) \nabla_x^\top \hat{f}_e(X_i^{e'}) \in \mathbb{R}^{d\times d}, \qquad \forall e, e' \in \mathcal{E}$$

Moreover, we define $\hat{\mathscr{L}}^{e,e'} = \operatorname{col}\left(\hat{M}^{e,e'}\right)$ to be an estimate for central subspace corresponding to $\mathbb{P}^e$.

3 Estimating the Central Subspace with Many Environments

In this section we first provide a naive approach for estimating the central subspace under the invariance assumption 1. We then elaborate on our proposed method and comment on its differences with the naive approach.

Consider a prior weight vector $w \in \mathbb{R}_+^{|\mathcal{E}|}$ over the environments satisfying $\sum_{e\in\mathcal{E}} w_e = 1$. In order to use the information from all environments and all models we aggregate the EGOP matrices using the prior w and define,

$$M^w := \sum_{e,e'\in\mathcal{E}} \sqrt{w_e}\sqrt{w_{e'}} M^{e,e'} = U^w \Lambda^w {V^w}^\top \in \mathbb{R}^{d\times d}$$

where the second equality is its singular value decomposition. In the following we establish that this aggreagation provides a feasible estimator for the central subspace.

Proposition 2. *Under the invariant assumption 1 the column space of the weighted aggregated EGOP M^w recovers the central subspace, i.e.*

$$\operatorname{col}(M^w) = \operatorname{col}(U^w) = \mathscr{L}$$

The proof is a simple extension of Proposition 1 and we defer it to appendix. Although, the choice of prior doesn't explicitly play a role in Proposition 2, however, it can be crucial in finite sample estimates. This is because one expects the convergence rates for EGOP estimates to their population

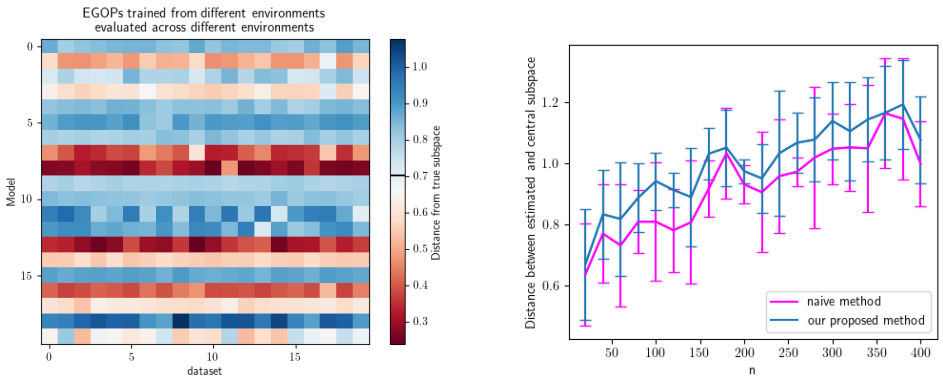

Figure 1: (left) The performance of subspace estimates $\hat{\mathscr{L}}^{e,e'}$ based on individual EGOPs $\hat{M}^{e,e'}$. Rows represent the model index and columns represent the dataset which is evaluated on. (right) The performance of our method on average dominates the naive method performance.

counterparts, i.e. $\hat{M}^w \to M^w$ when $n_e \to \infty$, to depend on the ill-conditioning of M^w, that is how far the k'th largest ingular value is from the largest. The optimal choice of prior can remedy this, yet it is unknown to the statistician apriori. We define our finite-sample naive estimator in the following using a fixed weighting choice.

Definition 4 (Naive estimator)**.** *Suppose $\hat{M}^{e,e'} \subset \mathbb{R}^{d\times d}$ to be EGOP estimates defined in 3 for a pair of environments $e, e' \in \mathcal{E}$. Then the naive estimator for central subspace can be obtained from column space of $\hat{M}^w$ with uniform prior $w_e = \frac{1}{T}$,*

$$\hat{\mathscr{L}}^{\text{naive}} = \text{col}\left(\hat{M}^w\right)$$

In order to put this estimator into perspective we resort to an optimisation representation of singular value decomposition. Recall that the i'th singular vector of a matrix can be iteratively constructed based on previous singular vectors using the following optimization problem,

$$(U_i^w, V_i^w) = \underset{\substack{u,v\in\mathbb{S}^d \\ \forall j<i,\, u^\top U_j^w = 0 \\ \forall j<i,\, v^\top V_j^w = 0}}{\arg\max} \; u^\top M^w v = \sum_{e,e'\in\mathcal{E}} \sqrt{w_e} u^\top M^{e,e'} v \sqrt{w_{e'}}$$

This is in contrast to our approach which solves a more flexible optimization problem with less constraints but tunes the weights,

$$(U_i^{\mathcal{E}}, V_i^{\mathcal{E}}) = \underset{\substack{u^e,v^e\in\mathbb{S}^d \\ \forall j<i,\, \sum_{e\in\mathcal{E}} u^{e\top} U_j^e = 0 \\ \forall j<i,\, \sum_{e'\in\mathcal{E}} v^{e'\top} V_j^{e'} = 0}}{\arg\max} \; \sum_{e,e'\in\mathcal{E}} u^{e\top} M^{e,e'} v^{e'}$$

In other words the naive estimator can be thought of as a form of parameter sharing where $u^e = \sqrt{w^e} u$ and $v^e = \sqrt{w^e} v$ for all envrionments $e \in \mathcal{E}$. This constraint might be benficial in the presence of a shared structure however the advantage of our method is that it also optimally tunes the weights to aggregate different environments.

4 Conclusion

In this paper we deal with the problem of reducing the dimenionality of the data and finding an invariant subspace among different environments. This problem can be indeed challenging as the regression function for each environment is nonlinear and may differ arbitrarily. Our setting also have close ties to multi-task and meta learning and inspired by these literatures we propose a novel method for finding such subspaces. Finally we empirically validate our method and compare it with classical methods from sufficient dimension reduction literature.

References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Etienne Boursier, Mikhail Konobeev, and Nicolas Flammarion. Trace norm regularization for multi-task learning with scarce data. In *Conference on Learning Theory*, pages 1303–1327. PMLR, 2022.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018.

Shubhendu Trivedi, Jialei Wang, Samory Kpotufe, and Gregory Shakhnarovich. A consistent estimator of the expected gradient outerproduct. In *UAI*, pages 819–828, 2014.

Qiang Wu, Justin Guinney, Mauro Maggioni, and Sayan Mukherjee. Learning gradients: predictive models that infer geometry and statistical dependence. *The Journal of Machine Learning Research*, 11:2175–2198, 2010.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3): 363–410, 2002.

Gan Yuan, Mingyue Xu, Samory Kpotufe, and Daniel Hsu. Efficient estimation of the central mean subspace via smoothed gradient outer products. *arXiv preprint arXiv:2312.15469*, 2023.

5 Appendix

5.1 Numerical Experiments

In this section we elaborate on our synthetic setting and provide evidence for admissability of our method in comparison to the naive method. Our codes can be found at https://github.com/scO0rpion/Fields-Experiment-Project.

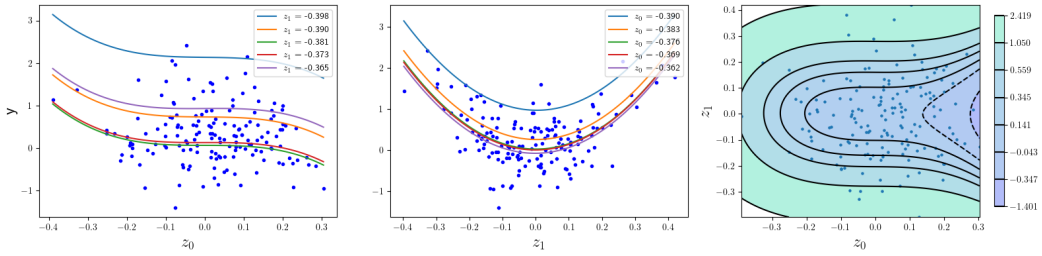

Figure 2: Demonstrates the regression function $f_e(z_0, z_1)$ for a particular environment. The dots correspond to latent data points (Z_i^e, Y_i^e) projected onto the underlying central subspace, i.e. $Z_i^e = B^\top X_i^e$. (left) The lines correspond to the graph of $z_0 \mapsto f(z_0, z_1)$ for different values of z_1. (middle) The lines correspond to the graph of $z_1 \mapsto f(z_0, z_1)$ for different values of z_0. (right) The contour plot of the regression function f_e.

Recall that our goal in this paper is to accurately recover the invariant central subspace using datasets from different environments. We generate our central subspace unfiromly at random among $k = 2$ dimenional subspaces in $\mathbb{R}^d$ with ambient dimension $d = 30$, i.e. $\mathscr{L} = \mathrm{col}(B)$ where entries of $B \in \mathbb{R}^{d\times k}$ are generated from standard gaussian. Note that we assume the knowledge of the instrinsic dimension k throughout. This value can also be estimated based on how fast the singular values of $M^{\mathcal{E}\times\mathcal{E}}$ drops (see Figure 5.1). We generate $T = 20$ environments for which their corresponding distribution $\mathbb{P}^e$ is obtained in the following way:

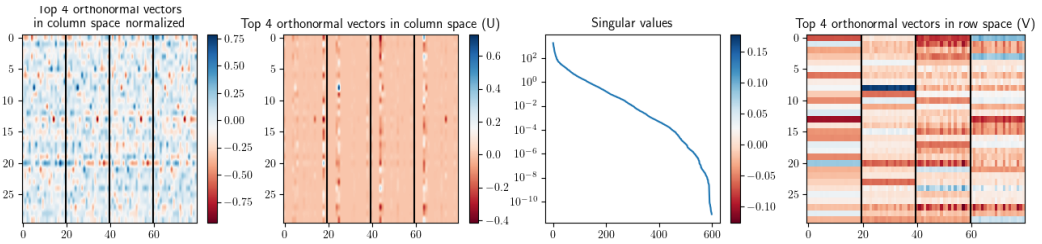

Figure 3: An illustration of singular value decomposition of $M^{\mathcal{E}\times\mathcal{E}} = U^{\mathcal{E}}\Lambda {V^{\mathcal{E}}}^\top$. Top 4 singular vectors separated by black lines. Each singular vector $U_i^{\mathcal{E}} \in \mathbb{R}^{dT}$ is reshaped into matrix in $\mathbb{R}^{d\times T}$ where e'th column represents U_i^e. (left) Each column is normalized to have length one. (left middle) Vectors in column space of $M^{\mathcal{E}\times\mathcal{E}}$. (right midle) Singular values demonstrate a sharp decay verifying $M^{\mathcal{E}\times\mathcal{E}}$ is of low rank. (right) Vectors in row space of $M^{\mathcal{E}\times\mathcal{E}}$.

1. **Random Design:** The covariates follows anisotropic gaussian $X^e \sim \mathcal{N}(0, \Sigma^e/\sqrt{d})$ where Σ^e is diagonal with diagonal elements following standard chi distribution. Similar results can be obtained when the covariates are isotropic.
2. **Polynomial Regression Function:** We consider a polynomial regression function with random coefficient generated from standard cauchy,
$$f_e(z) = a_0^e z_0^4 + a_1^e z_1^2 + a_2^e z_0 z_1$$
3. **Homogenous Noise:** Our response variable is obtained as $Y^e = f_e(B^\top X^e) + \epsilon^e$ where the noise is gaussian with variance $\sigma = 0.5$.

We generate datasets $\mathcal{D}_e$ with n samples from each environment and fit a regression function estimator $\hat{f}_e$ using kernel regression with a RBF kernel. Additionally we perform a 80-20 percent sample splitting to tune the hyperparameters. Finally, we evaluate the performance of our algorithm averaged over 20 experiments for each choice of sample size n and demonstrate the results in Figure 3.