Journal of Data Science 19 (4), 528–541
October 2021
DOI: 10.6339/21-JDS1017
Statistical Data Science

Do Predictor Envelopes Really Reduce Dimension?

Tate Jacobson[1] and Hui Zou[1,*]

[1]*School of Statistics, University of Minnesota*

Abstract

Predictor envelopes model the response variable by using a subspace of dimension d extracted from the full space of all p input variables. Predictor envelopes have a close connection to partial least squares and enjoy improved estimation efficiency in theory. As such, predictor envelopes have become increasingly popular in Chemometrics. Often, d is much smaller than p, which seemingly enhances the interpretability of the envelope model. However, the process of estimating the envelope subspace adds complexity to the final fitted model. To better understand the complexity of predictor envelopes, we study their effective degrees of freedom (EDF) in a variety of settings. We find that in many cases a d-dimensional predictor envelope model can have far more than $d + 1$ EDF and often has close to $p + 1$. However, the EDF of a predictor envelope depend heavily on the structure of the underlying data-generating model and there are settings under which predictor envelopes can have substantially reduced model complexity.

Keywords *dimension reduction; effective degrees of freedom; envelopes; Monte Carlo*

1 Introduction

Cook, Li, and Chiaromonte (2007) first introduced predictor envelopes as a method of sufficient dimension reduction for "large p" settings, where p denotes the number of input variables. Cook, Helland, and Su (2013) further developed the predictor envelope model, revealing its connection to partial least squares (PLS) and proving the efficiency gain of predictor envelopes over PLS. The predictor envelope model has become increasingly popular in Chemometrics—see Cook and Forzani (2020) which was published as the cover story of the *Journal of Chemometrics*. A substantial amount of work has been devoted to envelope models. We refer readers to Cook (2018) for a comprehensive treatment of the subject.

The predictor envelope model uses a subspace of dimension d extracted from the full space of p input variables, drawing on the response variable to find this d-dimensional subspace which is referred to as the envelope. Compared with the full model using all p variables, a predictor envelope model can achieve substantial efficiency gains. It is intuitive for users to interpret a predictor envelope with a d-dimensional subspace as having complexity $d + 1$ (the extra one corresponding to the intercept term in the model), but this could be misleading. For instance, best k-subset regression finds a subset regression model with k input variables that has the smallest residual sum of squares. The model complexity of best k-subset regression can be far greater than k because those k "best" input variables are chosen based on the data and the response variable is involved in the selection—see the simulation results in Janson, Fithian, and Hastie (2015). The same argument applies to predictor envelopes. If we knew the envelope subspace then the model complexity would be equal to that of fitting a linear model with d

*Corresponding author. Email: zouxx019@umn.edu.

Received March 9, 2021; Accepted June 8, 2021

features. In reality, we have to estimate the envelope subspace, so the actual model complexity can be different. To rigorously quantify this phenomenon, we use a more general measure of model complexity: the effective degrees of freedom (EDF) developed by Efron (1986) based on Stein's unbiased risk estimation theory. Critically, the EDF can be used to correct the downward bias of the training error as an estimator of in-sample prediction error. As such, accurate estimates of the EDF can prove invaluable to researchers seeking to compare and assess models' prediction performance.

In this paper, we explore the complexity of predictor envelopes as measured by their EDF. Due to the relatively complex nature of the predictor envelope estimator (See Section 2.2 for details), the analytical expression of its EDF is difficult to obtain. We opt to conduct a Monte Carlo study of the EDF under various model settings. Since linear models are the foundation for predictor envelopes, we limit ourselves to linear models: $Y = X\beta + \epsilon$ where $Y \in \mathbb{R}$ is the response variable, $X \in \mathbb{R}^p$ represents the set of covariates, β is the true regression coefficient and ϵ is the error term. We find that the number of predictors p and the dimension d of the envelope do not entirely determine an envelope's EDF. Rather, the structure of the predictors X and their relationship with the response Y influence the range of values that the EDF can take. On the one hand, in many but not all of our experiments, predictor envelopes are as complex as saturated linear models fit to the same p predictors, with $p + 1$ EDF. On the other hand, we identify two scenarios in which predictor envelopes indeed have much smaller EDF compared with the saturated linear model: (1) when the predictors that explain the variability in Y are highly correlated and (2) when the covariance matrix of X has a few dominating eigenvalues and the true data-generating β is linearly dependent on at least one of their corresponding eigenvectors. The EDF reductions in the first scenario are non-trivial, but fairly modest. The EDF reductions in the second scenario are far larger: a predictor envelope with the right settings can have EDF near $d + 1$, which suggests that the intuitive answer is correct in this scenario.

In Section 2 we briefly review the effective degrees of freedom and predictor envelopes. In Section 3 we discuss the ordinary Monte Carlo method for numerically computing the EDF as the design of our study at a high level. More concrete simulation models are described in Sections 4 and 5, where some "negative" and "positive" messages are reported, respectively. In Section 6 we confirm the same findings by using input variables from a real data example.

2 Preliminaries

2.1 Effective Degrees of Freedom

Suppose we fit a model to n observations $\{(X_i, Y_i)\}_{i=1}^n$, where $Y_i = f(X_i) + \epsilon_i$ and the ϵ_i have mean 0 and variance σ^2, and that this model gives us estimates $\widehat{Y}_i$. Ideally, we would like our model to make accurate predictions on new data. A model with a smaller training error may not have a smaller in-sample predictor error. It turns out that there is a model complexity term to the training error that can help correct the bias. The following theorem from Efron (1986) provides a foundation for doing so.

Theorem 1 (Efron, 1986)**.** *Let $Y_i = f(X_i) + \epsilon_i$ where the ϵ_i are independent identically distributed (i.i.d.) with mean 0 and variance σ^2 for $i = 1, \ldots, n$. Suppose that $\widehat{Y}_i$ are estimates of $f(X_i)$ from a model fit using the observed Y_i. Lastly, let $\tilde{Y}_i$ be new observations with the same*

covariate values X_i. Then

$$\mathrm{E}\left[\sum_{i=1}^{n}(\tilde{Y}_i - \widehat{Y}_i)^2\right] - \mathrm{E}\left[\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2\right] = 2\sum_{i=1}^{n}\mathrm{Cov}(Y_i, \widehat{Y}_i).$$

Define the *effective degrees of freedom* for any model-fitting method FIT_γ with tuning parameter γ as follows:

Definition 1 (Effective Degrees of Freedom)**.** Let $X \in \mathbb{R}^p$ and $Y = f(X) + \epsilon$, where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Suppose we have n observations (Y_i, X_i) and that we fit a model using FIT_γ. We define the effective degrees of freedom (EDF) for this model to be

$$\mathrm{df}(\mathbf{X}, f, \mathrm{FIT}_\gamma) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}(\widehat{Y}_i, Y_i).$$

An immediate corollary to Theorem 1 is that

$$\mathrm{E}\left[\sum_{i=1}^{n}(\tilde{Y}_i - \widehat{Y}_i)^2\right] = \mathrm{E}[\mathrm{RSS}] + 2\,\mathrm{df}(\mathbf{X}, f, \mathrm{FIT}_\gamma)\sigma^2.$$

Thus we see that $2\,\mathrm{df}(\mathbf{X}, f, \mathrm{FIT}_\gamma)$ plays the same role as $2\,\mathrm{df}$ in Mallows' C_p statistic (Mallows, 1973) for linear regression. In this way, the EDF provide a meaningful measure of model complexity and serve as a generalization of the degrees of freedom for a linear regression model. When FIT_γ is a linear smoother, meaning $\widehat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$ for some smoothing matrix $\mathbf{S}$ that does not depend on $\mathbf{Y}$, we have a closed-form expression for the EDF: $\mathrm{df}(\mathbf{X}, f, \mathrm{FIT}_\gamma) = \mathrm{tr}(\mathbf{S})$. For more sophisticated models, the analytical expression of the EDF can be difficult to obtain.

2.2 Predictor Envelopes

In this section we briefly review the predictor envelope estimation procedure introduced by Cook, Helland, and Su (2013) and further developed by Cook, Forzani, and Su (2016). Suppose we have a linear model

$$Y = \mu + \beta^T(X - \mu_x) + \epsilon \tag{1}$$

with $Y \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\epsilon \in \mathbb{R}$, $\beta \in \mathbb{R}^p$, and $X \in \mathbb{R}^p$. In this setting, X is a random vector with $\mathrm{E}[X] = \mu_x$ and $\mathrm{Var}(X) = \Sigma_X$, $\mathrm{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, and ϵ is independent of X denoted by $\epsilon \perp\!\!\!\perp X$. Before we delve into the details of how envelopes achieve predictor reduction, we need to establish some notation. Suppose we have a matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^p$. We define the $\mathbf{M}$-inner product as $\langle s, t\rangle_{\mathbf{M}} = s^T\mathbf{M}t$. We let $\mathbf{P}_{\mathcal{S}(\mathbf{M})}$ denote the projection onto $\mathcal{S}$ in the $\mathbf{M}$-inner product and let $\mathbf{Q}_{\mathcal{S}(\mathbf{M})} = \mathbf{I} - \mathbf{P}_{\mathcal{S}(\mathbf{M})}$. Note that when projecting onto a subspace in the usual inner product, we drop $\mathbf{M}$ from our notation. Lastly, let $\mathcal{S}^\perp$ denote the orthogonal complement of $\mathcal{S}$ in the usual inner product.

Suppose that we find a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that (i) $\mathrm{corr}(\mathbf{Q}_\mathcal{S}X, \mathbf{P}_\mathcal{S}X) = \mathbf{0}$ and (ii) $\mathrm{corr}(Y, \mathbf{Q}_\mathcal{S}X|\mathbf{P}_\mathcal{S}X) = \mathbf{0}$. Conditions (i) and (ii) tell us that $\mathbf{P}_\mathcal{S}X$ provides all of the information about Y that we can extract linearly from X while $\mathbf{Q}_\mathcal{S}X$ does not provide us with any additional information about either Y or $\mathbf{P}_\mathcal{S}X$. Cook, Helland, and Su (2013) proved that when (1) is the true model, conditions (i) and (ii) are equivalent to (a) $\Sigma_X\mathcal{S} \subseteq \mathcal{S}$ and $\Sigma_X\mathcal{S}^\perp \subseteq \mathcal{S}^\perp$ and (b) $\mathrm{span}(\beta) \subseteq \mathcal{S}$, respectively. The smallest subspace $\mathcal{S}$ satisfying these conditions is called

the Σ_X-envelope of $\mathcal{B} = \text{span}(\beta)$ and is denoted by $\mathcal{E}_{\Sigma_X}(\mathcal{B})$. The term "envelope" comes from the fact that $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ "envelopes" the column space of β in satisfying condition (b).

Let $d = \dim(\mathcal{E}_{\Sigma_X}(\mathcal{B}))$, where $d < p$, and $\Gamma \in \mathbb{R}^{p\times d}$ be a semiorthogonal basis matrix for $\mathcal{E}_{\Sigma_X}(\mathcal{B})$. If we know Γ, then we can express (1) as

$$Y = \mu + \alpha^T\{\Gamma^T(X - \mu_x)\} + \epsilon$$

and estimate $\mu \in \mathbb{R}$ and $\alpha \in \mathbb{R}^d$ as in standard linear regression with the reduced predictors $\Gamma^T(X - \mu_x)$. In practice, however, we need to estimate the basis matrix Γ. Let $C = (X^T, Y^T)^T$ and let $\mathbf{S}_C$ denote the sample version of $\Sigma_C = \text{Var}(C)$. Cook, Helland, and Su (2013) used a likelihood-based approach to estimate Σ_C, minimizing the objective function

$$F_d(\mathbf{S}_C, \Sigma_C) = \log|\Sigma_C| + \text{tr}(\mathbf{S}_C\Sigma_C^{-1}).$$

The structure of $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ allows us to express Σ_C as a function of Γ and a few other parameters. Let $\mathbf{S}_X$ denote the sample covariance matrix of X and s_Y^2 denote the sample of variance Y. Define $Z = s_Y^{-1}Y$ and let $\mathbf{S}_{XZ}$ denote the sample covariance between X and Z. After minimizing $F_d(\mathbf{S}_c, \Sigma_C)$ over every parameter except Γ, we only need to minimize

$$L_d(\mathbf{G}) = \log|\mathbf{G}^T(\mathbf{S}_X - \mathbf{S}_{XZ}\mathbf{S}_{XZ}^T)\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_X^{-1}\mathbf{G}|, \tag{2}$$

giving us the estimator

$$\hat{\Gamma} = \arg\min_{\mathbf{G}} L_d(\mathbf{G}).$$

For this study, we use the R package *Renvlp* (Lee and Su, 2020), which finds $\hat{\Gamma}$ using the algorithm outlined by Cook, Forzani, and Su (2016). In practice we do not know $d = \dim(\mathcal{E}_{\Sigma_X}(\mathcal{B}))$ and therefore treat d as a modeling parameter. Throughout this paper, we use Env_d to denote the predictor envelope model fitting procedure with the envelope dimension set to d.

The analytical expression of the EDF for a predictor envelope ($\text{df}(\mathbf{X}, f, \text{Env}_d)$) is difficult to obtain. Hence, in this study we use Monte Carlo to numerically compute the EDF of predictor envelopes in order to gain some insight into their complexity in different settings. We anticipate that $\text{df}(\mathbf{X}, f, \text{Env}_d)$ will vary considerably with d. If we set $d = p$, then the predictor envelope has no reduction and $\text{df}(\mathbf{X}, f, \text{Env}_p) = p + 1$. In actual applications of using predictor envelopes, d is often a small number ($d \ll p$). In such settings, we expect that $\text{df}(\mathbf{X}, f, \text{Env}_d)$ will fall between $d + 1$ and $p + 1$. If we knew Γ beforehand and fit a linear model to $\Gamma^T X$, that linear model would have $d + 1$ degrees of freedom. As such, we can think about decomposing $\text{df}(\mathbf{X}, f, \text{Env}_d)$ into the $d + 1$ degrees of freedom used to fit a linear model to $\hat{\Gamma}^T X$ and the additional degrees of freedom used in estimating the envelope.

3 Computing the EDF via Monte Carlo

We use Monte Carlo to numerically compute $\text{df}(\mathbf{X}, f, \text{FIT}_\gamma)$ for a given design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, data-generating model f, and model-fitting method FIT_γ. This is the standard approach for studying the EDF of a model when an analytical expression is not available (see, also, Janson, Fithian, and Hastie (2015)). We create M datasets with $\mathbf{X}$, f, and a pre-specified error variance $\text{Var}(\epsilon) = \sigma^2$ as follows: for each $j \in \{1, \ldots, M\}$ we generate $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$, then

set $Y_{ij} = f(X_i) + \epsilon_{ij}$. We then use method FIT_γ to obtain fitted values $\widehat{Y}_{ij}$. Using these, we estimate $\text{Cov}(\widehat{Y}_i, Y_i)$ over the M Monte Carlo iterations by

$$\widehat{\text{Cov}}(\widehat{Y}_i, Y_i) = \frac{1}{M-1}\sum_{j=1}^{M}(\widehat{Y}_{ij} - \overline{\widehat{Y}}_{i\cdot})(Y_{ij} - \overline{Y}_{i\cdot})$$

where $\overline{Y}_{i\cdot} = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$ and $\overline{\widehat{Y}}_{i\cdot} = \frac{1}{M}\sum_{j=1}^{M}\widehat{Y}_{ij}$. Since we specify σ^2 for the simulations, our estimate of $\text{df}(\mathbf{X}, f, \text{FIT}_\gamma)$ is simply

$$\widehat{\text{df}}(\mathbf{X}, f, \text{FIT}_\gamma) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\widehat{\text{Cov}}(\widehat{Y}_i, Y_i).$$

Throughout this study we set $M = 400$ and $\sigma^2 = 1$.

We use different design matrices $\mathbf{X}$ across our simulations. Though we limit ourselves to linear data-generating models, $f(X) = X\beta$, which is the foundation for predictor envelopes, we vary β across simulations as well. We want to understand how the structure of X and f impact the EDF rather than simply find the EDF for specific $\mathbf{X}$ and β. As such, we run each case 100 times, using the same simulation settings to generate different design matrices $\mathbf{X}_k$ and different coefficient vectors β_k for each of the 100 iterations. This gives us 100 different estimates $\widehat{\text{df}}(\mathbf{X}_k, \beta_k, \text{FIT}_\gamma)$ from design matrices and true coefficient vectors generated using the same settings. Throughout Sections 4, 5, and 6 we provide the mean and 1-standard deviation bars for the EDF for each set of simulation settings. To emphasize the role of the underlying structure of the data in shaping the EDF we report the mean of the estimates for each combination of settings for X and β:

$$\widehat{\text{df}}(X, \beta, \text{FIT}_\gamma) = \frac{1}{100}\sum_{k=1}^{100}\widehat{\text{df}}(\mathbf{X}_k, \beta_k, \text{FIT}_\gamma).$$

4 Envelope Complexity in Different Cases

In this section, we examine three cases to find how the structure of the joint distribution of (X, Y) and the envelope dimension impact $\text{df}(X, f, \text{Env}_d)$:

Case 1. X has a compound symmetric covariance matrix.
Case 2. X has a block covariance matrix with 5 predictors in each block.
Case 3. X has an order-1 auto-regressive (AR1) covariance matrix.

Let $\Sigma_\rho = \text{Cov}(X)$. We can express the first three correlation structures as follows:

Case 1. $(\Sigma_\rho)_{ij} = \rho$ if $i \neq j$, $(\Sigma_\rho)_{ii} = 1$ for all i.
Case 2. $(\Sigma_\rho)_{ij} = \rho$ if $i \neq j$ and $\lfloor\frac{i-1}{5}\rfloor = \lfloor\frac{j-1}{5}\rfloor$, $(\Sigma_\rho)_{ij} = 0$ if $i \neq j$ and $\lfloor\frac{i-1}{5}\rfloor \neq \lfloor\frac{j-1}{5}\rfloor$, and $(\Sigma_\rho)_{ii} = 1$ for all i.
Case 3. $(\Sigma_\rho)_{ij} = \rho^{|i-j|}$ for all i, j.

We vary parameter ρ from 0 to 0.8 in each case. In all three cases, we generate i.i.d. $X_i \sim N(0, \Sigma_\rho)$ and $\epsilon_i \sim N(0, 1)$ with $p = 50$ and $n = 150$. In the compound symmetric and AR1 cases, we generate $\beta_j \sim Gamma(2, 2)$ for 5 randomly selected $j \in \{1, \ldots, p\}$ and set the 45 remaining β_j to 0. In the block covariance case, we generate $\beta_j \sim Gamma(2, 2)$ for $j \in \{1, \ldots, 5\}$, so that the significant predictors all fall in the same block, and set the 45 remaining β_j to 0. In all three cases, we fit predictor envelopes with envelope dimensions ranging from $d = 1$ to 10.

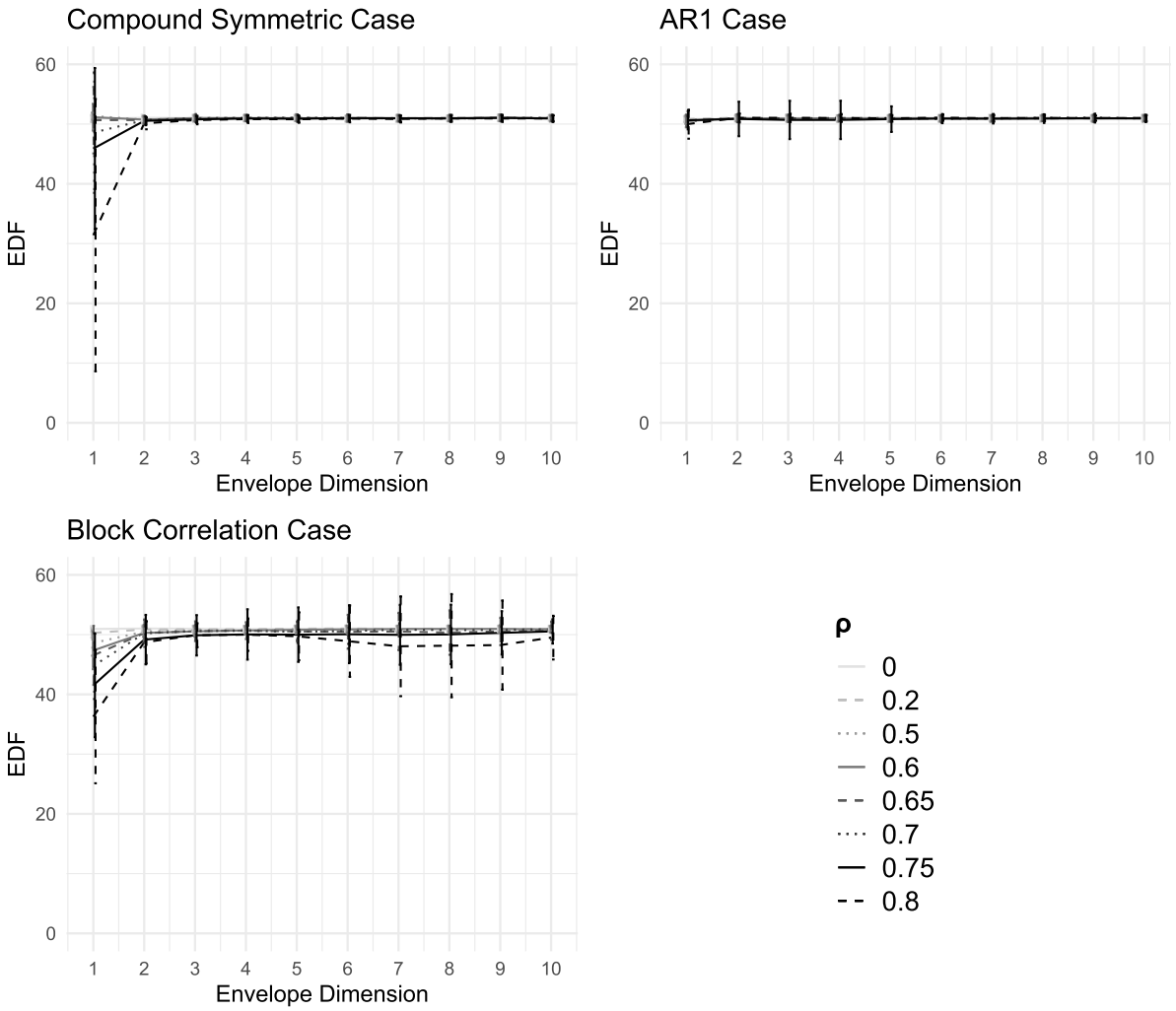

Figure 1: Envelope EDFs with different predictor correlation structures.

Figure 1 shows EDFs for predictor envelopes in the three cases outlined above. These plots include 1-standard deviation bars to show the variation across Monte Carlo iterations with different β_k. We see modest EDF reductions when $d = 1$, X has either a compound-symmetric or block covariance structure, and the informative predictors are highly correlated. All three of these conditions appear to be necessary here, as the envelopes' EDF reductions quickly disappear when either the informative predictors are not as strongly correlated or $d > 1$. Moreover, we can see that there is a considerable amount of variation in the EDF in the cases where we see some reduction, indicating that β_k also affects $\widehat{\text{df}}(\mathbf{X}_k, \beta_k, \text{Env}_d)$. It is precisely because of this variation that we choose to report the average EDF across X_k and β_k generated with the same settings: if we picked a single $\mathbf{X}$ and β to use across all 100 iterations, we might not be able to generalize beyond that case.

To ensure that these results are not limited to one choice of n and p, we rerun the compound symmetric case with $n = 200$ and $p = 100$. Figure 2 shows that the EDFs for predictor envelopes in this case follow the same pattern as when $n = 150$ and $p = 50$.

While there are a couple of scenarios in which EDF reductions are worth highlighting, we

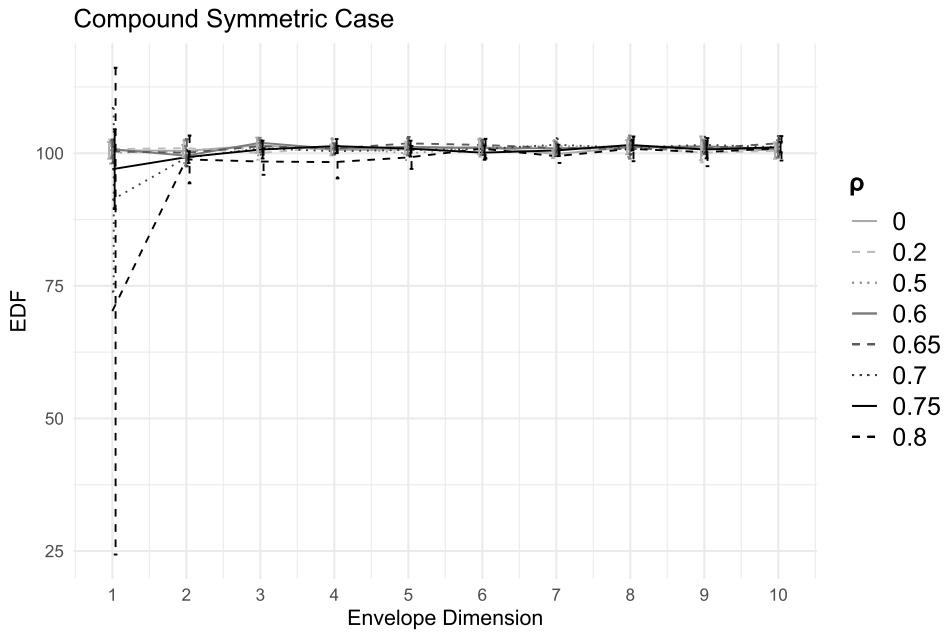

Figure 2: Envelope EDFs in the compound symmetric case with $n = 200,\ p = 100$.

should not overlook that $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ is approximately $p + 1$ in most of the settings. In the AR1 simulations, $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d) \approx p+1$ for all values of d, with little variation across simulations with different β_k. This suggests that the additional degrees of freedom used in estimating the envelope subspace $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ are nearly $p-d$ in these cases. Returning to Definition 1, this tells us that the algorithm used to estimate $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ is relying heavily on information from the response, as the fitted values $\widehat{Y}_i$ and the observed Y_i are highly correlated. In the compound symmetric and block correlation cases, on the other hand, the algorithm relies less on information from the response and more on information from the correlation structure of the predictors, leading to lower EDFs.

5 Settings for Reducing Envelope Complexity

Given the relatively "negative" message from Section 4, it is worth exploring whether there are scenarios in which a fitted predictor envelope model truly has substantially reduced EDF. In this section we consider the following set of three conditions: (1) a few eigenvalues of $\mathrm{Cov}(X) = \Sigma_X$ dominate its remaining eigenvalues, (2) the envelope dimension d is no bigger than the number of eigenvectors with dominating eigenvalues, and (3) the true data-generating β is linearly related to a least one of the eigenvectors with a dominating eigenvalue.

We recall from Section 2.2 that $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ is the smallest subspace $\mathcal{S}$ such that (a) $\Sigma_X\mathcal{S} \subseteq \mathcal{S}$ and $\Sigma_X\mathcal{S}^{\perp} \subseteq \mathcal{S}^{\perp}$ and (b) $\mathrm{span}(\beta) \subseteq \mathcal{S}$. A subspace $\mathcal{A}$ generated by orthogonal eigenvectors of Σ_X satisfies (a) (Cook, Helland, and Su, 2013). If β falls in the space generated by those eigenvectors, then (b) is satisfied as well. In this case, the envelope subspace $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ falls within $\mathcal{A}$ and the true envelope dimension d^* is equal to the number of generating eigenvectors. As such, we anticipate that the EDF of predictor envelopes can drop to near $d + 1$ when all three of our conditions are satisfied.

Suppose that Σ_X has p normalized, orthogonal eigenvectors $v_1, \ldots, v_p$ with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$. We simulate the following cases:

Case 1. $\beta = v_1$ and $\lambda_1 \gg \lambda_i$, $i = 2, \ldots, p$.
Case 2. $\beta = v_1 + v_2$ and $\lambda_1 = \lambda_2 \gg \lambda_i$, $i = 3, \ldots, p$.
Case 3. $\beta = v_1$ and $\lambda_1 = \lambda_2 \gg \lambda_i$, $i = 3, \ldots, p$.
Case 4. $\beta = v_1 + v_2$ and $\lambda_1 \gg \lambda_i$, $i = 2, \ldots, p$.
Case 5. $\beta = v_2$ and $\lambda_1 \gg \lambda_i$, $i = 2, \ldots, p$.

Within each of these scenarios, we vary the sizes of the dominating eigenvalues while fixing the remaining eigenvalues at $\lambda_i = \frac{1}{2}$. As before, we generate i.i.d. $X_i \sim N(0, \Sigma_X)$ and $\epsilon_i \sim N(0, 1)$ with $p = 50$ and $n = 150$ and estimate the EDF for predictor envelopes with dimension $d = 1, 2, \ldots, 10$.

We pick these five cases and the settings we vary within them to highlight a few meaningful contrasts. Let V denote the set of eigenvectors of Σ_X with dominating eigenvalues and let W denote the set of eigenvectors of Σ_X on which β is linearly dependent. With this notation, we can express the conditions under which we see substantial EDF reductions succinctly: (1) $V \neq \emptyset$, (2) $d \leqslant |V|$, and (3) $V \cap W \neq \emptyset$. We can verify the importance of the first two conditions by focusing on how $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ varies with λ_1 and d within each of the first four cases, as λ_1 and d determine whether $V \neq \emptyset$ and $d \leqslant |V|$. We must compare results across cases to establish the importance of the third condition, as only Case 5 is designed so that $V \cap W = \emptyset$.

Figure 3 shows the simulation results for the five cases outlined above. We start by focusing on the first four cases. A clear trend emerges among the simulations for which $\lambda_1 = 100$: $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ is near $d + 1$ when $d \leqslant |V|$ and near $p + 1$ for $d > |V|$. There is a clear fault line between $d = |V|$ and $d = |V| + 1$ in all four of these cases. We see that $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ is constant for $d \leqslant |V|$ in both Case 2 and Case 3 and that $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ increases slowly in d past $d = |V| + 1$. We see similar patterns for $\lambda_1 = 5, 10$, and 25, though the jumps in the EDF between $d = |V|$ and $d = |V| + 1$ are less dramatic.

Intuitively, we might also expect that the number of eigenvectors of Σ_X which provide us with information about β would similarly play a major role in determining the EDF for a predictor envelope. In Cases 3 and 4, however, we do not see major changes in $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ between $d = |W|$ and $d = |W| + 1$. Thus $|V|$, rather than $|W|$, appears to be the salient threshold for the envelope dimension when it comes to EDF reduction.

Next, we focus on making comparisons across different λ_1 for fixed d. We see that the relative size of λ_1 determines how far $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ drops below $p + 1$ for $d \in \{1, \ldots, |V|\}$. That is, the difference between $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_{|V|})$ and $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_{|V|+1})$ grows as the gap between λ_1 and $\lambda_{|V|+1}$ increases. In fact, when $\lambda_1 = 1$ predictor envelopes appear to hardly deliver any EDF reductions at all. Thus it seems necessary that $V \neq \emptyset$ for predictor envelopes to achieve substantial EDF reductions.

Lastly, we focus on the role of the eigenvectors of Σ_X which provide us with information about β. We set $V \cap W = \emptyset$ in Case 5. Here we see that even when $d = |V| = 1$, $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ is only slightly below $p + 1$. Contrast this result with Case 1. In both cases $|V| = |W| = 1$, yet we see dramatically different behavior when $d = 1$. This comparison suggests that at least one of the eigenvectors in V needs to provide some information about β for envelopes to achieve substantial EDF reductions.

As in the previous section, we rerun one of our simulations to ensure our findings generalize beyond our initial choice of n and p. Figure 4 shows the EDFs for predictor envelopes when $n = 200$, $p = 100$, and X and Y are generated as in Case 1. The overall pattern remains the same.

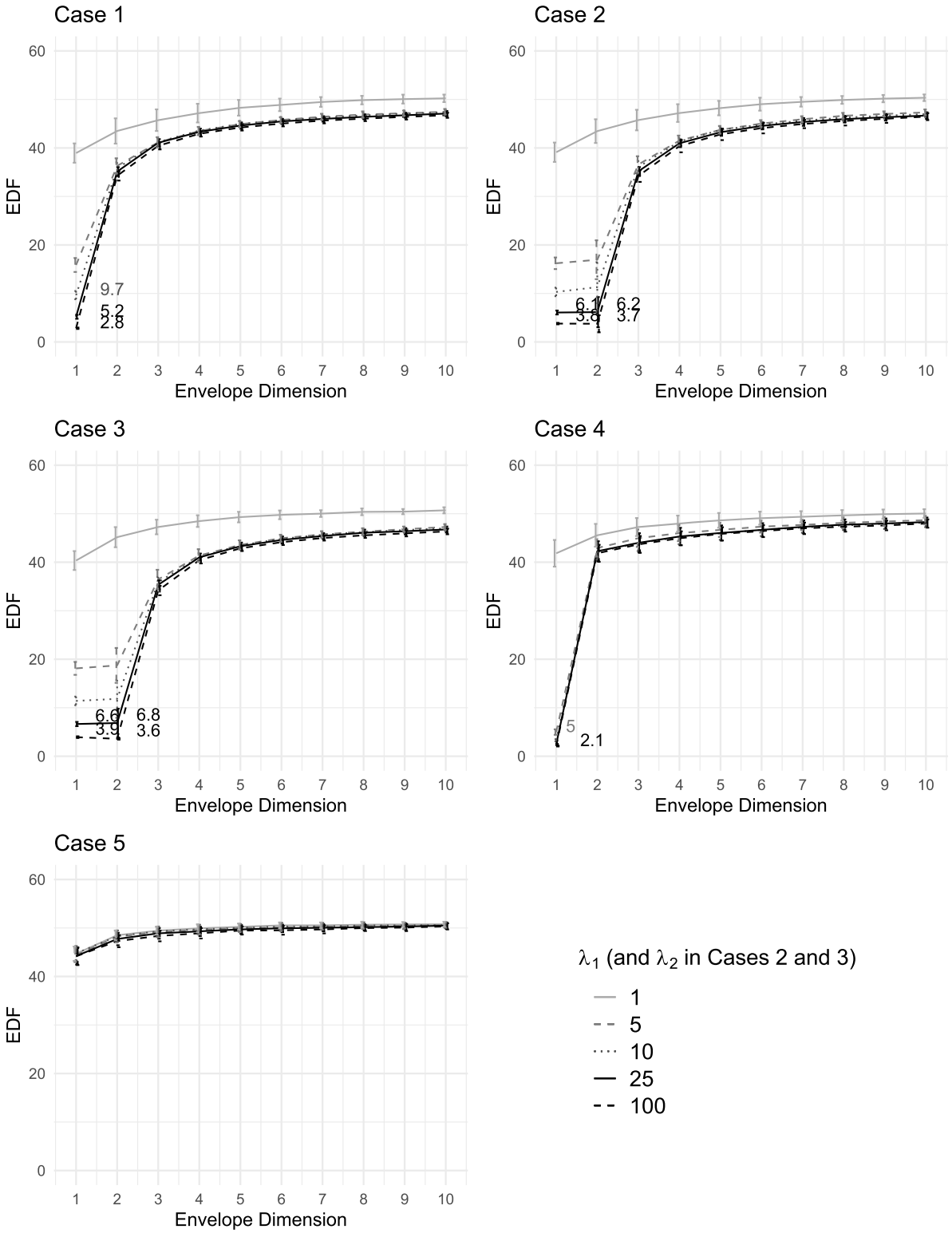

Figure 3: EDFs in settings for reducing envelope complexity.

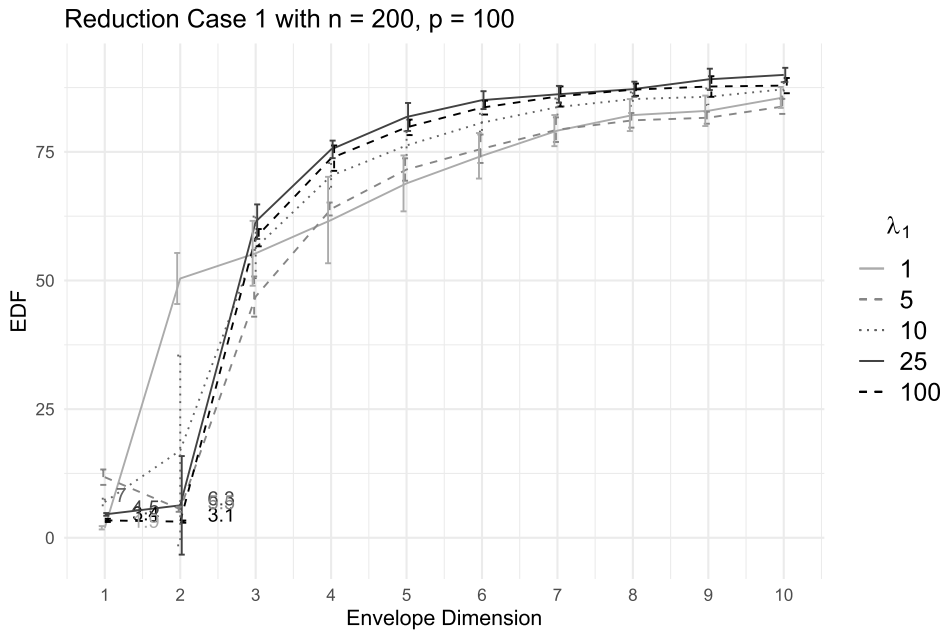

Figure 4: Envelope EDFs in Case 1 with $n = 200$, $p = 100$.

6 A Real Data Example

In this section, we turn to a real data example to study the EDF of predictor envelopes. Our data set characterizes the student populations of Minneapolis district schools in 1972 and contains $n = 63$ observations and $p = 9$ predictors (Cook, 1998).

In our first set of simulations, we only use the design matrix from this data set. We simulate our response variable from a linear model as described in Section 3, using different coefficient vectors β_k to generate Y in each of 100 simulation iterations. In these simulations, we generate $\beta_1, \ldots, \beta_5 \sim Gamma(2, 2)$ and set $\beta_6 = \cdots = \beta_9 = 0$. For our second set of simulations, we first fit an envelope model with dimension 1 to the complete data, then use the estimated regression coefficients $\hat{\beta}$ and response variance $\hat{\sigma}^2$ from that model as the "true" data-generating values for Monte Carlo simulations. In doing so, we treat the fitted envelope model as the "true" model for these simulations. In both sets of simulations, we fit predictor envelopes with envelope dimensions from $d = 1$ to 5.

Figure 5 shows the results from the first set of simulations. We see that $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d)$ is about 10 regardless of the envelope dimension. These results align with our findings from Section 4 for settings in which the significant predictors are not highly correlated. In the results for the second set of simulations, we see from Figure 6 that $\widehat{\mathrm{df}}(X, \beta, \mathrm{Env}_d) \approx 2$ for a predictor envelope with dimension 1, mirroring the results from the ideal setting simulations in Section 5. As in the ideal simulations, the EDF remains below $p + 1$ for models with higher envelope dimensions as well. These results reveal that when the "true" model for (X, Y) has an envelope structure (as it does in this case because we generated Y from such a model), estimating the envelope adds few degrees of freedom to the overall fitting procedure.

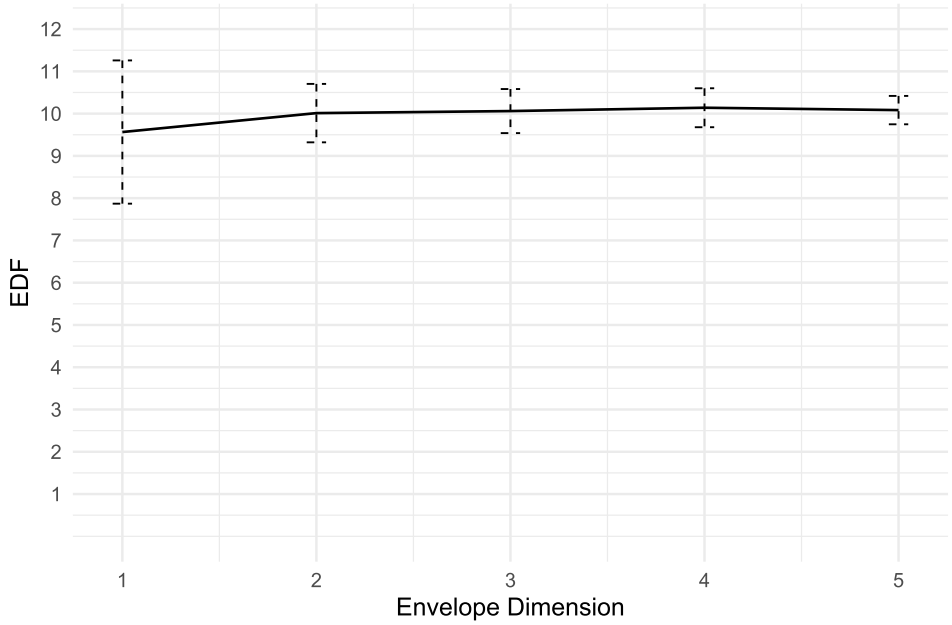

Figure 5: Envelope EDFs with predictors from Minneapolis school data.

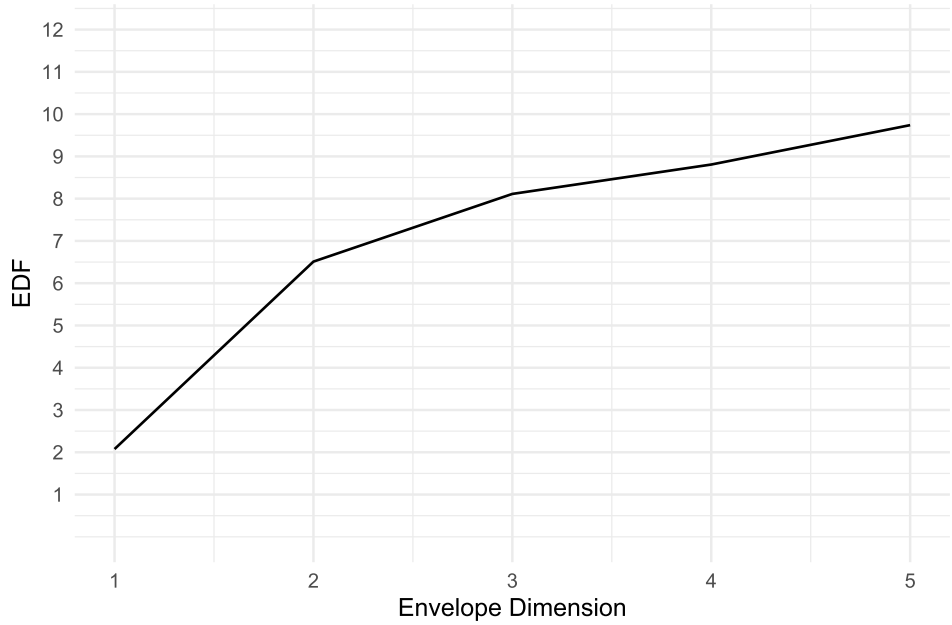

Figure 6: Envelope EDFs assuming an envelope model for Minneapolis school data.

7 Discussion

We have found that while the fundamental idea of predictor envelopes is to reduce the full space of the original p input variables down to a much-reduced subspace $\Gamma^T X \in \mathbb{R}^d$ for the purpose of modeling Y, the actual estimation process tends to have a dramatic impact on the complexity

of the fitted predictor envelope model. In many reasonable settings predictor envelopes tend to have far more than $d+1$ effective degrees of freedom. In fact, the EDF is often close to $p+1$. This suggests that under those settings estimating the subspace $\Gamma^T X$ adds nearly $p-d$ degrees of freedom to the overall fitting procedure. By comparing the EDF across several simulation settings, we have found that the structure of the joint distribution of (X, Y) plays a critically important role in the EDF.

The scenarios identified in Section 5 gave us results close to the usual perception that the complexity is $d+1$, suggesting that the process of estimating $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ adds few degrees of freedom in these settings. In particular, we found that the EDF is close to $d+1$ when the "true" model is a predictor envelope and $d \leqslant d^*$, the true envelope dimension. When we set an envelope model with dimension 1 as the "true" model for simulations in Section 6 we saw similar results: predictor envelopes with $d=1$ used only 2 degrees of freedom. In this regard, predictor envelopes follow a pattern seen in other dimension reduction techniques. Mukherjee et al. (2015) identified a similar phenomenon for reduced-rank estimators: their unbiased estimator for the EDF approaches the "naïve estimator," the number of free parameters in the model, when the model rank r is close to the true underlying rank r^*.

We know from Theorem 1 that the EDF measures how much the training RSS differs from the in-sample test RSS. As we have found, EDFs for predictor envelopes (and therefore the corrections needed for the training RSS) are often much larger than one might assume. We compute the EDF, RSS and in-sample test RSS for predictor envelopes in a few of our simulation studies. Figures 7 and 8 plot these values for the compound symmetric case with $\rho = 0.8$ (from Section 4) and the first reduction case with $\lambda_1 = 25$ (from Section 5), respectively. In both cases the RSS provides a gross underestimate of the in-sample test RSS. Moreover, in the latter case we see that different values of d minimize the RSS and the in-sample test RSS. The large EDF can be interpreted as the high complexity of the fitted envelope model, which implies high

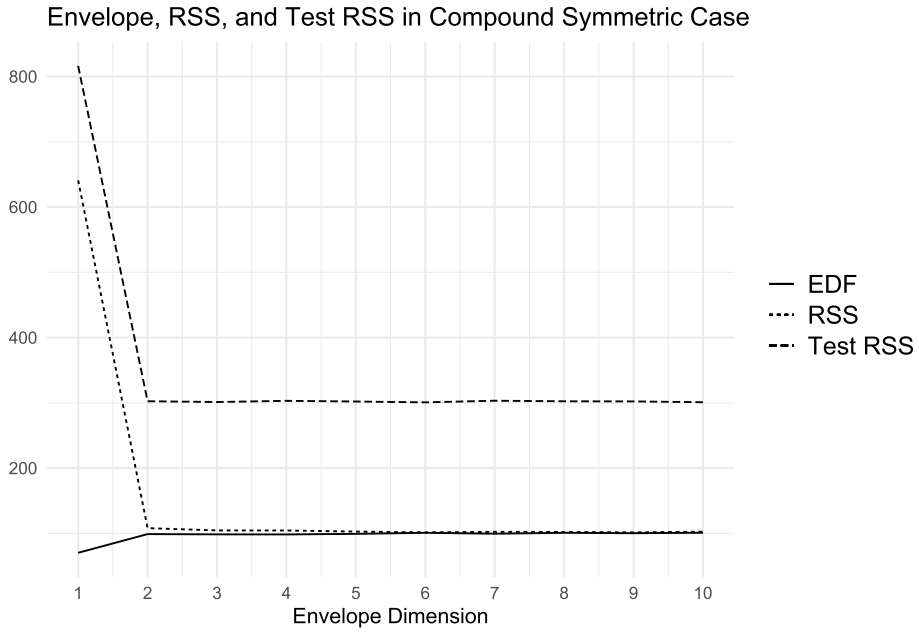

Figure 7: EDF, RSS, and test RSS in compound symmetric case with $n = 200$, $p = 100$.

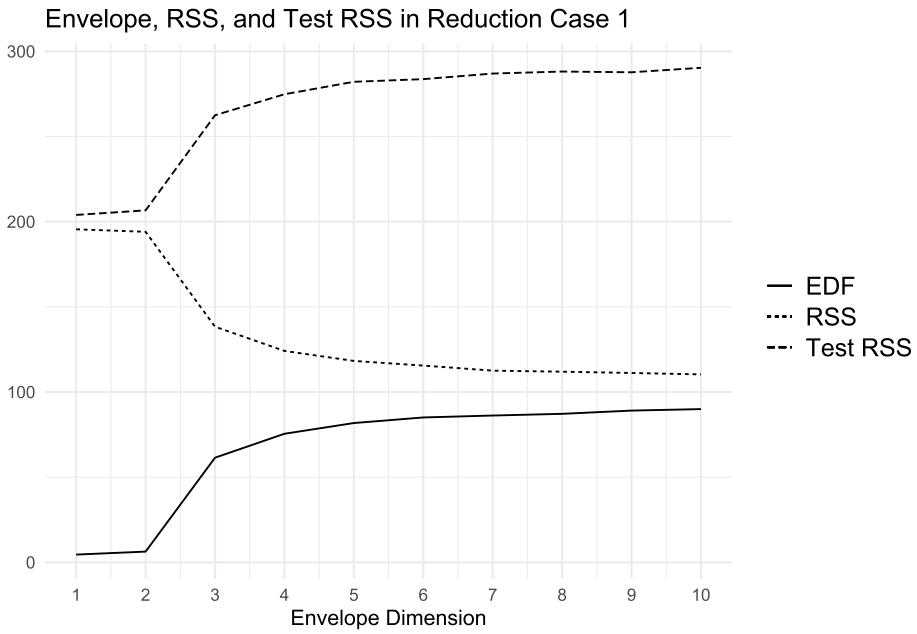

Figure 8: EDF, RSS, and test RSS in reduction Case 1 with $n = 200$, $p = 100$.

estimation variance contributing to the high test RSS. These results underscore that users need accurate estimates of the EDF to correctly compare and assess predictor envelopes.

Lastly, we point out that the EDF of partial least squares (PLS) was examined in Krämer and Sugiyama (2011) where it was observed that the EDF can be far greater than $d+1$ (here d denotes the number of components used in PLS) and often can be close to $p+1$. As mentioned earlier, predictor envelope models and PLS are related in the sense that they share the same population target, though the former uses a likelihood approach for estimation and has improved efficiency over the latter. Our findings that the correlation structure of the design matrix shapes the EDF of predictor envelopes align with the findings in Krämer and Sugiyama (2011).

Supplementary Material

Code and data for reproducing our results can be found at https://github.com/TateJacobson/Envelope-EDF. This repository contains the following folders:

- **Cleaning Output:** Contains an R script for cleaning saved simulation output and generating plots from it.
- **edf:** An R package for computing the effective degrees of freedom
- **Simulations:** Contains R scripts for the simulations run in "Do Predictor Envelopes Really Reduce Dimension?"

Acknowledgement

We thank the editor, AE and two referees for their helpful comments and suggestions.

Funding

This work is supported in part by NSF 1915842 and 2015120.

References

Cook RD (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons.

Cook RD (2018). *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. John Wiley & Sons.

Cook RD, Forzani L (2020). Envelopes: A new chapter in partial least squares regression. *Journal of Chemometrics*, 34(10), e3287, DOI: https://doi.org/10.1002/cem.3287.

Cook RD, Forzani L, Su Z (2016). A note on fast envelope estimation. *Journal of Multivariate Analysis*, 150: 42–54.

Cook RD, Helland IS, Su Z (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(5): 851–877.

Cook RD, Li B, Chiaromonte F (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3): 569–584.

Efron B (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394): 461–470.

Janson L, Fithian W, Hastie TJ (2015). Effective degrees of freedom: A flawed metaphor. *Biometrika*, 102(2): 479–485.

Krämer N, Sugiyama M (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, 106(494): 697–705.

Lee M, Su Z (2020). R package *Renvlp: Computing Envelope Estimators*. https://cran.r-project.org/web/packages/Renvlp/.

Mallows CL (1973). Some comments on C_p. *Technometrics*, 15(4): 661–675.

Mukherjee A, Chen K, Wang N, Zhu J (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2): 457–477.