One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models

David J. Olive and Lingling Zhang

School of Mathematical & Statistical Sciences
Department of Mathematics

Southern Illinois University
University at Albany

Carbondale, Illinois 62901-4408
Albany, New York 12222

dolive@siu.edu
lzhang28@albany.edu

Keywords Cox proportional hazards regression; GLM; Lasso; MMLE; OLS; PLS.

Mathematics Subject Classification Primary 62J05; Secondary 62J12.

Abstract

This paper gives large sample theory for the one component partial least squares estimator, including some hypothesis tests for high dimensional data, under much weaker conditions than those in the literature. Simple theory is also given for some data splitting estimators and the marginal maximum likelihood estimators. It is shown that lasso, one component partial least squares, and ordinary least squares often estimate different population multiple linear regression models. The paper also proves that there are often many valid population models for regression methods such as binary regression.

1. Introduction

This section reviews regression models, including variable selection and data splitting. Many regression models have a response variable Y that is independent of the $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. For the regression models, the conditioning and subscripts, such as i, will often be suppressed. The multiple linear regression model is $Y|\boldsymbol{x}^T\boldsymbol{\beta} = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ or $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Consider a parametric regression model $Y|\boldsymbol{x}^T\boldsymbol{\beta} \sim D(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ where D is a parametric distribution that depends

on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters. Three examples follow. The *binomial logistic regression model* is $Y_i \sim \text{binomial}\left(\text{m}_\text{i}, \rho(\text{SP}) = \frac{\text{e}^{\text{SP}}}{1+\text{e}^{\text{SP}}}\right)$. The binary logistic regression model has $m_i \equiv 1$ for $i = 1, ..., n$. A useful *Poisson regression model* is $Y \sim \text{Poisson}\left(\text{e}^{\text{SP}}\right)$. If the Y_i follow a Weibull regression model, then the $\log(Y_i)$ follow an accelerated failure time model: $\log(Y_i) = \delta + \boldsymbol{\beta}_A^T\boldsymbol{x}_i + \sigma e_i$. Let $\lambda_0 = \exp(-\delta/\sigma)$ and $\boldsymbol{\beta} = -\boldsymbol{\beta}_A/\sigma$. Then for $SP = \boldsymbol{\beta}^T\boldsymbol{x}$, the *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP))$$

where Y has a Weibull $W(\gamma, \lambda)$ distribution if the probability density function of Y is

$$f(y) = \lambda\gamma y^{\gamma-1}\exp[-\lambda y^\gamma] \text{ for } \text{y} > 0.$$

Variable selection estimators include forward selection or backward elimination. Sparse regression methods can also be used for variable selection even if n/p is not large: the regression submodel, such as a Nelder and Wedderburn (1972) generalized linear model (GLM), uses the predictors that had nonzero sparse regression estimated coefficients. These methods include least angle regression, lasso, relaxed lasso, elastic net, and sparse regression by projection. See Efron et al. (2004, p. 421), Meinshausen (2007, p. 376), Qi et al. (2015), Tay, Narasimhan, and Hastie (2023), Tibshirani (1996), and Zou and Hastie (2005).

A *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E \tag{1}$$

where $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Here E denotes the subset of terms that can be eliminated, without much loss of information, given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then $\boldsymbol{\beta}_S$ corresponds to the optimal reduced model. If $I \neq S$ and $S \subseteq I$, then $\boldsymbol{\beta}_I$ corresponds to a nonoptimal reduced model: overfitting, while $S \not\subseteq I$ corresponds to underfitting.

To clarify notation, suppose $p = 3$, a constant α is always in the model, and $S = I_2 = \{1\}$. Then the $J = 2^p = 8$ possible subsets of $\{1, 2, ..., p\}$ are $I_1 = \emptyset$, S, $I_3 = \{2\}$, $I_4 = \{3\}$,

$I_5 = \{1, 2\}$, $I_6 = \{1, 3\}$, $I_7 = \{2, 3\}$, and $I_8 = \{1, 2, 3\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_2, \hat{\beta}_3)^T$ and $\boldsymbol{x}_{I_7} = (x_2, x_3)^T$.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p$.

Theory for the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$ is complicated. See Pelawa Watagoda and Olive (2021) for multiple linear regression, and Rathnayake and Olive (2023) for models such as GLMs and Cox (1972) proportional hazards regression. For fixed p, these two papers showed that $\hat{\boldsymbol{\beta}}_{VS}$ is $\sqrt{n}$ consistent with a complicated nonnormal limiting distribution.

Principal components regression (PCR) and partial least squares (PLS) models use p conditional distributions $Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}, ..., \boldsymbol{\eta}_k^T\boldsymbol{x})$ for $k = 1, ..., p$. Estimating the $\boldsymbol{\eta}_i$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Denote the one component PLS (OPLS) estimator by $\hat{\boldsymbol{\beta}}_{OPLS}$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPLS}$. See Cook (2018) and Wold (1975) for more on these and related estimators.

For estimation with OLS, let the covariance matrix of $\boldsymbol{x}$ be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})E(\boldsymbol{x}^T)$ and $\boldsymbol{\eta} = \text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = E[(\boldsymbol{x} - E(\boldsymbol{X})(Y - E(Y))] = E(\boldsymbol{x}Y) - E(\boldsymbol{x})E(Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))Y] = E[\boldsymbol{x}(Y - E(Y))]$. Let

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_n = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \boldsymbol{S}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}), \tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\eta}}_n = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}),$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)^T, \text{ and } \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{\text{n}-1}{\text{n}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}.$$

Then the OLS estimators are $\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \overline{\boldsymbol{x}}$ and

$$\hat{\boldsymbol{\beta}}_{OLS} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\eta}}.$$

For a multiple linear regression model with independent, identically distributed (iid) cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ under mild regularity conditions, while $\hat{\alpha}_{OLS}$ is a consistent estimator of $E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$.

Cook, Helland, and Su (2013) showed that $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \quad \text{and} \quad \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}} \tag{2}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$, then $\boldsymbol{\beta}_{OPLS} = \mathbf{0}$. Let $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. Large sample theory for OPLS is given in Section 2, and see Section 3.1 for earlier theory.

The marginal maximum likelihood estimator (MMLE or marginal least squares estimator) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression of Y on x_i resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$. For multiple linear regression, the marginal estimators are the simple linear regression (SLR) estimators, and $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M}) = (\hat{\alpha}_{i,SLR}, \hat{\beta}_{i,SLR})$. Hence

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

If the $\boldsymbol{w}_i$ are the predictors standardized to have unit sample variances, then

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = \boldsymbol{I}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{w}, Y)$$

where $(\boldsymbol{w}, Y)$ denotes that Y was regressed on $\boldsymbol{w}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix.

Data splitting divides the training data set of n cases into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. An application of data splitting is to use a variable selection method, such as forward selection or lasso, on H to get submodel I_{min} with a predictors, then fit the selected model to the cases in the validation set V using standard inference. See, for example, Rinaldo et al. (2019).

High dimensional regression has n/p small. A fitted or population regression model is sparse if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is nonsparse. A high dimensional population full regression model is abundant or dense if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model.

Section 2 gives the large sample theory for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ and OPLS. Section 3 proves that there are a multitude of regression models and gives more theory for regression estimators. Section 4 explains a sequential data splitting method that was used for Section 5. The simulation in Section 5 shows that lasso with k-fold cross validation often selects models that are not the population generating model, but which are useful for prediction.

2. Large sample theory and testing

The following theorem gives the large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\mathrm{Cov}}(\boldsymbol{x}, Y)$. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ and let $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ be defined below where

$$\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = E[(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T (Y_i - \mu_Y)^2)] - \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T.$$

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$.

Theorem 1 *Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k \, Y_i^m)$ exist for $j = 1, ..., p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. Then a)*

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \ \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{3}$$

$$\text{and} \ \ \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

b) Let $\boldsymbol{z}_i = \boldsymbol{x}_i(Y_i - \overline{Y}_n)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$.

c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T). \tag{4}$$

Proof. a) Note that $\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ by the multivariate central limit theorem since the $\boldsymbol{w}_i$ are iid with $E(\boldsymbol{w}_i) = \boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y)$ and $\mathrm{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{w}}$. Now $n\tilde{\boldsymbol{\eta}}_n = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(Y_i - \mu_Y + \mu_Y - \overline{Y}) = \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$

$$+ \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\mu_Y - \overline{Y}) + (\boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}}) \sum_i (Y_i - \mu_Y) + n(\boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y})$$

$$= \sum_i \boldsymbol{w}_i - n\boldsymbol{a}_n - n\boldsymbol{a}_n + n\boldsymbol{a}_n = \sum_i \boldsymbol{w}_i - n(\boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y}).$$

$$\text{Thus } \sqrt{\mathrm{n}}\tilde{\boldsymbol{\eta}}_{\mathrm{n}} = \sqrt{\mathrm{n}}\frac{1}{\mathrm{n}}\sum_{\mathrm{i}} \boldsymbol{w}_{\mathrm{i}} - \frac{\sqrt{\mathrm{n}}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})\sqrt{\mathrm{n}}(\overline{\mathrm{Y}} - \mu_{\mathrm{Y}})}{\sqrt{\mathrm{n}}} = \sqrt{\mathrm{n}}\ \overline{\boldsymbol{w}}_{\mathrm{n}} + \mathrm{o_P}(1).$$

$$\text{Hence } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) = \sqrt{\mathrm{n}}(\overline{\boldsymbol{w}}_{\mathrm{n}} - \boldsymbol{\eta}) + \mathrm{o_P}(1).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$$

by Slutsky's theorem. Now

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \frac{n}{n-1}\boldsymbol{\eta} + \frac{n}{n-1}\boldsymbol{\eta} - \boldsymbol{\eta}\right)$$

$$= \sqrt{n}\frac{n}{n-1}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \sqrt{n}\left(\frac{\boldsymbol{\eta}}{n-1}\right).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

b) Now $\sum_{\mathrm{i}} \boldsymbol{w}_{\mathrm{i}} = \sum_{\mathrm{i}} (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}} + \overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(\mathrm{Y_i} - \overline{\mathrm{Y}} + \overline{\mathrm{Y}} - \mu_{\mathrm{Y}}) = \sum_{\mathrm{i}} (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})(\mathrm{Y_i} - \overline{\mathrm{Y}}) +$

$$\sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\overline{Y} - \mu_Y) + (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}}) \sum_i (Y_i - \overline{Y}) + n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(\overline{Y} - \mu_Y) =$$

$$\sum_i \boldsymbol{z}_i + n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(\overline{Y} - \mu_Y) = \sum_i \boldsymbol{z}_i + n\boldsymbol{a}_n = \sum_i (\boldsymbol{z}_i + \boldsymbol{a}_n).$$

$$\text{Hence } \sum_{\mathrm{i}} (\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})(\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})^{\mathrm{T}} = \sum_{\mathrm{i}} [(\boldsymbol{z}_{\mathrm{i}} + \boldsymbol{a}_{\mathrm{n}} - (\overline{\boldsymbol{z}}_{\mathrm{n}} + \boldsymbol{a}_{\mathrm{n}}))(\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}}_{\mathrm{n}})^{\mathrm{T}}] =$$

$$\sum_i (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_i - \overline{\boldsymbol{z}})^T.$$

$$\text{Thus } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i=1}}^{\mathrm{n}} (\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}})(\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}})^{\mathrm{T}} \text{ and } \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{\mathrm{n}-1}{\mathrm{n}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}.$$

c) If H_0 is true, then $\boldsymbol{A\eta} = \mathbf{0}$, and

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) = \sqrt{n}\boldsymbol{A}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \hat{\lambda}\boldsymbol{\eta} + \hat{\lambda}\boldsymbol{\eta} - \boldsymbol{\beta}_{OPLS}) =$$

$$\hat{\lambda}\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \boldsymbol{A}\sqrt{n}(\hat{\lambda} - \lambda)\boldsymbol{\eta} = \boldsymbol{Z}_n + \boldsymbol{b}_n \xrightarrow{D} N_k(\mathbf{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T)$$

since $\boldsymbol{b}_n = \mathbf{0}$ when H_0 is true. $\square$

In Theorems 1 and 2, the scalars λ and $\hat{\lambda}$ are given by Equation (2), $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$, and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{\Sigma}_{\boldsymbol{w}}$. Results from Su and Cook (2012), for example, show that elements of a sample covariance matrix can be stacked to get large sample theory. Then $\hat{\lambda}$ and $\hat{\boldsymbol{\eta}}$ can be stacked as in Theorem 2 by the multivariate delta method. Theorem 1 c) and Theorem 2 c) are equivalent with different notation.

Theorem 2 *Assume*

$$\sqrt{n}\left(\begin{pmatrix}\hat{\lambda}\\ \hat{\boldsymbol{\eta}}\end{pmatrix} - \begin{pmatrix}\lambda\\ \boldsymbol{\eta}\end{pmatrix}\right) \xrightarrow{D} N_{p+1}\left(\begin{pmatrix}0\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\Sigma_\lambda & \boldsymbol{\Sigma}_{\lambda\boldsymbol{\eta}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}\lambda} & \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\end{pmatrix}\right) \sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma}).$$

a) $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$.

b) $\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\right)$ *with* $\boldsymbol{D} = [\boldsymbol{\eta} \;\; \lambda\boldsymbol{I}_p]$ *where* $\boldsymbol{I}_p$ *is the* $p \times p$ *identity matrix.*

c) Let $\boldsymbol{A}$ *be a* $k \times p$ *full rank constant matrix with* $k \leq p$ *and* $\boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0} = \boldsymbol{A}\boldsymbol{\eta}$. *Then*

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OPLS} - \mathbf{0}) \xrightarrow{D} N_k\left(\mathbf{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T\right).$$

Proof. a) Follows by Equation (3) or since joint convergence in distribution implies marginal convergence in distribution.

b) Follows by the Multivariate Delta Method with

$$\boldsymbol{g}\begin{pmatrix}\lambda\\ \boldsymbol{\eta}\end{pmatrix} = \lambda\boldsymbol{\eta} =$$

$(\lambda\eta_1, ..., \lambda\eta_p)^T$, and the Jacobian matrix of partial derivatives $\boldsymbol{D} = \boldsymbol{D}_{\boldsymbol{g}}$.

c) By b), $\sqrt{\mathrm{n}}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{\mathrm{OPLS}} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N_k}\left(\mathbf{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)$,

but $\boldsymbol{AD} = [\boldsymbol{0} \;\; \lambda \boldsymbol{A}]$. Hence $\boldsymbol{AD\Sigma D}^T\boldsymbol{A}^T = \lambda^2 \boldsymbol{A\Sigma_\eta A}^T$. $\square$

REMARK 1: Notice that Theorems 1 and 2 depend on the theory of both the sample covariance vector and the sample covariance matrix, not on any other model such as linearity. It is possible that $Y|\boldsymbol{x}$ does not follow a linear model, but $Y|\boldsymbol{\beta}_E^T\boldsymbol{x}$ does follow a linear model. If the population generating model $Y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ is a linear model, then $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ is a linear model. Suppose the cases are iid, and the predictors have nonsingular covariance matrix $\boldsymbol{\Sigma_x}$. Suppose a linear model holds with $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$. If the iid errors e are independent of the predictors $\boldsymbol{x}$, then under mild conditions, linearity implies that $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$ and that the covariance structure is $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma_x\beta}_{OLS}$. Suppose $(\hat{\alpha}_E, \hat{\boldsymbol{\beta}}_E)$ estimates $(\alpha_E, \boldsymbol{\beta}_E)$. If $Y|\boldsymbol{x} = \alpha_E + \boldsymbol{\beta}_E^T\boldsymbol{x} + e$, then by the above discussion, $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_E$.

Some additional useful OPLS and OLS formulas are derived next if the cases are iid. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathrm{Cov}(\boldsymbol{x}, Y) = \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta} = \boldsymbol{\Sigma_x\beta}$. Since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma_x\beta}_{OLS}$,

$$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma_x\beta}_{OLS}, \;\; \boldsymbol{\beta}_{OPLS} = \lambda\mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS}, \text{ and } \boldsymbol{\beta}_{\mathrm{OLS}} = \frac{1}{\lambda}[\mathrm{Cov}(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{\mathrm{OPLS}}.$$

2.1 High dimensional tests

The following simple testing method reduces a possibly high dimensional problem to a low dimensional problem. Testing $H_0 : \boldsymbol{A\beta}_{OPLS} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A\beta}_{OPLS} \neq \boldsymbol{0}$ is equivalent to testing $H_0 : \boldsymbol{A\eta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A\eta} \neq \boldsymbol{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix. Let $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}) = \mathrm{Cov}(\hat{\boldsymbol{\eta}}) = \boldsymbol{\Sigma_w}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y})$, but we can get good nonsingular estimators of $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}) = \mathrm{Cov}((\hat{\eta}_{i1}, ..., \hat{\eta}_{ik})^T)$ with $\boldsymbol{u} = (x_{i1}, ..., x_{ik})^T$ where $n \geq Jk$ with $J \geq 10$. (Values of J much larger than 10 may be needed if some of the k predictors and/or Y are skewed.) Simply apply Theorem 1 to the predictors $\boldsymbol{u}$ used in the hypothesis test, and thus use the sample covariance matrix of the vectors $\boldsymbol{u}_i(Y_i - \overline{Y})$. Hence we can test hypotheses like $H_0 : \beta_i - \beta_j = 0$. In particular, testing $H_0 : \beta_i = 0$ is equivalent to testing $H_0 : \eta_i = \sigma_{x_i,Y} = 0$ where $\sigma_{x_i,Y} = \mathrm{Cov}(x_i, Y)$.

Note that the tests with $\hat{\boldsymbol{\eta}}$ using k predictors x_{ij} do not depend on other predictors, including important predictors that were left out of the model (underfitting). Hence the tests can have considerable resistance to underfitting and overfitting.

3. The multitude of models

This section shows that there are often a multitude of population regression models that are estimating different population parameters. Note that when j predictors each satisfy a marginal regression model with the response Y (such as simple linear regression), then subsets of those j predictors will often satisfy a regression model with the response Y (such as multiple linear regression). Under multivariate normality, it is known that $Y|\boldsymbol{x}_I$ follows a multiple linear regression model where $\boldsymbol{x}_I = (x_{i1}, ..., x_{ik})^T$ is a vector corresponding to a subset of the predictors. Theorem 3a) gives a similar result for every linear combination of the predictors $\boldsymbol{\eta}^T\boldsymbol{x}$, including sparse and nonsparse models. Let $\Sigma_Y = \sigma_Y^2$.

Theorem 3 *Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution.*
a) If the joint distribution of $(Y, \boldsymbol{x}^T)^T$ is multivariate normal,

$$\begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+1}\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \end{pmatrix} \right),$$

then $Y|\boldsymbol{x} \sim Y|(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}) \sim N(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}, \sigma^2)$ follows a multiple linear regression model, but so does $Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim N(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}, \sigma_O^2)$ where $\alpha_O = \mu_Y - \boldsymbol{\beta}_O^T\boldsymbol{\mu_x}$, $\boldsymbol{\beta}_O = \lambda\boldsymbol{\eta}$, $\sigma_O^2 = \Sigma_Y - \boldsymbol{\beta}_O^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{\Sigma_x}\boldsymbol{\eta}}.$$

b) If the response Y is binary, then $Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}) \sim$ binomial$(m = 1, \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}))$ where $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}) = P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})]$. Hence every linear combination of the predictors satisfies a binary regression model.

Proof. a)

$$\begin{pmatrix} 1 & \boldsymbol{0}^T \\ 0 & \boldsymbol{\eta}^T \end{pmatrix} \begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} Y \\ \boldsymbol{\eta}^T\boldsymbol{x} \end{pmatrix}$$

$$\sim N_2\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\eta}^T\boldsymbol{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta} \\ \boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\eta}^T\boldsymbol{\Sigma_x}\boldsymbol{\eta} \end{pmatrix} \right).$$

Hence $W = Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim N(\mu_W, \sigma_W^2)$ where

$$\mu_W = \mu_Y + \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{\Sigma_x}\boldsymbol{\eta}}(\boldsymbol{\eta}^T\boldsymbol{x} - \boldsymbol{\eta}^T\boldsymbol{\mu_x}) = \mu_Y - \lambda\boldsymbol{\eta}^T\boldsymbol{\mu_x} + \lambda\boldsymbol{\eta}^T\boldsymbol{x},$$

and

$$\sigma_W^2 = \sigma_O^2 = \sigma_Y^2 - \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta}\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} = \sigma_Y^2 - \frac{(\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta})^2}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} = \sigma_Y^2 - \lambda \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

b) $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] = 0P[Y = 0|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] + 1P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})]$

$= P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})$. $\square$

For multiple linear regression, note that $\sigma_O^2 < \sigma_Y^2 = \Sigma_Y$ unless $\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = 0$. If $\boldsymbol{\eta} = \boldsymbol{\beta}_{OLS}$, then $\lambda = 1$ and $\sigma_O^2 = \sigma_Y^2 - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The population quantity estimated by the one component partial least squares estimator corresponds to $\boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

3.1 Consequences

Although Theorems 1–3 have simple proofs, the theorems have important consequences. One consequence is the testing theory in Section 2.1.

Data splitting: To help understand data splitting when the cases in H are randomly selected, let I denote the predictors selected using H, possibly after variable selection or after looking at the data and building the model. Let $\hat{\boldsymbol{\beta}}_E(\boldsymbol{x}_I, Y)$ be the estimator obtained by regressing Y on $\boldsymbol{x}_I$ using the cases in V. Then $\hat{\boldsymbol{\beta}}_E(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$. For example, if the cases are iid with enough low order moments, then $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \boldsymbol{\Sigma}_{\boldsymbol{x}_I}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}_I,Y}$ while $\hat{\boldsymbol{\beta}}_{OPLS}(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I,Y}$. If the model is sparse, check the fitted model with the same checks used for low dimensional data. For data splitting in low dimensions, if the full model is good, then often model (1) works well in that we can eliminate predictors and often do nearly as well or better than the full model. In high dimensions, we often do not know if the full model, that regresses Y on $\boldsymbol{x}$, is good. The data splitting and high dimensional regression literature often claims that $\boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y) = \boldsymbol{\beta}_E(\boldsymbol{x}, Y)$. For example, $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$, or model (1) holds with $S \subseteq I_{min}$ and $\boldsymbol{\beta}_{I_{min}}$ a $k \times 1$ vector with $a_S \leq k \leq n/10$. While these claims can be true, the regularity conditions often become too strong as $n/p \to 0$.

MMLE and the oracle property: The MMLE is interesting since if each predictor satisfies a marginal model, then the marginal model theory can be used to find a confidence interval for β_i for $i = 1, ..., p$ where β_i is the ith component of $\boldsymbol{\beta}_{MMLE}$. For multiple linear regression, let $\boldsymbol{V} = diag(\boldsymbol{\Sigma}_{\boldsymbol{x}}) = diag(\sigma_1^2, ..., \sigma_p^2)$. For iid cases, $\boldsymbol{\beta}_{MMLE} = \boldsymbol{V}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x},Y} =$

$\boldsymbol{V}^{-1}\boldsymbol{\Sigma_x}\boldsymbol{\beta}_{OLS}$, and $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS}$ if $\boldsymbol{\beta}_{OLS} = \mathbf{0}$, or if $(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1})\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathbf{0}$, or if $\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{V}^{-1}\boldsymbol{\Sigma_x}$ with eigenvalue 1.

For standardized predictors, let s_j and σ_j be the sample and population standard deviations of x_j. Let $\boldsymbol{w}_i = \hat{\boldsymbol{D}}\boldsymbol{x}_i = diag(1/s_1, ..., 1/s_p)\boldsymbol{x}_i$ and $\boldsymbol{u}_i = \boldsymbol{D}\boldsymbol{x}_i = diag(1/\sigma_1, ..., 1/\sigma_p)\boldsymbol{x}_i$. Note that $\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) = \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y}) + \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) = O_P(1) + \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y})$ under mild regularity conditions for iid cases. Hence $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{u},Y}$ that is not asymptotically equivalent to $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y}$ unless $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathbf{0}$. The algebra given in the following theorem proves the theorem. Note that $\boldsymbol{\Sigma_u}$ is the correlation matrix of $\boldsymbol{x}$.

Theorem 4 *Consider the MMLE for multiple linear regression. Suppose the cases* $(Y_i, \boldsymbol{x}_i^T)^T$ *are iid from some distribution. Let* $\boldsymbol{w}_i$ *be the standardized predictors and assume* $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$ *and* $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \xrightarrow{P} \boldsymbol{\Sigma_u}$ *where the* $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$ *are nonsingular for large enough* n *and* $\boldsymbol{\Sigma_u}$ *is nonsingular.*

$$a)\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{w}, Y) \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u},Y} =$$

$$\boldsymbol{\eta}_{OPLS}(\boldsymbol{u}, Y) = \boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma_u}[\boldsymbol{\Sigma_u}]^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = \boldsymbol{\Sigma_u}\boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y).$$

b) Let $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y)$. *Then* $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma_u}\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}$ *if* $\boldsymbol{\beta}_{OLS} = \mathbf{0}$ *or if* $\boldsymbol{\beta}_{OLS}$ *is an eigenvector of* $\boldsymbol{\Sigma_u}$ *with eigenvalue = 1.*

The oracle property for model selection, including variable selection, is $P(I_{min} = S) \to 1$ as $n \to \infty$ for model (1). For this property to hold, S needs to be one of the subsets considered by the model selection method with probability going to 1 as $n \to \infty$. For fixed p and "fast" estimators such as lasso and forward selection, the oracle property tends to hold if the predictors are nearly orthogonal. See Wieczorek and Lei (2022) for references. The MMLE can be used for variable selection with OLS by taking the k predictors with the largest $|\hat{\beta}_{j,MMLE}|$. The oracle property for the MMLE tends not to hold for correlated predictors by Theorem 4. MMLE variable selection often gives a useful submodel since predictors that satisfy a marginal regression model with the response Y (such as SLR) will often satisfy a regression model with the response Y (such as multiple linear regression).

OPLS and OLS: Chun and Keleş (2010) suggested that $\hat{\boldsymbol{\beta}}_{OPLS}$ only estimates $\boldsymbol{\beta}_{OLS}$ under very strong regularity conditions. Cook and Forzani (2018, 2019) showed that the regularity condition is $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, in which case $\sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{C})$.

Table 1: OPLS Results Under Theorem 1 Assumptions

General	$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\beta}_{OPLS}$
$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\lambda}[Cov(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}$	$\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$	$\boldsymbol{\beta}_{OPLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$	$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{kPLS}$	$\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{OLS}$

In much of the OPLS literature, an assumption is $Y|\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$. Then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ by the Remark 1 in Section 2, and the results in Table 1 hold. To see some problems with the assumption, consider multiple linear regression with $\text{Cov}(\boldsymbol{x}) = diag(1, 2, ..., p)$. First consider OPLS with $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$. Then at most one element of $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is nonzero since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\text{Cov}(\boldsymbol{x})$. Hence at most one predictor is correlated with Y, regardless of the value of p. This restriction is too strong.

If the cases are iid from a multivariate normal distribution, then $Y|\boldsymbol{x} = \alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} + e$ and $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ are both linear models by Theorem 3 where e depends on the model. Since $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ forces $\boldsymbol{\beta}_{OLS}$ to be an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, *if* $\boldsymbol{\beta}_{OLS} \neq \mathbf{0}$ *is not an eigenvector of* $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, *then* $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. For a computational example, let $\boldsymbol{x} \sim N_p(\mathbf{0}, diag(1, 2, 3, 4))$ with $\boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(1, 2, 3, 4)$, and let the population generating model be $Y_i = x_{i1} + x_{i2} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, 1)$ and independent of the $\boldsymbol{x}_i$. Then $\alpha = 0$ and $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. Hence $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta} = (1, 1, 0, 0)^T$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS} = (1, 2, 0, 0)^T$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}} = 5/9.$$

Thus $\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS} = (5/9, 10/9, 0, 0)^T \neq \boldsymbol{\beta}_{OLS}$. Thus OLS and OPLS usually give different valid population multiple linear regression models with $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$.

However, model $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ is often a useful multiple linear regression model with large sample theory given in Section 2. The claims in the OPLS literature that $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS} =$ an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ under mild regularity conditions are incorrect. See, for example, Basa et al. (2022), Cook and Forzani (2018, 2019), and Cook, Helland and Su (2013). In the OLS literature, $\boldsymbol{\beta}_{OLS}$ can be any vector in $\mathbb{R}^p$. If $\boldsymbol{\beta}_{OLS}$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, and $\boldsymbol{\beta}_{OPLS}$ were restricted to be eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the OLS and OPLS estimators would often not fit the data well.

The Bet on Sparsity Principle: Hastie, Tibshirani, and Wainwright (2015, p. 2) state that the "bet on sparsity principle" is *use a procedure that does well in sparse problems, since no procedure does well in dense problems.* Here the dense (or abundant) problem refers to the population generating model. Estimating the optimal population generating model or the model $Y|\boldsymbol{x}$ may be too difficult for a given dense problem, but many suboptimal models, including sparse fitted models, may be useful. For regression models with iid cases, the $Y_1, ..., Y_n$ are iid, and the useful suboptimal *null model* omits all of the predictors. For high dimensional data, a reasonable goal is to find a regression model that greatly outperforms the null model.

Next, consider sparse high dimensional estimators with $\boldsymbol{\beta}_E = \boldsymbol{\beta}_{OLS}$, such as E=lasso. Suppose model (1) holds with iid cases, $\text{Cov}(\boldsymbol{x}) = diag(1, 2, ..., p)$, and $n \geq 10a_S$. Hence $p - a_S$ of the elements of $\boldsymbol{\beta}_{OLS}$ are equal to zero. Then $\text{Cov}(\boldsymbol{x}, Y) = \text{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$, and at least 90% of the predictors are uncorrelated with Y. If $p > 100n$, then for lasso, at least 99% of the predictors are uncorrelated with Y since lasso uses at most $a = n$ predictors. Hence for sparse models, often $\boldsymbol{\beta}_E \neq \boldsymbol{\beta}_{OLS}$ for high dimensional data. However, if data splitting with lasso variable selection is used to find model I, the model $Y|\boldsymbol{\beta}_I^T\boldsymbol{x}_I$ will often be useful. Rathnayake and Olive (2023) proved that for fixed p and model (1), lasso and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}_{OLS}$ if lasso and elastic net are consistent estimators of $\boldsymbol{\beta}_{OLS}$.

Theorem 3 showed that sparse fitted models can do well in dense problems. The multitude of models result also helps explain why sparse fitted models can be useful even when the

population generating model is not sparse. Sparse variable selection models are interesting, since data splitting can be used for testing and confidence regions, and the submodel can often be checked with response and residual plots. See Olive (2013). The sparse fitted lasso model I can be more useful than the sparse lasso variable selection model if that model is ill conditioned. For example for multiple linear regression, if $(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$ is ill conditioned.

The following two sections help illustrate that k-fold cross validation with lasso often selects a model useful for prediction. Also see Chetverikov, Liao, and Chernozhukov (2022).

4. Sequential data splitting

The sequential data splitting algorithm is simple. Let $\lfloor x \rfloor$ be the integer part of x, e.g. $\lfloor 7.7 \rfloor = 7$. Denote the ceiling function by $\lceil x \rceil$, e.g. $\lceil 7.7 \rceil = 8$. Initially, randomly divide the data set into two sets: H_1 with $n_1 \leq n/2$ cases and V_1 with $n - n_1$ cases. Apply lasso on H_1 to get a set of a_1 predictors, including a constant if a constant is in the model. If $n_1 \geq 10a_1$, set $H = H_1$ and $V = V_1$. Otherwise, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Apply lasso on H_2 to get a set of a_2 predictors. If $n_2 \geq 10a_2$, set $H = H_2$ and $V = V_2$. Continue in this manner, forming sets $(H_1, V_1), (H_2, V_2), ..., (H_d, V_d)$ where H_i has $n_i = in_1$. Stop when $n_d \geq 10a_d$ or $n_{d+1} > \lfloor (n - J)/2 \rfloor$ where $J = 5$ was often used in the simulations. For the second case, use $n_d = \lfloor (n - J)/2 \rfloor$. Then $H = H_d$ and $V = V_d$. Use the model I_d with a_d predictors as the full model for inference with the data in $V = V_d$.

Lasso uses up to n_d active predictors and a constant. If J is an integer between 0 and 5, set $n_1 = \max(1, \lfloor (n - J)/2 \rfloor)$ if $n < 40$. Otherwise, we often used $n_1 = 30$, but changed n_1 to $\lfloor n/2000 \rfloor$ if initially $\lfloor n/(2n_1) \rfloor > 1000$. If $n >> p$, let $n_1 = Kp$ with K a positive integer, such as $K = 10$ or $K = 20$, or use $n_1 \approx Kp \approx n/(2M)$ with $M = \lceil n/(2Kp) \rceil$. If n/p is not large, options include $M = 10$ or $n_1 = Ka_0$ where a_0 is, for example, a guess of a lower bound for the number of active predictors.

5. Example and simulation

EXAMPLE. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. Let $Y =$ the *number of women married to civilians* in the district with a constant

and predictors x_1 = the *population of the district in 1843*, x_2 = the *number of married civilian men* in the district, x_3 = the *number of married men in the military* in the district, and x_4 = the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y and x_2 are highly correlated but not equal. Similarly, x_3 and x_4 are highly correlated but not equal. Then $\hat{\boldsymbol{\beta}}_{OLS} = (0.00035, 0.9995, -0.2328, 0.1531)^T$, forward selection with OLS and the C_p criterion used $\hat{\boldsymbol{\beta}}_{I,0} = (0, 1.0010, 0, 0)^T$, lasso had $\hat{\boldsymbol{\beta}}_L = (0.0015, 0.9605, 0, 0)^T$, lasso variable selection $\hat{\boldsymbol{\beta}}_{LVS} = (0.00007, 1.006, 0, 0)^T$, $\hat{\boldsymbol{\beta}}_{MMLE} = (0.1782, 1.0010, 48.5630, 51.5513)^T$, and $\hat{\boldsymbol{\beta}}_{OPLS} = (0.1727, 0.0311, 0.00018, 0.00018)^T$. The estimators had $\hat{\boldsymbol{\beta}}_3 \approx \hat{\boldsymbol{\beta}}_4$, and all six estimators produced fitted values $\hat{Y}_i$ that are very highly correlated with the response Y_i. For OPLS, the largest $|\hat{\beta}_i|$ corresponds to the largest $|\widehat{Cov}(x_i, Y)|$, and $\hat{\beta}_i/\hat{\beta}_j = \widehat{Cov}(x_i, Y)/\widehat{Cov}(x_j, Y)$ does not depend on any other variables that may be in or out of the model. Similar properties hold for the OPLS population β_i. This example illustrates that the OLS, OPLS, and MMLE estimators $\hat{\boldsymbol{\beta}}_E$ are quite different, as expected from Theorems 1 and 4.

Next we did a small simulation study to illustrate that *the model I selected by lasso was often good for prediction* even when underfitting was common ($S \not\subseteq I$), since the prediction intervals still had good coverage with short length. *The simulation also illustrates the multitude of models.* Underfitting occurs when a predictor that generated the full model was not selected. The sequential data splitting of Section 4 was used with $n_1 = 30$. The programs give the mean n_d (*mnnd*): the number of cases used in H_d, and the mean a_d (*mnad*): the number of nonzero lasso coefficients $\hat{\beta}_i$, including the constant if the model contains a constant, for lasso applied to the n_d cases in H_d. The program computed the Olive, Rathnayake, and Haile (2022) large sample 95% prediction intervals (PIs) for lasso applied to all n cases (*lsapi*), lasso variable selection applied to all n cases (*LVSpi*), lasso applied to V_d (*lsplitpi*), and the model selected using H_d applied to V_d (*splitpi*). The second and fourth models used OLS, a GLM, or Weibull regression applied to the n cases or the cases in V_d. Two lines per run are shown in each table. The first line gives the average coverage (*cov*) of the prediction

intervals while the second line gives the average length (*len*). The value of *undfit* gives the proportion of times that the lasso model I underfit: $S \not\subseteq I$. Since 5000 runs were used, if that proportion is 0.05, in 250 of the 5000 runs, lasso underfit, while then the proportion of times that lasso did not underfit is 0.95. More simulations are in Zhang (2022).

The prediction intervals were computed roughly as follows. If $Y \sim D(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\theta})$, then apply a prediction interval to a bootstrap sample of size B: $Y_1^*, ..., Y_B^*$ where the Y_i^* are iid $D(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$. For multiple linear regression, obtain the n_c residuals r_j and apply a prediction interval to $\hat{Y}_f + r_1, ..., \hat{Y}_f + r_{n_c}$ where $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$ and $n_c = n$ or $n = n_V$ depending on whether all n cases or data splitting was used for the prediction interval.

Table 2: Poisson regression data splitting: underfitting and PI coverage and length

n	p/k	psi= ψ	mnnd/mnad	cov/len	lsapi	LVSpi	lsplitpi	splitpi	undfit
100	20	0.6000	38.6768	cov	0.9986	0.9909	0.9898	0.9783	0.2332
	1		3.3482	len	8.1733	8.2308	8.7990	8.2294	
100	100	0.3000	44.5573	cov	0.9799	0.9819	0.9870	0.9580	0.1670
	1		7.1632	len	8.0935	7.1896	7.8179	7.6183	
100	20	0.0000	43.6966	cov	0.9841	0.9733	0.9685	0.9574	0.5224
	19		10.6884	len	8.1987	7.7602	8.7657	7.9307	
1000	20	0.5000	56.4071	cov	0.9834	0.9915	0.9849	1.0000	0.3164
	1		3.9243	len	7.9746	8.1575	7.1918	7.2256	
1000	1000	0.1000	110.0411	cov	0.9880	0.9884	0.9902	0.9881	0.3520
	1		8.8354	len	7.3550	8.5021	8.2190	8.1900	
1000	10	0.3160	74.3965	cov	0.9920	0.9920	0.9899	0.9836	0.9952
	9		7.0638	len	7.5158	7.9454	8.1729	8.1484	

The full model was simulated as in Pelawa Watagoda and Olive (2021) and Olive, Rathnayake, and Haile (2022). This section and the programs use a change in notation: if $\boldsymbol{\beta}_c = (\alpha\ \boldsymbol{\beta}^T)^T$ and $\boldsymbol{w} = (1\ \boldsymbol{x}^T)^T$ in Section 1 of this paper, then the program notation is $\boldsymbol{\beta} = \boldsymbol{\beta}_c$ and $\boldsymbol{x} = \boldsymbol{w}$ are $p \times 1$ vectors, $\beta_1 = \alpha$, and $\boldsymbol{u} = \boldsymbol{x}$ is a $(p-1) \times 1$ vector. For

Table 3: Binomial regression data splitting: underfitting and PI coverage and length, int=1, a=4/3, m=4, B=1000

n	p/k	psi= ψ	mnnd/mnad	cov/len	lsapi	LVSpi	lsplitpi	splitpi	undfit
100	20	0.0000	40.9616	cov	0.9932	0.9850	0.9904	0.9752	0.0018
	1		4.6866	len	2.8414	2.6524	2.8588	2.5972	
100	100	0.2000	45.6434	cov	0.9948	0.9782	0.9886	0.9624	0.2238
	1		8.7554	len	2.8426	2.6696	2.8244	2.5558	
1000	20	0.5000	56.0040	cov	0.9874	0.9872	0.9890	0.9878	0.3164
	1		4.0666	len	2.4560	2.4374	2.4614	2.4392	
1000	1000	0.0000	95.0734	cov	0.9902	0.9820	0.9898	0.9822	0.0422
	1		5.3892	len	2.6302	2.4922	2.6320	2.4834	
1000	10	0.4000	63.4080	cov	0.9870	0.9856	0.9862	0.9858	0.9992
	9		5.2826	len	2.5050	2.4854	2.5008	2.4890	

the simulations, generating $\boldsymbol{x}^T\boldsymbol{\beta}$ is important for regression models other than multiple linear regression. For example, for binomial logistic regression, typically $-5 \leq \boldsymbol{x}^T\boldsymbol{\beta} \leq 5$ or there can be problems with the maximum likelihood estimator. Let $\boldsymbol{x} = (1\ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $m = p-1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\text{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (m-2)\psi^2]$. Hence the correlations are $cor(z_i, z_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. For multiple linear regression, let $\boldsymbol{u} = \boldsymbol{z}$. For the other regression models, let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence

$\boldsymbol{\beta} = (\beta_1, 1, .., 1, 0, ..., 0)^T$ with β_1, k ones and $p-k-1$ zeros. The default settings for Poisson regression use $\beta_1 = 1 = a$. The default settings for binomial regression with $m = 4$ trials use $\beta_1 = 1$ and $a = 4/3$. In the Table 3 caption, these values correspond to int=1, $a = 4/3$, and $m = 4$ while psi = ψ. The bootstrap sample for the prediction intervals had size $B = 1000$.

The terms lsapi, LVSpi, lsplitpi, splitpi, and noundfit appear on the first line of Tables 2, 3, and 5. Table 4 only used the Weibull regression prediction intervals. For the first two lines of numbers in Table 2, $n = 100$, $p = 20$ is the number of predictors including a constant, $k = 1$ nontrivial predictors were active, and psi=$\psi = 0.6$. The ψ value controls the correlation of the predictors and $\psi = 0$ means the predictors are uncorrelated. For $n = 100$, $n_d = 30$ if $a_d \leq 3$, and $n_d = 47$, otherwise. The value $mnad = 3.3482$ indicates the average number of fitted predictors, including a constant, in the simulation. Since $mnnd = 38.68$ is the average of the n_d in the simulation, typically data splitting used $n_d = 30$ with $a_d \leq 3$, but occasionally used $n_d = 47$. The prediction interval coverage is the proportion of the large sample 95% prediction intervals that contained Y_f where the test data case is $(\boldsymbol{x}_f, Y_f)$. With 5000 runs, a coverage < 0.94 indicates that the prediction interval was too short. For the first two lines of numbers in Table 2, the coverages were near 0.99 and the average prediction interval lengths were between 8.17 and 8.8. The proportion of the 5000 runs with underfitting was 0.2332.

For the Weibull regression model, there is no constant since the constant appears in the corresponding accelerated failure time model, which is a multiple linear regression model with right censored response $\log(Y)$. The data was generated as for the Poisson and Binomial regression, but replace $\boldsymbol{u}$ by $\boldsymbol{x}$ and $p-1$ by p. Let $SP = \boldsymbol{x}_i^T\boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations use $a = 1$ where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p - k$ zeros. The right censored Weibull regression data was generated in a manner similar to Zhou (2001) with $\gamma = 1$. The caption in Table 4 gives $a = 1$. The values gam= γ and clam in the caption control the Weibull distribution and the amount of right censoring.

Data splitting is useful for hypothesis testing and confidence intervals. The nominal 95% prediction intervals were used as a check for whether lasso was finding a useful model for

Table 4: Weibull regression data splitting: underfitting and PI coverage and length, n=100, a=1, gam=1, B=1000, clam=0.1

n	p/k	psi= ψ	mnnd/mnad	cov/len	LVSpi	splitpi	undfit
100	4	0.00	31.7646	cov	0.9550	0.9552	0.0174
	1		1.8314	len	5.5483	5.5033	
100	4	0.80	30.7004	cov	0.9574	0.9576	0.9688
	1		1.6076	len	5.5956	5.5384	
100	20	0.00	36.3172	cov	0.9326	0.9328	0.0506
	1		2.7178	len	5.9745	25.4093	
100	20	0.60	33.4238	cov	0.9510	0.9512	0.7422
	1		2.2570	len	5.7760	9.8008	
100	10	0.00	39.1528	cov	0.9518	0.9520	0.8784
	9		4.2938	len	6.9368	6.6001	
100	50	0.00	35.1850	cov	0.7750	0.7752	1.0
	19		2.1098	len	346.39	332.04	

Table 5: Multiple linear regression data splitting: underfitting and PI coverage and length, J=5, type=3

n	p/k	psi= ψ	mnnd/mnad	cov/len	lsapi	LVSpi	lsplitpi	splitpi	undfit
100	4	0.8000	33.3354	cov	0.9676	0.9672	0.9768	0.9764	0.1388
	1		2.6874	len	4.0502	4.0545	4.6570	4.6614	
1000	4	0.8000	36.0180	cov	0.9558	0.9564	0.9562	0.9562	0.1358
	1		2.6694	len	3.1302	3.1306	3.1333	3.1333	
1000	20	0.0000	55.7820	cov	0.9516	0.9474	0.9514	0.9490	0.0170
	1		3.5346	len	3.2048	3.2349	3.2139	3.2449	
1000	20	0.5000	64.9500	cov	0.9548	0.9548	0.9536	0.9528	0.0098
	1		4.7108	len	3.1909	3.1942	3.2011	3.2054	
1000	1000	0.0000	82.3860	cov	0.9558	0.9460	0.9572	0.9440	0.0898
	1		4.5362	len	3.3778	3.4472	3.4008	3.4644	

prediction (coverage near 0.95) even if underfitting was present. This result could occur for at least two reasons. First, as ψ increases to 1, the predictor variables are roughly $x_i = x_j + e_{ij}$ where the error magnitude rapidly gets close to 0 as $\psi \to 1$. Hence omitting some good predictors may not be a problem for prediction. Second, for some regression models, there are many linear combinations that give a good fit. See Theorem 3.

For multiple linear regression, the zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) $-$ 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Only distribution iii) is not symmetric. The lengths of the asymptotically optimal 95% PIs are i) $3.92 = 2(1.96)$, ii) 6.365, iii) 2.996, iv) $1.90 = 2(0.95)$, and v) 13.490.

For the regression methods, first consider $k = 1$. Often there was little underfit for $\psi = 0$. The amount of underfitting tended to increase with ψ, and to be worse with larger p. With $n = 1000$, not much more than 10% of the cases were used for H. For larger values of k, lasso often underfit, especially if $k = p - 1$ and $n/k < 10$. See Table 2 for Poisson regression, see Table 3 for Binomial regression, where with m=4, a 100% PI for Y_f is [0,4] with length

4. The nominal 95% PIs were shorter than 4 in Table 3. See Table 4 for proportional hazards regression where the two prediction intervals were made for Weibull regression. In Table 4, sometimes the two PI lengths differed. For the last two lines of Table 4, there was serious underfitting with low PI coverage and large PI length. See Table 5 for multiple linear regression where usually the data splitting PI and PI using all n cases had similar average lengths, but there were data configurations where using all n cases can give a much smaller length and better coverage.

6. Conclusions

Regression models, such as $Y|\boldsymbol{\beta}^T\boldsymbol{x}$, tend to be useful when they fit the data well. This paper shows that nonsparse models, such as OPLS, can be useful for inference even when n/p is not large. There are many problems with assuming that the regression model estimates a population generating model. Removing this assumption greatly increases the scope of data splitting, sparse fitted models, and nonsparse dimension reduction model selection estimators such as partial least squares. In particular, sparse fitted models, like lasso, tend to give poor approximations to a nonsparse population generating model, but this paper shows that the sparse fitted model can still be useful if data splitting is used.

Table 6: Regression Summary

	low dimensions	data splitting with sparse I	high dim. regularity conditions are too strong
general:	$\boldsymbol{\beta}(\boldsymbol{x},Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I,Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I,Y)$	$\boldsymbol{\beta}(\boldsymbol{x},Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I,Y)$
data splitting:	$\boldsymbol{\beta}(\boldsymbol{x},Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I,Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I,Y)$	$\boldsymbol{\beta}(\boldsymbol{x},Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I,Y)$
lasso:	$\boldsymbol{\beta}_{lasso}$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I,Y)$	$\boldsymbol{\beta}(\boldsymbol{x},Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I,Y)$
OPLS:	$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$	$\boldsymbol{\beta}_{I,OPLS} = \lambda_I\boldsymbol{\Sigma}_{\boldsymbol{x}_I,Y}$	$\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$
MMLE:	$\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$	$\boldsymbol{\beta}_{I,MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}_I,Y}$	$\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS}$

The multitude of models result is useful and simple. For fixed p, lasso in `glmnet` tends to be at best $n^{1/4}$ consistent for multiple linear regression, while large sample theory for lasso and elastic net does not appear to be available for GLMs and Cox regression. See Guan

and Tibshirani (2020). For fixed p, Rathnayake and Olive (2023) have the interesting result that if the sparse estimator is consistent for $\boldsymbol{\beta}$, then the sparse variable selection estimator (that applies OLS, the GLM, or the Cox regression estimator to the predictors with nonzero coefficients) is $\sqrt{n}$ consistent for $\boldsymbol{\beta}$. Thus $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$.

Table 6 summarizes what the regression estimators tend to estimate in low dimensions or after data splitting with a sparse fitted model I. The third column of Table 6 gives some results in the high dimensional literature where the regularity conditions are often too strong. In particular, often the regularity conditions are too strong for low dimensional results to hold in high dimensions.

Yüzbaşi, Arashi, and Ahmed (2020) has an interesting test. Taavoni and Arashi (2021) has useful references for high dimensional statistics.

Simulations were done in *R*. See R Core Team (2020). The collection of Olive (2023) *R* functions *slpack*, available from (http://parker.ad.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The functions for regression data splitting are `mlrsplitsim`, `prsplit`, `brsplitsim`, and `PHsplitsim`. These functions used the Friedman et al. (2015) `glmnet` package. The data set for the Hebbler (1847) example is available from the Olive (2017) website (http://parker.ad.siu.edu/Olive/lregdata.txt).

Acknowledgments

The authors thank the Editor, Associate Editor, and the referees for their work.

References

Basa, J., R. D. Cook, L. Forzani, and M. Marcos. 2022. Asymptotic distribution of one-component partial least squares regression estimators in high dimensions. *The Canadian Journal of Statistics* to appear. doi:10.1002/cjs.11755.

Chetverikov, D., Z. Liao, and V. Chernozhukov. 2022. On cross validated lasso in high dimensions. *The Annals of Statistics* 49 (3):1300-1317. doi:10.1214/20-AOS2000.

Chun, H., and S. Keleş. 2010. Sparse partial least squares regression for simultaneous dimension reduction and predictor selection. *Journal of the Royal Statistical Society: Series*

B (Statistical Methodology) 72 (1):3-25. doi:10.1111/j.1467-9868.2009.00723.x.

Cook, R. D. 2018. *An introduction to envelopes: Dimension reduction for efficient estimation in multivariate statistics*. Hoboken, NJ: Wiley.

Cook, R. D., and L. Forzani. 2018. Big data and partial least squares prediction. *The Canadian Journal of Statistics* 46 (1):62-78. doi:10.1002/cjs.11316.

Cook R. D., and L. Forzani. 2019. Partial least squares prediction in high-dimensional regression. *The Annals of Statistics* 47 (2):884-908. doi:10.1214/18-AOS1681.

Cook, R. D., I. S. Helland, and Z. Su. 2013. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (5):851-877. doi:10.1111/rssb.12018.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 34 (2):187-220. doi:10.1111/j.2517-6161.1972.tb00899.x.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression (with discussion). *The Annals of Statistics* 32 (2):407-451. doi:10.1214/009053604000000067.

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (5):849-911. doi:10.1111/j.1467-9868.2008.00674.x.

Fan, J., and R. Song. 2010. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* 38 (6):3217-3841. doi:10.1214/10-AOS798.

Friedman, J., T. Hastie, N. Simon, and R. Tibshirani. 2015. *glmnet*: Lasso and elastic-net regularized generalized linear models. *R* package version 2.0. http://cran.r-project.org/package=glmnet.

Guan, L., and R. Tibshirani. 2020. Post model-fitting exploration via a "next-door" analysis. *The Canadian Journal of Statistics* 48 (3):447-470. doi:10.1002/cjs.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press Taylor & Francis.

Hebbler, B. 1847. Statistics of Prussia. *Journal of the Royal Statistical Society, A* 10 (2):154-186. doi:10.2307/2337688.

Meinshausen, N. 2007. Relaxed lasso. *Computational Statistics & Data Analysis* 52 (1):374-393. doi:10.1016/j.csda.2006.12.019.

Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society, A* 135 (3):370-380. doi:10.2307/2344614.

Olive, D. J. 2013. Plots for generalized additive models. *Communications in Statistics - Theory and Methods* 42 (18):2610-2628. doi:10.1080/03610926.2011.628772.

Olive, D. J. 2017. *Linear regression*. New York: Springer.

Olive, D. J. 2023. *Prediction and statistical learning*. Online course notes. http://parker.ad.siu.edu/Olive/slearnbk.htm.

Olive, D. J., R. C. Rathnayake, and M. G. Haile. 2022. Prediction intervals for GLMs, GAMs, and some survival regression models. *Communications in Statistics - Theory and Methods* 51 (22):8012-8026. doi:10.1080/03610926.2021.1887238.

Pelawa Watagoda, L. C. R., and D. J. Olive. 2021. Comparing six shrinkage estimators with large sample theory and asymptotically optimal prediction intervals. *Statistical Papers* 62 (5):2407-2431. doi:10.1007/s00362-020-01193-1.

Qi, X., R. Luo, R. J. Carroll, and H. Zhao. 2015. Sparse regression by projection and sparse discriminant analysis. *Journal of Computational and Graphical Statistics* 24 (2):416-438. doi:10.1080/10618600.2014.907094.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. www.R-project.org.

Rathnayake, R. C., and D. J. Olive. 2023. Bootstrapping some GLMs and survival regression models after variable selection. *Communications in Statistics - Theory and Methods* 52 (3):2625-2645. doi:10.1080/03610926.2021.1955389.

Rinaldo, A., L. Wasserman, and M. G'Sell. 2019. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *The Annals of Statistics* 47 (6):3438-3469. doi:10.1214/18-AOS1784.

Su, Z., and R. D. Cook. 2012. Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika* 99 (3):687-702. doi:10.1093/biomet/ass024.

Taavoni, M., and M. Arashi. 2021. High-dimensional generalized semiparametric model for longitudinal data. *Statistics* 55 (4):831-850. doi:10.1080/02331888.2021.1977304.

Tay, J. K., B. Narasimhan, and T. Hastie. 2023. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software* 106 (1):1-31. doi:10.18637/jss.v106.i01.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58 (1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Wieczorek, J., and J. Lei. 2022. Model-selection properties of forward selection and sequential cross-validation for high-dimensional regression. *Canadian Journal of Statistics* 50 (2):454-470. doi:10.1002/cjs.11635.

Wold, H. 1975. Soft modelling by latent variables: the non-linear partial least squares (NIPALS) approach. *Journal of Applied Probability* 12 (S1):117-142. doi:10.1017/S0021900200047604.

Yüzbaşi, B., M. Arashi, and S. E. Ahmed. 2020. Shrinkage estimation strategies in generalized ridge regression models: low/high-dimension regime. *International Statistical Review* 88 (1):229-251. doi:10.1111/insr.12351.

Zhang, L. (2022). Data Splitting Inference. (Ph.D. Thesis), Southern Illinois University, USA, at (http://parker.ad.siu.edu/Olive/slinglingphd.pdf).

Zhou, M. 2001. Understanding the Cox regression models with time–change covariates. *The American Statistician* 55 (2):153-155. doi:10.1198/000313001750358491.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301-320. doi:10.1111/j.1467-9868.2005.00503.x.