

# Regularized Partial Least Squares Estimation of an Approximate Envelope Model for Supervised Dimension Reduction in Time Series Forecasting\*

John C. Chao<sup>1</sup> and Norman R. Swanson<sup>2</sup>

<sup>1</sup>University of Maryland and <sup>2</sup>Rutgers University

First Version: March 10, 2025

This Version: September 25, 2026

## Abstract

This paper introduces regularized partial least squares (RPLS) estimators for time series forecasting using a many-predictor regression model that has a so-called “approximate envelope” structure. The proposed envelope framework allows for dimension reduction in an inherently supervised or targeted manner that takes into account possible relationships between the predictor variables and the target variable of interest. The envelope structure not only generalizes the standard many-predictor regression framework, but also a diffusion model framework. In particular, we establish that given an approximate envelope structure, the conditional mean function of the many-predictor regression is asymptotically equivalent to the conditional mean function of a diffusion index forecast equation, when the latter is the true underlying model. This and other asymptotic consistency and convergence rate results given in the paper establish that RPLS procedures are effective for estimating the conditional mean function of both of these types of models. Findings also indicate that RPLS estimators benefit from additional key features such as adaptability to alternative statistical environments not treatable using the unregularized PLS estimator. For example, our estimators are able to not only take advantage of the possibility for dimension reduction due to the presence of a factor structure but can also make use of parameter sparsity that may be present in the underlying data generating mechanism. An empirical example based on forecasting various macroeconomic variables illustrates the use of the new estimators.

*Keywords:* Envelope model, factor analysis, forecasting, partial least squares, regularization, adaptive Lasso, hard thresholding, SCAD, soft thresholding.

*JEL classification:* C32, C33, C38, C52, C53, C55, C58.

\*John C. Chao, Department of Economics, 7343 Preinkert Drive, University of Maryland, jcchao@umd.edu. Norman R. Swanson, Department of Economics, 9500 Hamilton Street, Rutgers University, nswanson@rutgers.edu. We are grateful to Marine Carrasco, Saraswata Chaudhuri, Jean-Marie Dufour, Yuan Liao, Laura Liu, Yixiao Sun as well as participants of the 2025 IAAE Annual Conference, the 2025 Econometric Society World Congress, the 2025 NBER-NSF Time Series Conference, the 2025 Conference Celebrating Don Andrews, the 18th Greater New York Metropolitan Area Econometrics Colloquium, and the Montreal Econometrics Seminar for their helpful comments. Chao and Swanson would like to thank the University of Maryland and Rutgers University, respectively, for research support.

# 1 Introduction

The impressive number of high dimensional and high frequency datasets now available to applied practitioners has been a driver for many recent advances in the areas of machine learning, data mining, variable selection, and dimension reduction. In the area of dimension reduction, many frequently utilized methods center around the use of factor models, whereby the variation in a high-dimensional data vector is assumed to be well explained by a small number of latent factors. The large sample properties of principal component analysis methods used for estimation in such models has been studied in a literature that includes important contributions by Stock and Watson (2002), Bai (2003), Forni, Hallin, Lippi, and Reichlin (2004, 2005), and many others. It is well documented that the original method of principal component analysis (PCA) is an unsupervised procedure in the sense that factor extraction based on PCA often focuses on selecting factors that explain the most variation in the predictor variables, without taking into account whether the available predictor variables or any linear combination of them actually covary with a target variable being predicted. Needless to say, when the objective is forecasting it is desirable to carry out supervised dimension reduction that takes into account possible relationships between the predictor variables and the target variable of interest. One such targeted procedure is partial least squares (PLS), as recently discussed in Carrasco and Rossi (2016), Groen and Kapetanios (2016), and Ahn and Bae (2022), who present various interesting theoretical, simulation, and empirical results showing how PLS can be useful for macroeconomic and financial forecasting in data rich environments. These papers consider unregularized PLS estimation. In this paper, we introduce a so-called approximate envelope model which not only nests models such as the many-predictors regression and a standard diffusion index model, but also motivates our development of new regularized PLS estimators. These regularized PLS estimators are adaptive to different ways in which the underlying model may be parsimonious, a feature which can be used to facilitate improved forecasting performance, especially in high-dimensional settings.

To provide a bit of background on PLS-type procedures, note that although the original (unregularized) PLS estimator was first introduced in the early 1980s by Herman Wold in a series of papers (see Wold (1982, 1985) and Wold, Martens, and Wold (1983)), and although variants of the method have been applied successfully in many areas of applied science, PLS-type procedures were arguably somewhat slow to be adopted in both the econometrics and the statistics literatures. One reason for this, as noted in Cook, Helland, and Su (2013) and Cook and Forzani (2019), is that PLS was originally presented as an algorithm without the associated population modeling framework needed for analyzing many of its theoretical properties. Recently, however, an interesting new paradigm for targeted or supervised dimension reduction has been proposed by Cook, Li, and Chiaromonte (2010) and Cook, Helland, and Su (2013) that lays the foundation for such a population model. These authors propose what they call the envelope model. In the context of the many-predictors regression setup, their model provides a framework for estimating the smallest reducing subspace that contains information about the coefficient vector of a high-dimensional regression. This

smallest reducing subspace is called the envelope because it “envelopes” the coefficient vector of the high-dimensional regression, and the coefficient vector contains the essential parameters that define the relationship between the predictors and the target variable in a linear regression. Dimension reduction in this context is achieved by estimating the smallest such subspace. Cook and Forzani (2019) analyze the asymptotic properties of the (unregularized) PLS estimator in the context of an envelope model with *i.i.d.* data, thus not allowing for the kind of temporal dependence and possible heterogeneity in the second moments that typically characterizes the kind of high-dimensional time series environments often considered relevant for forecasting economic time series. In addition, the predictor envelope model studied in these papers does not accommodate the situation where the high-dimensional vector of predictor variables is generated by an approximate factor model whose error component exhibits general cross-sectional dependence, in addition to temporal dependence. This leads to one of the contributions of this paper. Namely, we generalize the envelope setup to the types of complicated data rich environments often encountered in time series forecasting by proposing what we call an “approximate envelope model” that not only allows for temporal dependence and heterogeneity of second moments but also allows, in the factor case, for the error vector of the factor model to have a general covariance matrix, leading to a (matrix) perturbation of the usual envelope representation.

Using this approximate envelope setup as the population modeling framework for PLS-type procedures, we provide results which yield new insights into PLS-type procedures and motivate our introduction of new regularized PLS estimators. We also provide results establishing that under the additional structure imposed by our approximate envelope model, a many-predictors forecasting equation is closely related to a diffusion index forecast equation, where the target variable depends on latent factors, and not directly on the predictor variables. More precisely, we show that the conditional mean function of the former converges to that of the latter when the coefficient vector of the many-predictors forecasting regression has an envelope structure and is modeled as being in a  $1/\sqrt{\underline{\lambda}_1}$  neighborhood of zero, where  $\underline{\lambda}_1$  is the largest eigenvalue of the covariance matrix associated with the common component of the underlying factor model. This result implies that the conditional mean function of the diffusion index forecast equation can be well approximated by the conditional mean function of a many-predictors forecast regression that has an approximate envelope structure, even though the two models may look very different at first glance. Moreover, under the  $1/\sqrt{\underline{\lambda}_1}$  local-to-zero parameterization, the asymptotic correspondence between the diffusion index forecast equation and the many-predictors forecast equation is general enough so that it holds not only under the conventional strong factor assumption, where  $\underline{\lambda}_1$  grows at the rate  $N$ , but also under many situations where factors are weaker than stipulated by a conventional factor pervasiveness assumption, i.e.,  $\underline{\lambda}_1$  may grow at a rate less than  $N$ . This result, taken together with a number of other results given in this paper, yields insights about why PLS-type procedures can be effective in recovering the conditional mean function of diffusion index forecast equation.

Our regularized PLS estimators are constructed by using regularized estimators of the covariance matrix of  $X_t$  and of the covariance vector between  $X_t$  and  $y_{t+h}$ , instead of their

sample analogues, where  $X_t$  and  $y_{t+h}$  denote, respectively, the vector of predictor variables and the target variable of interest. A key benefit of our regularized PLS estimators is that they are more adaptive to alternative statistical environments, since they are designed to be able to take advantage not only of the type of dimension reduction that is available due to the presence of a factor structure, but also of any parameter sparsity, exact or approximate, that may be present<sup>1</sup>. In summary, we propose two different classes of regularized PLS estimators: one is referred to as the class of doubly regularized PLS (or DRPLS) estimators and the other as the class of generalized-inverse-based PLS (or GRPLS) estimators. As explained in Section 6, DRPLS and GRPLS are both designed to facilitate application to real world data where one may encounter situations in which PLS-type estimators based on the ordinary inverse can be affected by certain numerical instability issues when the procedure is being trained<sup>2</sup>. Consistency and convergence rate result are discussed for both estimators. Interestingly, the rates of convergence of these estimators is shown to clearly reflect the trade-off between the amount of parameter sparsity and the strength of the underlying factors, implying that the presence of parameter sparsity can help to compensate for situations where there are weaker factors. Similar consistency and convergence rate results are also given for point forecasts constructed using DRPLS and GRPLS. These asymptotic results on point forecasts reveal a certain robustness of our regularized PLS procedures in the sense that point forecasts constructed from DRPLS and GRPLS estimators can consistently estimate the conditional mean function of the underlying forecast equation regardless of whether the true forecast equation is a many-predictors regression or a diffusion index model<sup>3</sup>.

The rest of this paper is organized as follows. In Section 2, we provide motivation and background on the envelope framework. Section 3 provides discussion of the relationship between the diffusion index forecast equation and the many-predictors forecast equation that has an envelope structure and gives formal results showing that the conditional mean function of the former can be well approximated by that of the latter. Section 4 describes the approximate envelope model and discusses the assumptions that we make on this model. Section 5 explores the connections between PLS-type procedures and the approximate envelope model, and gives results which yield insights into why PLS-type procedures can be effective in estimating both the many-predictors regression and the diffusion index model. In Section 6, we introduce DRPLS and GRPLS, and provide results characterizing the asymp-

---

<sup>1</sup>It should be emphasized that our regularized PLS procedures can take advantage not only of situations where there is exact sparsity but also of situations where there is approximate sparsity. As discussed in detail in Remark 6.2 (b), approximate sparsity is a situation where a significant number of parameters are not exactly zero but may be relatively small in magnitude. Hence, unlike the situation with exact sparsity, a model with approximate sparsity can be viewed as a type of dense model in the terminology given in the well-known paper by Giannone, Lenza, and Primiceri (2021).

<sup>2</sup>Refer to part (b) of Remark 6.1 for further discussion of this issue.

<sup>3</sup>Although we give results showing that the conditional mean function of the diffusion index forecast equation is asymptotically equivalent to that of an envelope-type many-predictors forecast equation, one might still worry that there may be some errors-in-variables type issues that can arise in approximating the diffusion index model with a many-predictors regression. It turns out that, unlike OLS, PLS-type estimation procedures are less susceptible to the type of errors-in-variables issues that can arise in our context, as explained in Section 6.

otic properties of these estimators and associated point forecasts both in the case where the true forecast equation is a many-predictors regression with an envelope structure and also in the case where the true model is a diffusion index model. Section 7 presents an empirical illustration comparing the forecast results based on our regularized PLS procedures, PCA, unregularized PLS, and a benchmark AR model. Section 8 contains concluding remarks. Proofs of Theorems 3, 5(a), and 6(a) are given in a Supplemental Appendix whereas the proofs of Theorems 5(b) and 6(b) as well as all the remaining theorems and supporting lemmas are given in an additional Online Appendix<sup>4</sup>. Before proceeding further, we will say a few words on some of the commonly used notation in the sequel. We use  $\mathbb{R}$  to denote the set of real numbers and  $\mathbb{Z}$  to denote the set of integers. For our matrix notation, we use  $\lambda_{\min}(A)$ ,  $\lambda_{\max}(A)$ , and  $\text{tr}(A)$  to denote, respectively, the minimal eigenvalue, the maximal eigenvalue, and the trace of a square matrix  $A$ .  $A'$  denotes the transpose of a (not necessarily square) matrix  $A$ .  $\|a\|_2$  denotes the usual Euclidean norm when applied to a (finite-dimensional) vector  $a$ . Moreover, for a matrix  $A$ ,  $\|A\|_2 \equiv \sqrt{\lambda_{\max}(A'A)}$  denotes the matrix spectral norm while  $\|A\|_F \equiv \sqrt{\text{tr}\{A'A\}}$  denotes the Frobenius norm. Additionally, an  $N \times \kappa$  matrix  $L$ , with  $N > \kappa$ , is said to be semi-orthogonal if  $L'L = I_\kappa$ . We take  $\text{diag}(a_1, \dots, a_\kappa)$  to denote a  $\kappa \times \kappa$  diagonal matrix whose diagonal elements are given by  $a_1, \dots, a_\kappa$ . We write  $f(N, T) \asymp g(N, T)$  if  $f(N, T)/g(N, T) = O(1)$  and  $g(N, T)/f(N, T) = O(1)$  as  $N, T \rightarrow \infty$ ; we write  $f(N, T) \ll g(N, T)$  if  $f(N, T)/g(N, T) \rightarrow 0$  as  $N, T \rightarrow \infty$ ; and we write  $f(N, T) \lesssim g(N, T)$  if either  $f(N, T) \ll g(N, T)$  or  $f(N, T) \asymp g(N, T)$ . Finally, note that the abbreviation w.p.a.1 stands for “with probability approaching one”; and we let  $\sigma(X_{t+m}, X_{t+m+1}, X_{t+m+2}, \dots)$  denote the  $\sigma$ -field generated by  $X_{t+m}, X_{t+m+1}, X_{t+m+2}, \dots$ .

## 2 Motivation and Some Background on the Envelope Model

We begin by providing some background on the envelope model in order to motivate the setup used in this paper. Consider the many-predictors forecasting equation given by:

$$y_{t+h} = \alpha + (X_t - \mu_X)' \beta_{\text{env}} + \varepsilon_{t+h}, \quad (1)$$

where  $h$ , the forecast horizon, is taken to be a fixed positive integer. Here,  $X_t$  denotes an  $N \times 1$  vector of predictors. The order of magnitude of  $N$  may be smaller, the same, or even larger than  $T$ . The goal of the envelope framework, in the context of this forecasting problem, is to try to remove that part of the variation of  $X_t$  which is immaterial or irrelevant for the prediction of  $y_{t+h}$  by focusing attention on the smallest reducing subspace that contains information about  $\beta_{\text{env}}$ , called the envelope by Cook, Li, and Chiaromonte (2010).

To discuss things in a bit more detail, suppose that  $\Sigma_X = E[(X_t - \mu_X)(X_t - \mu_X)']$  does not vary across  $t$ , an assumption that we will later relax. Now, let  $\mathcal{R}$  denote a reducing subspace of  $\Sigma_X$ , i.e.  $\mathcal{R}$  is a subspace of  $\mathbb{R}^N$  such that the following two conditions are

---

<sup>4</sup>Both the Supplemental Appendix and the Online Appendix can be viewed at the authors' websites.

satisfied:

- (i)  $v \in \mathcal{R}$  implies that  $\Sigma_X v \in \mathcal{R}$ ,
- (ii)  $v \in \mathcal{R}^\perp$  implies that  $\Sigma_X v \in \mathcal{R}^\perp$ .

In addition, let  $\mathcal{B}$  denote the span of  $\beta_{\text{env}}$ . Then, the  $\Sigma_X$ -envelope of  $\mathcal{B}$ , typically denoted by  $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ , is defined as the smallest reducing subspace that contains  $\mathcal{B}$ , or, alternatively, the intersection of all reducing subspaces of  $\Sigma_X$  that contain  $\mathcal{B}$ . Let  $\kappa$  be the dimension of  $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ . It can be shown that  $\mathcal{E}_{\Sigma_X}(\mathcal{B})$  is the  $\Sigma_X$ -envelope of  $\mathcal{B}$  if and only if we can write  $\Sigma_X$  in terms of the following decomposition:

$$\Sigma_X = L_1 \Delta_1 L_1' + L_2 \Delta_2 L_2', \quad (2)$$

where  $L_1$  is a  $N \times \kappa$  semi-orthogonal matrix whose columns are those orthonormal eigenvectors of  $\Sigma_X$  which span  $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ . Also,  $L_2$  is a  $N \times (N - \kappa)$  semi-orthogonal matrix whose columns are those orthonormal eigenvectors of  $\Sigma_X$  which span the orthogonal complement of  $\mathcal{E}_{\Sigma_X}(\mathcal{B})$ . Finally,  $\Delta_1 = \text{diag}(\lambda_1, \dots, \lambda_\kappa)$  and  $\Delta_2 = \text{diag}(\lambda_{\kappa+1}, \dots, \lambda_N)$  contain the eigenvalues of  $\Sigma_X$  associated with the columns of  $L_1$  and  $L_2$ , respectively.<sup>5</sup> It is easy to see that expression (2) is just a rearrangement of the spectral decomposition of  $\Sigma_X$  into two groups of eigenvalues/eigenvectors: a group that provides useful information for the purpose of forecasting and a group that does not. This rearrangement is meaningful because it formalizes a setting where a relationship exists between the regression parameter  $\beta_{\text{env}}$  and certain eigenvectors of  $\Sigma_X$ , which, in turn, opens the door for doing targeted or supervised dimension reduction. Indeed, under the above framework,  $\beta_{\text{env}} \in \mathcal{E}_{\Sigma_X}(\mathcal{B}) = \text{span}(L_1)$ , from which it follows that there exists a  $\kappa \times 1$  vector  $\varphi$  such that  $\beta_{\text{env}} = L_1 \varphi$ . Hence, in the case where  $\kappa$  is smaller than  $N$ , we have reduced the dimension of the unknown parameter vector from the  $N \times 1$  vector  $\beta_{\text{env}}$  to a  $\kappa \times 1$  vector  $\varphi$ , provided we can properly deal with the basis matrix  $L_1$ , as we will explain later. The implication of this reduction for forecasting is illustrated by the following proposition, which is a slightly modified version of Lemma 4.1 part (i) of Cook (2018) and also of part (ii) of Proposition 1 of Cook, Helland, and Su (2013) and is given here for the purpose of motivation.

**Proposition 1:** Consider the many-predictors forecast equation given in expression (1). Suppose that the following assumptions hold: (P1)  $\{X_t\}$  is covariance stationary with mean vector  $\mu_X$  and covariance matrix  $\Sigma_X$ . (P2) Let  $\Sigma_X = L_1 \Delta_1 L_1' + L_2 \Delta_2 L_2'$  be the decomposition defined in expression (2) and suppose that  $\beta_{\text{env}} \in \text{span}(L_1)$ . (P3)  $E[\varepsilon_{t+h}|X_t] = 0$  a.s., for all  $t \in \mathbb{Z}$ . Under these conditions,  $\text{cov}(y_{t+h}, L_2' X_t | L_1' X_t) = 0$ .

**Remark 2.1:** A clear implication of the result given in Proposition 1 is that conditional on  $L_1' X_t$ ,  $L_2' X_t$  will not be helpful in predicting  $y_{t+h}$ , so that dimension reduction can be

---

<sup>5</sup> A formal statement and proof of this result is given in Lemma D-8 in Appendix D of the Online Appendix. This lemma is related to Proposition 2.1 and Lemma S1.3 of Cook, Li, and Chiaromonte (2010), which we have modified slightly to adapt to our setup, and which is included in the Online Appendix for expositional purposes only.

achieved by focusing only on  $L_1'X_t$ . Note also that this approach to dimension reduction is targeted since it makes use of the relationship between  $\beta_{\text{env}}$  and the eigen-structure of  $\Sigma_X$ . As mentioned in the introduction, a natural way to estimate the many-predictors forecast equation (1) when it has an envelope structure is to use partial least squares (PLS) type estimators. It is also worth noting that PLS-type estimation in this case has a convenient structure that ensures that we do not actually have to explicitly identify and estimate the basis matrix  $L_1$ . See Sections 5 and 6 for a more detailed discussion of the benefits of using PLS-type procedures in this context. Finally, there turns out to be a close connection between many-predictors forecast regression with an envelope structure and the diffusion index model of Stock and Watson (1998, 2002), in spite of the fact that they may look very different on the surface. We will explore this relationship in greater depth in the sequel.

**Remark 2.2:** (a) It should be noted that the decomposition given in expression (2) is not general enough for certain applications in economics. This is particularly true if the vector of predictor variables,  $X_t$ , is generated by an approximate factor model given by  $X_t = \mu_X + \Gamma(F_t - \mu_F) + u_t$ , where  $F_t$  is a vector of latent factors and the error vector,  $u_t$ , exhibits general cross-sectional dependence. In this case, under standard conditions discussed in the next section,  $\Sigma_X = \Sigma_{\text{com}} + \Sigma_u$ , where  $\Sigma_{\text{com}}$  is the covariance matrix associated with the common component  $\Gamma(F_t - \mu_F)$ , and where  $\Sigma_u$  is the covariance matrix associated with the error component  $u_t$ , with nonzero off-diagonal elements due to cross-sectional dependence. Under this scenario, it is reasonable to model supervised or targeted dimension reduction in terms of selecting a small subset of the eigenvectors of  $\Sigma_{\text{com}}$  that contains information about  $\beta_{\text{env}}$ , since in this setup  $\Sigma_{\text{com}}$  is the covariance matrix associated with the signal component and since the error component is not likely to contain useful information about  $\beta_{\text{env}}$  and, in addition, no dimension reduction is usually considered to be achievable with respect to the error component. This means that when specifying the underlying population model, the columns of the matrix  $L_1$  should be taken to be a subset of the eigenvectors of the covariance matrix  $\Sigma_{\text{com}}$ , and not of  $\Sigma_X$ , which is contaminated by the presence of  $\Sigma_u$ . However, by proceeding in this manner we are not, in general, able to obtain a representation of the form given by expression (2). Thus, in Section 4 we introduce what we refer to as the approximate envelope model, and its corresponding envelope decomposition

$$\Sigma_X = L_1\Delta_1L_1' + L_2\Delta_2L_2' + \Sigma_u. \quad (3)$$

Note that expression (3) is similar in form to the decomposition given in (2), but is perturbed by  $\Sigma_u$ .<sup>6</sup> To differentiate between the two, we refer to the decomposition (2) as an exact envelope decomposition and the decomposition (3) as an approximate envelope decomposition.

---

<sup>6</sup>To give further intuition on expression (3), consider the conventional situation where  $X_t$  is generated by an approximate factor model with all available factors being useful for predicting some target variable of interest  $y_{t+h}$ . Specializing to this case, it turns out that expression (3) reduces to  $\Sigma_X = L_1\Delta_1L_1' + \Sigma_u$ . This is because, in this case, the columns of  $L_2$  are the eigenvectors associated with the 0 eigenvalues of  $\Sigma_{\text{com}}$ . On the other hand, the diagonal elements of  $\Delta_1$  will be the nonzero eigenvalues of  $\Sigma_{\text{com}}$  and the columns of  $L_1$  will be the associated eigenvectors. A more detailed discussion of the relationship between envelope-type models and factor models is given in Section 3.

(b) It should also be noted that our analysis focuses on an approximate version of what is called the predictor envelope model. As discussed above, we aim to reduce the dimensionality of the vector of predictor variables  $X_t$  by focusing only on those linear transformations  $L_1' X_t$  that contain information about the target variable of interest  $y_{t+h}$ . There are also papers in the literature that study the so-called response envelope model. Unlike predictor envelope models, a response envelope model focuses on the problem of reducing the dimensionality of  $Y_t$ , where  $Y_t$  can denote either a vector of dependent variables in a multivariate regression or even a matrix of dependent variables. Some interesting recent papers that propose novel response envelope specifications in the context of large VAR models include Samadi and Herath (2024), wherein a reduced-rank envelope VAR (REVAR) model is introduced, and Samadi and De Alwis (2025) wherein the envelope matrix autoregressive (EMAR) model is introduced. We do not consider response envelope models because they do not fall within the class of problems addressed in this paper.

### 3 Connections between the Many-Predictors Regression and the Diffusion Index Model

Before we formally discuss the approximate envelope model, it is worthwhile to first explore the connections between a many-predictors regression with an envelope-type structure and the diffusion index model. Exploring these connections allows us to provide insight into why PLS-type procedures are useful for estimating not only many-predictors regressions but also diffusion index models. To proceed, consider a diffusion index type model where both the forecasting equation and the vector of predictors,  $X_t$ , depend on a  $K \times 1$  vector of latent factors,  $F_t$ :

$$y_{t+h} = \alpha + (F_t - \mu_F)' \varphi + \varepsilon_{t+h}, \quad (4)$$

$$X_t = \mu_X + \Gamma (F_t - \mu_F) + u_t, \text{ , for } t = 1, \dots, T_h, \text{ with } T_h = T - h. \quad (5)$$

Let  $\underline{\Lambda} = \text{diag}(\underline{\lambda}_1, \dots, \underline{\lambda}_K)$ , where the diagonal elements of  $\underline{\Lambda}$  are the nonzero eigenvalues of the common component covariance matrix  $\Gamma \bar{\Sigma}_F \Gamma'$ , and where  $\bar{\Sigma}_F = T_h^{-1} \sum_{t=1}^{T_h} E[(F_t - \mu_F)(F_t - \mu_F)']$ . Assume that:

$$\underline{\lambda}_1 \asymp \dots \asymp \underline{\lambda}_K \text{ and } \underline{\lambda}_1 > \dots > \underline{\lambda}_K > 0, \quad (6)$$

so that the eigenvalues associated with the common component are all of the same order of magnitude and have algebraic multiplicity equal to 1. Also, assume that:

$$\frac{\Gamma' \Gamma}{\underline{\lambda}_1} = I_K \text{ and } \bar{\Sigma}_F = \frac{\underline{\Lambda}}{\underline{\lambda}_1}. \quad (7)$$

Assumptions similar to those given in expressions (6) and (7) are imposed in Stock and Watson (2002) and Ahn and Bae (2022). Since estimation by partial least squares does not require that we separately identify  $\Gamma$  and  $F_t$ , the normalization assumptions given in expres-

sion (7) are not restrictive. This is so because even if we initially start with a specification:

$$y_{t+h} = \alpha + \left( \underline{F}_t - \underline{\mu}_F \right)' \varphi + \varepsilon_{t+h}, \quad (8)$$

$$X_t = \mu_X + \underline{\Gamma} \left( \underline{F}_t - \underline{\mu}_F \right) + u_t, \quad (9)$$

where the factor loading matrix  $\underline{\Gamma}$  and the factor covariance matrix

$\overline{\Sigma}_F = T_h^{-1} \sum_{t=1}^{T_h} E \left[ (\underline{F}_t - \underline{\mu}_F) (\underline{F}_t - \underline{\mu}_F)' \right]$  do not satisfy the conditions given in expression (7), we can always transform the model in an appropriate way so that these conditions will be satisfied. We show this explicitly in the proof of Lemma D-1 in Appendix D of the Online Appendix. One difference between the normalization given on the factor loading matrix in expression (7) and that given in Stock and Watson (2002) and Ahn and Bae (2022) is that we divide  $\Gamma' \Gamma$  by the largest eigenvalue  $\underline{\lambda}_1$ , whereas they divide  $\Gamma' \Gamma$  by  $N$ . We do this to allow for the possibility that the factors may be weaker than what is typically assumed under a factor pervasiveness or strong factor condition. In the case where the factor pervasiveness holds,  $\underline{\lambda}_1$  will grow at rate  $N$ . In addition, note that the normalization of Stock and Watson (2002) and Ahn and Bae (2022) holds asymptotically, i.e.,  $\Gamma' \Gamma / N \rightarrow I_K$ . The results in part (c) of Lemma D-1 show that this is not necessary, since the factor model can always be transformed so that  $\Gamma' \Gamma / \underline{\lambda}_1 = I_K$  holds exactly.

Now, given the model specified in expressions (4)-(7), it is easy to see that under the additional assumption that  $E[(F_t - \mu_F) u_t'] = 0 \forall t$ , we can write the covariance matrix of the predictors as follows:

$$\begin{aligned} \overline{\Sigma}_X &= \frac{1}{T_h} \sum_{t=1}^{T_h} E \left[ (X_t - \mu_X) (X_t - \mu_X)' \right] \\ &= \Gamma \left( \frac{1}{T_h} \sum_{t=1}^{T_h} E \left[ (F_t - \mu_F) (F_t - \mu_F)' \right] \right) \Gamma' + \frac{1}{T_h} \sum_{t=1}^{T_h} E [u_t u_t'] \\ &= \Gamma \overline{\Sigma}_F \Gamma' + \overline{\Sigma}_u \\ &= \Gamma \left( \frac{\underline{\Lambda}}{\underline{\lambda}_1} \right) \Gamma' + \overline{\Sigma}_u \\ &= \Sigma_{\text{com}} + \overline{\Sigma}_u, \end{aligned} \quad (10)$$

where  $T_h = T - h$  and where

$$\Sigma_{\text{com}} = \Gamma \left( \frac{\underline{\Lambda}}{\underline{\lambda}_1} \right) \Gamma' = \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} \underline{\Lambda} \frac{\Gamma'}{\sqrt{\underline{\lambda}_1}} \quad (11)$$

denotes the contribution of the common component to the covariance matrix  $\overline{\Sigma}_X$ . Now, the columns of the matrix  $\Gamma / \sqrt{\underline{\lambda}_1}$  are clearly the orthonormal eigenvectors of the matrix  $\Sigma_{\text{com}}$

associated with its nonzero eigenvalues  $\underline{\lambda}_1, \dots, \underline{\lambda}_K$  since

$$\Sigma_{\text{com}} \cdot \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} = \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} \underline{\Lambda} \frac{\Gamma'}{\sqrt{\underline{\lambda}_1}} \cdot \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} = \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} \underline{\Lambda} \frac{\Gamma' \Gamma}{\underline{\lambda}_1} = \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} \underline{\Lambda}.$$

It turns out that there is an interesting connection between the diffusion index model described above and the many-predictors regression model given by

$$\begin{aligned} y_{t+h} &= \alpha + (X_t - \mu_X)' \beta + \varepsilon_{t+h} \\ X_t &= \mu_X + \Gamma (F_t - \mu_F) + u_t, \end{aligned}$$

if the coefficient vector  $\beta$  has a certain local-to-zero envelope structure. As mentioned in Section 2, a key feature of a predictor envelope model is that the coefficient vector  $\beta$  of the model has a representation as a linear combination of the columns of  $L_1$ , where  $L_1$  is a semi-orthogonal matrix whose column vectors together form a basis that spans the envelope subspace. Now, in the particular case where  $X_t$  is generated by a factor model, the columns of  $L_1$  are just those leading eigenvectors of  $\Sigma_{\text{com}}$  (alternatively, those columns of  $\Gamma/\sqrt{\underline{\lambda}_1}$ ) which have predictive content for the target variable  $y_{t+h}$ . In this context, we begin by considering the case where all  $K$  eigenvectors associated with the nonzero eigenvalues of  $\Sigma_{\text{com}}$  are useful for predicting  $y_{t+h}$ . For this case,  $L_1 = \Gamma/\sqrt{\underline{\lambda}_1}$ . Now, suppose that we take:

$$\beta = \frac{\beta_{\text{env}}}{\sqrt{\underline{\lambda}_1}} = \frac{L_1 \varphi}{\sqrt{\underline{\lambda}_1}} = \left( \frac{\Gamma}{\sqrt{\underline{\lambda}_1}} \right) \left( \frac{\varphi}{\sqrt{\underline{\lambda}_1}} \right), \quad (12)$$

so that  $\beta$  is a linear combination of the columns of  $L_1$  but normalized by having been divided by  $\sqrt{\underline{\lambda}_1}$ , where  $\underline{\lambda}_1$  is the largest eigenvalue of  $\Sigma_{\text{com}}$ . Given this envelope representation of  $\beta$ , we can write the conditional mean function of the many-predictors regression in this case as  $(X_t - \mu_X)' \beta = (X_t - \mu_X)' L_1 \varphi / \sqrt{\underline{\lambda}_1}$ , where, to simplify our discussion, we have disregarded the intercept parameter  $\alpha$ . Now, let  $\ell_{1,j}$  for  $j \in \{1, \dots, K\}$  be the  $j^{\text{th}}$  column of  $L_1$  and define  $p_{jt} = (X_t - \mu_X)' \ell_{1,j}$  (for  $j = 1, \dots, K$ ) so that  $p_{1t}, \dots, p_{Kt}$  are approximately the (leading) principal components of  $X_t$  after we have recentered  $X_t$  by subtracting out its mean vector  $\mu_X$ .<sup>7</sup> It follows that we can further write

$$(X_t - \mu_X)' \beta = \sum_{j=1}^K \frac{(X_t - \mu_X)' \ell_{1,j}}{\sqrt{\underline{\lambda}_1}} \varphi_j = \sum_{j=1}^K \left( \frac{p_{jt}}{\sqrt{\underline{\lambda}_1}} \right) \varphi_j, \quad (13)$$

where  $\sum_{j=1}^K (p_{jt}/\sqrt{\underline{\lambda}_1}) \varphi_j$  can be interpreted as the conditional mean function of an appro-

---

<sup>7</sup>We say that  $p_1, \dots, p_K$  are approximately the principal components of  $X_t$  because they are constructed using the eigenvectors of the common component covariance matrix  $\Sigma_{\text{com}}$ , and not of  $\Sigma_X$ . Asymptotically, however, they will be the  $K$  leading principal components under the usual assumption that the order of magnitude of the  $K$  leading eigenvalues of  $\Sigma_{\text{com}}$  is a lot larger than that of the largest eigenvalue of the error covariance matrix  $\Sigma_u$ .

proportionately normalized principal component (PC) regression. Expression (13), thus, suggests a connection between the conditional mean function of the many-predictors regression and that of a PC regression in the case where the former has a certain envelope structure. Moreover, expression (13) also shows that, when viewed from the perspective of a PC regression,  $\varphi = (\varphi_1, \dots, \varphi_K)'$ , the vector of linear combination weights in the envelope representation given in expression (12), can be reinterpreted as the vector of coefficients associated with the principal component regressors.

The preceding discussion suggests that there is a connection between the many-predictors regression with an envelope structure and a PC-type regression. Moreover, given that PC regression is obviously closely related to the diffusion index model, this further suggests that there may also exist a connection between the many-predictors regression with an envelope structure and the diffusion index model, even though the two models may look very different at first glance. Indeed, as we will show formally in Theorem 1 below, the conditional mean function of the many-predictors model turns out to be asymptotically equivalent to that of the diffusion index forecast equation, when  $X_t$  is generated by the factor model given in expression (5) and when  $\beta$  has a certain local-to-zero envelope structure.

**Theorem 1:** Suppose that the following assumptions are satisfied. (i) Let the vector of predictors  $X_t$  be generated by the factor model (5) satisfying the conditions given by expressions (6) and (7). In addition, let  $E[u_t] = 0$  for all  $t$  and let  $\Sigma_{u,t} = E[u_t u_t']$ ; and suppose that there exists some constant  $C^* > 0$ , such that  $\sup_t \lambda_{\max}(\Sigma_{u,t}) \leq C^* < \infty$ . (ii)  $\underline{\lambda}_1 \rightarrow \infty$ , as  $N, T \rightarrow \infty$ , and there exists some constant  $\bar{c} > 1$ , such that  $0 < (1/\bar{c}) \leq \min_{1 \leq g \leq K} |\varphi_{\cdot g}| \leq \max_{1 \leq g \leq K} |\varphi_{\cdot g}| \leq \bar{c} < \infty$ , where  $\varphi_{\cdot g}$  denotes the  $g^{th}$  component of the  $K \times 1$  coefficient vector  $\varphi$ . (iii) Let  $\beta = \beta_{\text{env}} / \sqrt{\underline{\lambda}_1}$ , where  $\beta_{\text{env}} = L_1 \varphi$ , and where  $L_1 = \Gamma / \sqrt{\underline{\lambda}_1}$ . Then,

$$(X_T - \mu_X)' \beta = (F_T - \mu_F)' \varphi + O_p \left( \frac{1}{\sqrt{\underline{\lambda}_1}} \right) = (F_T - \mu_F)' \varphi + O_p \left( \frac{1}{\sqrt{\lambda_1}} \right), \quad (14)$$

where  $\lambda_1 = \lambda_{\max}(\bar{\Sigma}_X)$  and  $\underline{\lambda}_1 = \lambda_{\max}(\Gamma \bar{\Sigma}_F \Gamma')$ .

**Remark 3.1:** Theorem 1 shows that the conditional mean function of the diffusion index model is well approximated by the conditional mean function of a many-predictors regression if the high-dimensional coefficient vector  $\beta$  of the latter has an envelope structure and a local-to-zero parameterization given by  $\beta = \beta_{\text{env}} / \sqrt{\underline{\lambda}_1}$ . Namely,  $\beta$  is modeled to be in a  $1/\sqrt{\underline{\lambda}_1}$  neighborhood of zero. When  $\underline{\lambda}_1 \asymp N$ , i.e. in the conventional case where factors are assumed to be pervasive, this local-to-zero parameterization corresponds to the usual Pitman–drift-type shrinkage which decays at the rate  $1/\sqrt{N}$ . However, the local-to-zero parameterization that we specify here allows for greater generality, as it also allows for the possibility that the factors may be weak, provided that the factors are not so weak that  $\underline{\lambda}_1$  does not diverge at all, i.e.  $\underline{\lambda}_1 = O(1)$ .

The result given in Theorem 1, thus, helps to explain why a PLS-type procedure can be effective in recovering the conditional mean function of a diffusion index forecast equation, even though it would seem more suited for estimating a many-predictors regression given

that it is a least-squares type procedure. In Section 6, we give a formal result as well as some additional discussion on why PLS-type procedures can consistently estimate the conditional mean function of a diffusion index forecast equation<sup>8</sup>.

## 4 The Approximate Envelope Model

We now introduce what we refer to as the approximate envelope model. To proceed, consider the following many-predictors forecast model

$$\begin{matrix} y_{t+h} \\ 1 \times 1 \end{matrix} = \alpha + (X_t - \mu_X)' \beta + \varepsilon_{t+h}, \quad (15)$$

$$\begin{matrix} X_t \\ N \times 1 \end{matrix} = \mu_X + W_t + u_t, \quad (16)$$

where  $t \in \mathbb{Z}$  and where  $h$ , the forecast horizon, is a fixed (finite) integer such that  $h \geq 1$ . Here,  $W_t$  is not assumed to be observed, and in the special case that  $X_t$  is generated by a factor model, we take  $W_t = \Gamma(F_t - \mu_F)$ , where  $F_t$  is the latent factor vector,  $\Gamma$  is the factor loading matrix, and  $\mu_F = E[F_t]$ . The following assumptions are used in the sequel.

**Assumption 4-1:**  $E[W_t] = 0$ ,  $E[u_t] = 0$ , and  $E[W_t u_t'] = 0$ , for all  $t$ . In addition, there exists a constant  $\bar{C} > 0$ , such that  $\sup_t \lambda_{\max}(\Sigma_{u,t}) \leq \bar{C} < \infty$ , where  $\Sigma_{u,t} = E[u_t u_t']$ .

**Assumption 4-2:** Let  $\bar{\Sigma}_X = T_h^{-1} \sum_{t=1}^{T_h} \Sigma_{X,t} = T_h^{-1} \sum_{t=1}^{T_h} E[(X_t - \mu_X)(X_t - \mu_X)']$ , and let  $\bar{\Sigma}_X$  have the approximate envelope decomposition

$$\bar{\Sigma}_X = \bar{\Sigma}_W + \bar{\Sigma}_u = L_1 \underline{\Delta}_1 L_1' + L_2 \underline{\Delta}_2 L_2' + \bar{\Sigma}_u, \quad (17)$$

where  $\bar{\Sigma}_W = T_h^{-1} \sum_{t=1}^{T_h} \Sigma_{W,t} = T_h^{-1} \sum_{t=1}^{T_h} E[W_t W_t']$  and  $\bar{\Sigma}_u = T_h^{-1} \sum_{t=1}^{T_h} \Sigma_{u,t} = T_h^{-1} \sum_{t=1}^{T_h} E[u_t u_t']$ , with  $T_h = T - h$ . Also, similar to the notation defined for the exact envelope model given in Section 2,  $L_1$  denotes an  $N \times \kappa$  semi-orthogonal matrix whose columns are those orthonormal eigenvectors of  $\bar{\Sigma}_W$  which span  $\mathcal{E}_{\bar{\Sigma}_W}(\mathcal{B})$ ;  $L_2$  denotes an  $N \times (N - \kappa)$  semi-orthogonal matrix whose columns are those orthonormal eigenvectors of  $\bar{\Sigma}_W$  which span the orthogonal complement of  $\mathcal{E}_{\bar{\Sigma}_W}(\mathcal{B})$ ; and  $\mathcal{E}_{\bar{\Sigma}_W}(\mathcal{B})$  denotes the  $\bar{\Sigma}_W$ -envelope of  $\mathcal{B}$ , defined as the intersection of all reducing subspaces of  $\bar{\Sigma}_W$  that contain  $\mathcal{B} = \text{span}(\beta)$ . In addition, let  $\underline{\Delta}_1 = \text{diag}(\lambda_{1,1}, \dots, \lambda_{1,\kappa})$  and  $\underline{\Delta}_2 = \text{diag}(\lambda_{2,1}, \dots, \lambda_{2,N-\kappa})$ , where the diagonal elements of  $\underline{\Delta}_1$  and  $\underline{\Delta}_2$  are the  $N$  eigenvalues of  $\bar{\Sigma}_W$  divided into two groups, with  $\lambda_{1,1}, \dots, \lambda_{1,\kappa}$  being the eigenvalues corresponding to those eigenvectors which form the columns of  $L_1$ , and with  $\lambda_{2,1}, \dots, \lambda_{2,N-\kappa}$  being the eigenvalues corresponding to the eigenvectors which form the columns of  $L_2$ . Suppose that these eigenvalues satisfy the following conditions. (a) For the first group of eigenvalues  $\lambda_{1,1}, \dots, \lambda_{1,\kappa}$ , suppose that

$$\lambda_{1,i} \asymp \lambda_{1,j} \text{ for all } i, j \in \{1, \dots, \kappa\}, \quad (18)$$

---

<sup>8</sup>For further discussion, refer to Theorem 7 and the discussion which precedes it.

where  $\kappa$  is a fixed positive integer. Suppose also that  $\underline{\lambda}_{1,j} \rightarrow \infty$ , as  $N, T \rightarrow \infty$  (for  $j = 1, \dots, \kappa$ ), and that there exists a constant  $0 < \underline{c} < 1$ , such that:

$$\frac{\underline{\lambda}_{1,i} - \underline{\lambda}_{1,j}}{\underline{\lambda}_{1,1}} \geq \underline{c}, \text{ for all } i < j \text{ with } i \in \{1, \dots, \kappa - 1\} \text{ and } j \in \{2, \dots, \kappa\}, \quad (19)$$

for all  $N$  and  $T$  sufficiently large. In the case where  $\kappa = 1$ , this assumption reduces to  $\underline{\lambda}_{1,1} \rightarrow \infty$ , as  $N, T \rightarrow \infty$ . (b) Let  $\varrho$  be a fixed non-negative integer. For the case where  $\varrho \geq 1$ , the eigenvalues  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,N-\kappa}$  are assumed to be divided into two groups,  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,\varrho}$  and  $\underline{\lambda}_{2,\varrho+1}, \dots, \underline{\lambda}_{2,N-\kappa}$ , with the second group having the local-to-zero parameterization

$$\underline{\lambda}_{2,j} = \frac{c_j}{m(N)} \text{ for } j = \varrho + 1, \dots, N - \kappa. \quad (20)$$

for  $c_j \geq 0$ . Assume that  $m(N) \rightarrow \infty$ , as  $N \rightarrow \infty$ , and that there exists an absolute constant,  $C_1 > 0$ , such that:

$$\frac{N}{\underline{\lambda}_{1,1}} \sup_{\varrho+1 \leq j \leq N-\kappa} \underline{\lambda}_{2,j} = \frac{N}{\underline{\lambda}_{1,1} m(N)} \sup_{\varrho+1 \leq j \leq N-\kappa} c_j \leq C_1 < \infty, \text{ for all } N \text{ and } T \text{ sufficiently large.} \quad (21)$$

In addition, there exists an absolute constant,  $C_2 > 0$ , such that:

$$\sum_{j=1}^{\varrho} \frac{\underline{\lambda}_{2,j}}{\underline{\lambda}_{1,1}} \leq C_2 < \infty, \text{ for all } N \text{ and } T \text{ sufficiently large.} \quad (22)$$

For the special case where  $\varrho = 0$ , the eigenvalues  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,N-\kappa}$  are all assumed to satisfy the local-to-zero parameterization specified in expressions (20) and (21).

**Assumption 4-3:** Let  $\gamma$ ,  $\gamma_1$ , and  $\gamma_2$  be positive constants satisfying the relationship  $\frac{1}{\gamma} = \frac{1}{\gamma_1} + \frac{1}{\gamma_2}$  or  $\gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}$ , such that  $0 < \gamma < 1$ . Suppose that the following conditions hold.

(a) There exists a constant  $b_1 > 0$ , such that for any  $s > 0$ , for any  $\delta \in \mathbb{R}^N$  with  $\|\delta\|_2 = 1$  and for all  $i$ ,

$$\begin{aligned} \sup_{t \in \mathbb{Z}} P(|W_{it}| > s) &\leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \\ \sup_{t \in \mathbb{Z}} P(|u'_t \delta| > s) &\leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \\ \sup_{t \in \mathbb{Z}} P(|\varepsilon_t| > s) &\leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \end{aligned}$$

where  $\bar{C}$  is a positive constant such that  $\bar{C} \geq \exp\{1\}$ . (b) Let  $V_t = (\varepsilon_t, W'_t, u'_t)'$ . Define  $\mathcal{F}_{-\infty}^t = \sigma(\dots, V_{t-2}, V_{t-1}, V_t)$  and  $\mathcal{F}_{t+m}^\infty = \sigma(V_{t+m}, V_{t+m+1}, V_{t+m+2}, \dots)$ . Suppose that there

exist positive constants  $\gamma_2$ ,  $a$ , and  $c$ , such that:

$$\alpha(m) = \sup_t \alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty) \leq a \exp(-cm^{\gamma_2}), \text{ for all } m \geq 1,$$

where  $\alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty)$  denotes the strong (or  $\alpha$ -) mixing coefficient given by

$$\alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty) = \sup_{G \in \mathcal{F}_{-\infty}^t, H \in \mathcal{F}_{t+m}^\infty} |P(G \cap H) - P(G)P(H)|.$$

(c)  $E[\varepsilon_t] = 0$  for all  $t \in \mathbb{Z}$ . In addition, let  $h$  be the finite positive integer previously defined, and assume that  $E[W_t \varepsilon_{t+h}] = 0$  for all  $t$  and  $E[u_{it} \varepsilon_{t+h}] = 0$  for all  $i$  and  $t$ .

**Assumption 4-4:** (a) Let:

$$\beta = \frac{\beta_{\text{env}}}{\sqrt{\underline{\Delta}_{1,1}}}, \text{ where } \beta_{\text{env}} = L_1 \phi \text{ and where } \phi \text{ is a } \kappa \times 1 \text{ parameter vector,} \quad (23)$$

and suppose that there exists a positive constant  $\bar{c} > 1$  such that  $0 < \frac{1}{\bar{c}} \leq \min_{1 \leq g \leq \kappa} |\phi_g| \leq \max_{1 \leq g \leq \kappa} |\phi_g| \leq \bar{c} < \infty$ . (b) Suppose that, for all  $s > 0$ ,

$$\sup_{t \in \mathbb{Z}} P(|W'_t \beta| > s) \leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}$$

where  $\beta$  is as defined in expression (23) above and where  $b_1$ ,  $\gamma_1$ , and  $\bar{C}$  are the constants defined in part (a) of Assumption 4-3.

**Remark 4.1:** (a) Note that it is without loss of generality that we take  $E[W_t] = 0$  in Assumption 4-1. This is so because if the original model takes the form  $X_t = \mu + \underline{W}_t + u_t$ , where  $E[\underline{W}_t] = \mu_{\underline{W}} \neq 0$ , we can always rewrite this equation as  $X_t = (\mu + \mu_{\underline{W}}) + (\underline{W}_t - \mu_{\underline{W}}) + u_t = \mu_X + W_t + u_t$ , where  $\mu_X = \mu + \mu_{\underline{W}}$  and  $W_t = \underline{W}_t - \mu_{\underline{W}}$ . In this paper, we are not interested in identifying and estimating  $\underline{W}_t$ , so that no additional identifying restriction is needed. As we will show later, even in the factor case where  $W_t = \Gamma(F_t - \mu_F)$ ; under a PLS-type procedure it is not necessary for us to estimate  $F_t$  or  $W_t$  if our sole purpose is to consistently estimate the conditional mean function of the forecast equation (15) in order to obtain a point forecast for  $y_{t+h}$ .

(b) Comparing expression (17) in Assumption 4-2 with expression (2) of Section 2, we see that there are two main differences between the approximate envelope decomposition considered here and the exact envelope decomposition given in expression (2). First, note that, here, we allow for the possibility that  $\Sigma_{X,t}$ , the covariance matrix of  $X_t$ , may exhibit heterogeneity across  $t$ , and for this reason we have specified the envelope relationship in terms of time averages of  $\Sigma_{X,t}$ , for  $t = 1, \dots, T_h$ . Hence,  $L_1$ ,  $L_2$ , and the eigenvalues will all depend on both  $N$  and  $T$ , in general. To keep our notation from being overly complicated, we have suppressed these dependencies. Second, expression (17) does not exactly satisfy an envelope relationship of the form  $\bar{\Sigma}_X = L_1 \underline{\Delta}_1 L_1' + L_2 \underline{\Delta}_2 L_2'$ , where the columns of  $L_1$  span the smallest reducing space that contains the span of  $\beta$ . Rather, the relationship is “perturbed”

by an additional matrix,  $\bar{\Sigma}_u$ , and it is in this sense that we think of our setup as having an approximate (or, alternatively, a perturbed) envelope structure. As discussed in Section 2, an important reason why we allow for the additional perturbation term is to accommodate situations such as the case where  $X_t$  is generated by an approximate factor model whose error covariance matrix may be non-diagonal due to the presence of cross-sectional dependence<sup>9</sup>. In that case, we can take  $\Sigma_{u,t} = E[u_t u_t']$  and  $\bar{\Sigma}_u = T_h^{-1} \sum_{t=1}^{T_h} \Sigma_{u,t}$ , where  $u_t$  is the error or idiosyncratic component of the factor model and where  $\bar{\Sigma}_u$  may have nonzero off-diagonal elements. Moreover, within a factor model setup,  $\bar{\Sigma}_W$  can be interpreted as the part of  $\bar{\Sigma}_X$  that is associated with the common component which, in turn, is further decomposed into the part that contains information about  $\beta$  and the part that does not.

(c) Observe that in Assumption 4-2(a) we require all of the eigenvalues of  $\bar{\Sigma}_W$  in the first group, i.e.  $\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}$ , to diverge to infinity, but not necessarily at the rate  $N$ . It should be noted that if  $W_t = \Gamma(F_t - \mu_F)$ , then having the eigenvalues  $\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}$  diverge at the rate  $N$  corresponds to the case where the conventional assumption of factor pervasiveness holds. Hence, in the special case where  $X_t$  is generated by a factor model, Assumption 4-2(a) allows for the possibility that factors are weaker than that assumed under the condition of factor pervasiveness. However, we do not consider the case where  $\underline{\lambda}_{1,j} = O(1)$ , for  $j = 1, \dots, \kappa$ , since in this case the signals will be so weak that consistent estimation of the conditional mean function of the forecast equation will not be possible.

(d) Note that Assumption 4-2(b) allows the largest  $\varrho$  eigenvalues in the second group, i.e.  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,\varrho}$ , to either diverge at the same rate as the eigenvalues of the first group, i.e.  $\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}$ , or at a slower rate, as long as the condition given in expression (22) is satisfied. Note also that Assumption 4-2(b) does not require all eigenvalues in the set  $\{\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,\varrho}\}$  to diverge at the same rate so it is general enough to accommodate situations where some of the eigenvalues in the set  $\{\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,\varrho}\}$  may diverge at the same rate as the eigenvalues in the first group, whereas others may diverge at a slower rate.

(e) Expression (19) of Assumption 4-2(a) implies that the eigenvalues in the first group satisfy the condition

$$1 = \frac{\underline{\lambda}_{1,1}}{\underline{\lambda}_{1,1}} > \frac{\underline{\lambda}_{1,2}}{\underline{\lambda}_{1,1}} > \dots > \frac{\underline{\lambda}_{1,\kappa}}{\underline{\lambda}_{1,1}}, \text{ for all } N \text{ and } T \text{ sufficiently large,}$$

so that although the first group of eigenvalues, i.e.  $\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}$ , all have the same order of magnitude, they are all distinct in the sense that when standardized by the largest of them, i.e.  $\underline{\lambda}_{1,1}$ , they are not equal to each other - not even asymptotically. However, note that Assumption 4-2 does not actually require  $\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}$  to all be eigenvalues with algebraic multiplicity equal to 1. Rather, even if a particular member of the set  $\{\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}\}$ , say  $\underline{\lambda}_{1,1}$ , has algebraic multiplicity equal to  $d$ , for integer  $d \geq 2$ , a decomposition of the form given by expression (17) can still be obtained by allowing  $\underline{\lambda}_{1,1}$  to appear only once along the diagonal of the matrix  $\underline{\Delta}_1$ , but appear  $d - 1$  times along the diagonal of the matrix  $\underline{\Delta}_2$

---

<sup>9</sup>Some related discussion of this issue is given in Remark A.1 of Appendix A of the Online Appendix.

(noting that Assumption 4-2 does not actually require  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,\varrho}$  to have values different from  $\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}$ , nor does it require  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,\varrho}$  to all have distinct values). In fact, what we have just described is also the correct envelope decomposition for this case, since, as illustrated in Example 1 given in the Online Appendix, if one of the top eigenvalues, say  $\underline{\lambda}_{1,1}$ , has algebraic multiplicity greater than 1, then associated with it will be multiple eigenvectors. However, under an appropriate change of basis, only one of these eigenvectors will actually be part of a basis that spans the eigen-subspace containing  $\beta$ , while all other eigenvectors will be orthogonal to the eigen-subspace containing  $\beta$  and thus will not carry meaningful information for the prediction of  $y_{t+h}$ . Hence, if one or more of the leading eigenvalues have algebraic multiplicity greater than 1, then the dimension of the envelope can actually be less than the total number of eigenvectors associated with the leading eigenvalues, resulting in the opportunity to achieve further dimension reduction under the envelope setup<sup>10</sup>.

(f) Under Assumption 4-2(b), the diagonal elements of  $\underline{\Delta}_2$  also include the eigenvalues  $\underline{\lambda}_{2,\varrho+1}, \dots, \underline{\lambda}_{2,N-\kappa}$ , which have a local-to-zero structure satisfying the conditions given by (20) and (21). When specialized to the case where  $X_t$  is generated by a factor model, the eigen-structure specified in Assumption 4-2 is more than general enough to cover most factor model specifications that have been presented in the econometrics literature, since the covariance matrix of the common component of the factor model is typically assumed to have a finite number of eigenvalues which might diverge as sample size increases, and the remaining eigenvalues are assumed to be zero rather than local-to-zero.

(g) Part (a) of Assumption 4-3 requires that  $\{W_t\}$  and  $\{\varepsilon_t\}$  have the kind of exponential-type tail behavior commonly assumed in the econometrics and statistics literature on estimation and inference for high-dimensional models. This assumption allows us to apply a version of the Bernstein inequality for weakly dependent processes that has been obtained by Merlevède, Peligrad, and Rio (2011). In addition, this assumption is in accord with what people have found empirically about the tail behavior of many economic time series. For example, the study by Fagiolo, Napoletano, and Roventini (2008) shows that the distributions of aggregate output growth rate time series of a majority of OECD countries are well-approximated by symmetric exponential-power densities, with tail behavior corresponding to our assumption.

(h) Part (a) of Assumption 4-4 allows the approximate envelope coefficient vector  $\beta$  to have a local-to-zero parameterization. As discussed previously, specifying  $\beta$  in this way is motivated by the result of Theorem 1 above. To motivate part (b) of Assumption 4-4, i.e., the condition that there exist positive constants  $b_1$ ,  $\gamma_1$ , and  $\overline{C}$  such that, for all  $s > 0$ ,

$$\sup_{t \in \mathbb{Z}} P(|W_t' \beta| > s) \leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \overline{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\};$$

it is useful to consider the special case where the vector of predictor variables  $X_t$  is generated

---

<sup>10</sup>See Remark A.2 and Example 1, given at the end of Appendix A of the Online Appendix for further discussion and technical details about this case. The Online Appendix can be viewed at the authors' websites.

by a factor model so that  $W_t = \Gamma (F_t - \mu_F)$ . In this case,

$$W_t' \beta = (F_t - \mu_F)' \Gamma' \frac{\Gamma}{\sqrt{\Delta_{1,1}}} \frac{\varphi}{\sqrt{\Delta_{1,1}}} = (F_t - \mu_F)' \frac{\Gamma' \Gamma}{\Delta_{1,1}} \varphi = (F_t - \mu_F)' \varphi,$$

and it is easy to see that part (b) of Assumption 4-4 will hold given the condition

$$\sup_{t \in \mathbb{Z}} P(|F_{t,g} - \mu_{F,g}| > s) \leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \overline{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \quad (24)$$

for  $g \in \{1, \dots, K\}$ , as specified in Section 6 when we analyze the particular case of a diffusion index model. See part (a) of Assumption (v) in that section. In fact, we show explicitly in Lemma E-7 part (a) for the case where  $\kappa = K$  and in Lemma E-6 part (a) for the case where  $1 \leq \kappa < K$  that, given expression (24), there exist positive constants  $b_2$  and  $\gamma_1$  not depending on  $t$  such that, for all  $t \in \mathbb{Z}^{11}$ ,

$$P(|(F_t - \mu_F)' \varphi| > s) \leq \exp \left\{ 1 - \left( \frac{s}{b_2} \right)^{\gamma_1} \right\} \leq \overline{C} \exp \left\{ - \left( \frac{s}{b_2} \right)^{\gamma_1} \right\}.$$

Part (b) of Assumption 4-4. thus, provides a more general condition which can be explicitly verified to hold under the factor model case.

(i) Although part (b) of Assumption 4-3 above allows  $\{\varepsilon_t\}$  to be a dependent process that satisfies a mixing condition, it should be noted that the perhaps more common situation where  $\{\varepsilon_t\}$  is an independent random sequence is clearly a special case of our more general assumption. The results we will present are general enough so that they hold regardless of whether  $\{\varepsilon_t\}$  is a dependent or an independent process.

## 5 Connections between the Approximate Envelope Model and Partial Least Squares

In this section, we motivate why PLS-type procedures are natural candidates for estimating the envelope model and its variants. Although, as mentioned previously, the PLS procedure has often been described in terms of iterated algorithms, papers by Helland (1988, 1990) develop a closed-form representation for the estimator. To simplify our discussion, we initially consider the case where the underlying model is covariance stationary since this is in accord with what has been assumed in much of the PLS literature. In the case where there is no heterogeneity in the second moments, Helland (1990) shows that the population version of the PLS estimator is:

$$\overline{\beta} = G (G' \Sigma_X G)^{-1} G' \sigma_{Xy}, \quad (25)$$

---

<sup>11</sup>Lemmas E-6 and E-7 and their proofs are both given in Appendix E of the Online Appendix.

where  $G = \begin{pmatrix} \sigma_{Xy} & \Sigma_X \sigma_{Xy} & \cdots & \Sigma_X^{\kappa-1} \sigma_{Xy} \end{pmatrix}$ , and  $\Sigma_X$  and  $\sigma_{Xy}$  do not depend on  $t$ . It is easy to see why expression (25) is the population analogue of the PLS estimator. Namely, consider the special case of equation (15) where  $\mu_X = 0$  and  $\alpha = 0$  and plug sample analogues for  $\Sigma_X$  and  $\sigma_{Xy}$  into (25). This yields the sample PLS estimator:

$$\begin{aligned} \hat{\beta} &= \hat{G} \left( \hat{G}' \hat{\Sigma}_X \hat{G} \right)^{-1} \hat{G}' \hat{\sigma}_{Xy} = V_\kappa D_{\kappa,T}^{-1} \left( D_{\kappa,T}^{-1} V_\kappa' \frac{X'X}{T} V_\kappa D_{\kappa,T}^{-1} \right)^{-1} D_{\kappa,T}^{-1} V_\kappa' \frac{X'y}{T} \\ &= V_\kappa (V_\kappa' X'X V_\kappa)^{-1} V_\kappa' X'y, \end{aligned}$$

where  $\hat{G} = \begin{pmatrix} \hat{\sigma}_{Xy} & \hat{\Sigma}_X \hat{\sigma}_{Xy} & \cdots & \hat{\Sigma}_X^{\kappa-1} \hat{\sigma}_{Xy} \end{pmatrix}$ ,  $\hat{\Sigma}_X = X'X/T$ ,  $\hat{\sigma}_{Xy} = X'y/T$ , and  $\frac{X}{T \times N} = \begin{pmatrix} X_1 & X_2 & \cdots & X_T \end{pmatrix}'$  and where  $V_\kappa = \begin{pmatrix} X'y & (X'X)X'y & \cdots & (X'X)^{\kappa-1}X'y \end{pmatrix}$  and  $D_{\kappa,T} = \text{diag}(T, \dots, T^\kappa)$ .<sup>12</sup> Since we allow for the possibility of heterogeneity in the second moments, we consider a generalization of the representation for the population PLS estimator given in expression (25). More precisely, our generalized formula for the population PLS estimator takes the form:

$$\bar{\beta} = G (G' \bar{\Sigma}_X G)^{-1} G' \bar{\sigma}_{Xy}, \quad (26)$$

where  $G = \begin{pmatrix} \bar{\sigma}_{Xy} & \bar{\Sigma}_X \bar{\sigma}_{Xy} & \cdots & \bar{\Sigma}_X^{\kappa-1} \bar{\sigma}_{Xy} \end{pmatrix}$  and where

$$\bar{\Sigma}_X = T_h^{-1} \sum_{t=1}^{T_h} E[(X_t - \mu_X)(X_t - \mu_X)'] \text{ and } \bar{\sigma}_{Xy} = T_h^{-1} \sum_{t=1}^{T_h} E\{(X_t - \mu_X)(y_{t+h} - \alpha)\}.$$

The following theorem characterizes the relationship between  $\beta$ , the envelope model coefficient vector, and  $\bar{\beta}$ , the population version of the PLS estimator, under both the exact envelope model and the approximate envelope model.

**Theorem 2:** Let  $\bar{\beta}$  and  $\beta$  be defined by expressions (26) and (23), respectively. Consider the model given by expressions (15) and (16), and suppose that Assumptions 4-1, 4-2, 4-3(c), and 4-4(a) hold. Then, the following statements are true.

- (a) For the exact envelope model with  $\bar{\Sigma}_u = 0$ ,  $\bar{\beta} = G (G' \bar{\Sigma}_X G)^{-1} G' \bar{\sigma}_{Xy} = L_1 \phi / \sqrt{\lambda_{1,1}} = \beta$ .
- (b) For the approximate envelope model where  $\bar{\Sigma}_u$  is a general symmetric and positive semi-definite matrix satisfying Assumption 4-1,  $\|\bar{\beta} - \beta\|_2 = O\left(1/\lambda_1^{\frac{3}{2}}\right)$ , where  $\lambda_1 = \lambda_{\max}(\bar{\Sigma}_X)$ .

**Remark 5.1:** Note that part (a) of Theorem 2 shows that the population version of the PLS estimator is exactly equal to the envelope coefficient vector when the envelope model is exact. A similar result has previously been obtained by Cook, Helland, and Su (2013), but for the case where data is assumed to be *i.i.d.*. Our result allows for heterogeneity in the second moments across  $t$ . Part (b) of Theorem 2 is a novel result which shows that  $\bar{\beta}$  converges to  $\beta$  in the  $\ell_2$  norm, even under an approximate envelope model. An obvious implication of this result is that any consistent estimator of the population analogue  $\bar{\beta}$  will also be a consistent estimator of the envelope parameter vector  $\beta$ , as long as  $\lambda_1 \rightarrow \infty$ , which will be true if  $\lambda_{1,1} \rightarrow \infty$ , as can be seen by applying the result given in part (a) of Lemma

<sup>12</sup>Note that Carrasco and Rossi (2016), Groen and Kapetanios (2016), and Ahn and Bae (2022) give a similar representation for the sample PLS estimator.

D-3 given in Appendix D of the Online Appendix.

Our next two results provide some further development of the connection between diffusion index models and many-predictors regressions with an envelope structure. Before discussing these results, note first that, in the context of a diffusion index model, the envelope dimension  $\kappa$  may be smaller than the total number of factors  $K$  if not all the available factors are relevant for forecasting the target variable of interest  $y_{t+h}$ . More precisely, in this case, ignoring the intercept term, we can write the conditional mean function of the diffusion index forecast equation given in expression (4) as

$$\begin{aligned} (F_t - \mu_F)' \varphi &= (F_{1,t} - \mu_{F,1})' \varphi_1 + (F_{2,t} - \mu_{F,2})' \varphi_2 \\ &= (F_{1,t} - \mu_{F,1})' \varphi_1, \end{aligned} \quad (27)$$

where  $F_{1,t}$  is a  $\kappa \times 1$  vector of factors whose components are useful for the purpose of forecasting  $y_{t+h}$ , whereas the components of the  $(K - \kappa) \times 1$  vector  $F_{2,t}$  are those factors which are not useful for forecasting because its vector of coefficients  $\varphi_2 = 0$ . Moreover, the factor equation (5) in this  $1 \leq \kappa < K$  case can be rewritten in the partitioned form

$$\begin{aligned} X_t &= \mu_X + \Gamma (F_t - \mu_F) + u_t \\ &= \mu_X + \Gamma_1 (F_{1,t} - \mu_{F,1}) + \Gamma_2 (F_{2,t} - \mu_{F,2}) + u_t, \end{aligned} \quad (28)$$

where the columns of the  $N \times \kappa$  matrix  $\Gamma_1$  are those (un-normalized) eigenvectors which contain useful information for forecasting  $y_{t+h}$  since sample analogues of  $\Gamma_1$  can be used to construct principal component estimates of those factors which have nonzero coefficients in the forecast equation. On the other hand, the columns of the  $N \times (K - \kappa)$  matrix  $\Gamma_2$  are those (un-normalized) eigenvectors which are irrelevant for the purpose of forecasting since sample analogues of  $\Gamma_2$  are only useful for estimating those factors which have zero coefficients in the forecast equation<sup>13</sup>.

Next, recall that, in Section 3, we specify the set of nonzero eigenvalues of the common component covariance matrix  $\Gamma \bar{\Sigma}_F \Gamma'$  to be the set  $\{\lambda_1, \dots, \lambda_K\}$ . We now split this set of eigenvalues into two groups, where  $\{\lambda_{1,1}, \dots, \lambda_{1,\kappa}\}$  is the set of eigenvalues associated with the columns of  $\Gamma_1$  and  $\{\lambda_{2,1}, \dots, \lambda_{2,K-\kappa}\}$  is the set of eigenvalues associated with the columns of  $\Gamma_2$  and where  $\{\lambda_{1,1}, \dots, \lambda_{1,\kappa}\} \cup \{\lambda_{2,1}, \dots, \lambda_{2,K-\kappa}\} = \{\lambda_1, \dots, \lambda_K\}$ . The reason for this slight change of notation is that, here, we do not require  $\lambda_{1,1}, \dots, \lambda_{1,\kappa}$  to be the same as  $\lambda_1, \dots, \lambda_\kappa$  so that  $\lambda_{1,1}, \dots, \lambda_{1,\kappa}$  need not be the  $\kappa$  largest elements of the original set  $\{\lambda_1, \dots, \lambda_K\}$ . We

---

<sup>13</sup>It is without loss of generality that in the representation given by expressions (27) and (28) above that we seem to be taking the relevant factors to be the first  $\kappa$  components of  $F_t$ . This can always be achieved by reordering the components of  $F_t$  via transformation by an appropriate permutation matrix. We provide additional details to show that this is true in Appendix E of the Online Appendix. Please see the discussion and calculations given in Remark E-17 which can be found at the end of Appendix E following the proof of Lemma E-16.

also write in this case

$$\underline{\Lambda}_{K \times K} = \begin{pmatrix} \underline{\Lambda}_1 & 0 \\ 0 & \underline{\Lambda}_2 \end{pmatrix},$$

where  $\underline{\Lambda}_1 = \text{diag}(\lambda_{1,1}, \dots, \lambda_{1,\kappa})$  and  $\underline{\Lambda}_2 = \text{diag}(\lambda_{2,1}, \dots, \lambda_{2,K-\kappa})$ .

Interpreting the  $1 \leq \kappa < K$  case within the approximate envelope setup, it seems reasonable to try to approximate this case using an envelope that is spanned by the  $\kappa$  columns of the semi-orthogonal submatrix  $L_1 = \Gamma_1 / \sqrt{\Delta_{1,1}}$ , so that the envelope model coefficient vector  $\beta$  in this case is given by

$$\beta = L_1 \frac{\varphi_1}{\sqrt{\Delta_{1,1}}} = \left( \frac{\Gamma_1}{\sqrt{\Delta_{1,1}}} \right) \frac{\varphi_1}{\sqrt{\Delta_{1,1}}}, \quad (29)$$

which follows naturally from specializing the previous formula, i.e.,

$$\beta = \left( \frac{\Gamma}{\sqrt{\Delta_{1,1}}} \right) \frac{\varphi}{\sqrt{\Delta_{1,1}}} = \left( \frac{\Gamma_1}{\sqrt{\Delta_{1,1}}} \right) \frac{\varphi_1}{\sqrt{\Delta_{1,1}}} + \left( \frac{\Gamma_2}{\sqrt{\Delta_{1,1}}} \right) \frac{\varphi_2}{\sqrt{\Delta_{1,1}}},$$

to the case where  $\varphi_2 = 0$ . We show in Theorem 3 below that the conditional mean function of a many-predictors regression with coefficient vector  $\beta$  given by expression (29) is indeed a good asymptotic approximation of the conditional mean function of the diffusion index forecast equation in the case where  $1 \leq \kappa < K$ .

On the other hand, in the case where  $\kappa = K$ , there is obviously no need to split the nonzero eigenvalues of  $\Gamma \bar{\Sigma}_F \Gamma'$  into two groups, nor is there a need in this case to partition  $\Gamma$  and  $\underline{\Lambda}$ , as done above. However, to make our notation in the  $\kappa = K$  case as close as possible to that of the  $1 \leq \kappa < K$  case so that we can give essentially the same proof to cover both cases for many of our results, we often use the double subscript notation even in the  $\kappa = K$  case. Hence, in the  $\kappa = K$  case, we take the set of nonzero eigenvalues of  $\Gamma \bar{\Sigma}_F \Gamma'$  to be the set  $\{\lambda_{1,1}, \dots, \lambda_{1,K}\}$ , all members of which are now associated with eigenvectors that help to span the envelope. Having a first subscript set to 1, thus, can be interpreted as signifying that this is an eigenvalue associated with an eigenvector that contains useful information for forecasting the target variable of interest  $y_{t+h}$ . In a similar vain, in the  $\kappa = K$  case, set  $\underline{\Lambda} = \underline{\Lambda}_1 = \text{diag}(\lambda_{1,1}, \dots, \lambda_{1,K})$  and also  $\varphi = (\varphi_{1,1}, \dots, \varphi_{1,K})'$ , noting that under this notation for an element of the vector  $\varphi$  to have a first subscript equal to 1 means that it is nonzero.

Given this background and based on the notation introduced above, we shall consider two different scenarios when considering the diffusion index case:

**Scenario 1 ( $1 \leq \kappa < K$  Case).** Assume that the following conditions hold:

**Assumption S1-1:** There exists a constant  $\bar{c} > 1$  such that  $0 < \frac{1}{\bar{c}} \leq \min_{1 \leq g \leq \kappa} |\varphi_{1,g}| \leq \max_{1 \leq g \leq \kappa} |\varphi_{1,g}| \leq \bar{c} < \infty$ , whereas  $\varphi_2 = (\varphi_{2,1}, \dots, \varphi_{2,K-\kappa})' = 0$ .

**Assumption S1-2:** (a) Suppose that  $\lambda_{1,g} \rightarrow \infty$  and  $\lambda_{2,j} \rightarrow \infty$ , as  $N, T \rightarrow \infty$  (for  $g = 1, \dots, \kappa$  and  $j = 1, \dots, K - \kappa$ ) such that  $\lambda_{1,1} \asymp \dots \asymp \lambda_{1,\kappa} \asymp \lambda_{2,1} \asymp \dots \asymp \lambda_{2,K-\kappa}$ . (b)

Suppose also that there exists a constant,  $\underline{c}$ , with  $0 < \underline{c} < 1$ , such that  $(\underline{\lambda}_{1,g} - \underline{\lambda}_{1,k}) / \underline{\lambda}_{1,1} \geq \underline{c}$ , for all  $g < k$ , with  $g \in \{1, \dots, \kappa - 1\}$  and  $k \in \{2, \dots, \kappa\}$  and for all  $N$  and  $T$  sufficiently large.

**Scenario 2 ( $\kappa = K$  Case).** Assume that the following conditions hold:

**Assumption S2-1:** There exists a constant  $\bar{c} > 1$  such that  $0 < \frac{1}{\bar{c}} \leq \min_{1 \leq g \leq K} |\varphi_{1,g}| \leq \max_{1 \leq g \leq K} |\varphi_{1,g}| \leq \bar{c} < \infty$ .

**Assumption S2-2** (a) Suppose that  $\underline{\lambda}_{1,g} \rightarrow \infty$  as  $N, T \rightarrow \infty$  (for  $g = 1, \dots, K$ ) such that  $\underline{\lambda}_{1,1} \asymp \dots \asymp \underline{\lambda}_{1,K}$ . (b) Suppose also that there exists a constant,  $\underline{c}$ , with  $0 < \underline{c} < 1$ , such that  $(\underline{\lambda}_{1,g} - \underline{\lambda}_{1,k}) / \underline{\lambda}_{1,1} \geq \underline{c}$ , for all  $g < k$ , with  $g \in \{1, \dots, K - 1\}$  and  $k \in \{2, \dots, K\}$  and for all  $N$  and  $T$  sufficiently large.

**Remark:** Note that one difference between Assumption S1-2 and Assumption S2-2 is that similar to the condition specified in Assumption 4-2 part (a), in the  $1 \leq \kappa < K$  case, the eigenvalues in the second group (i.e.,  $\underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,K-\kappa}$ ) need not have algebraic multiplicity equal to 1. See Remark A.2 and Example 1 in Appendix A of the Online Appendix for further discussion of why the assumption is specified in this way.

In Theorem 1, we showed that the conditional mean function of the many-predictors model with an approximate envelope structure is asymptotically equivalent to the conditional mean function of the diffusion index forecast equation, when  $\kappa = K$ . We now extend this result to the case where  $1 \leq \kappa < K$  (i.e., Scenario 1).

**Theorem 3:** Let the vector of predictors  $X_t$  be generated by the factor model

$$X_t = \mu_X + \Gamma (F_t - \mu_F) + u_t. \quad (30)$$

Suppose that Assumption S1-1 and part (a) of Assumption S1-2 hold. In addition, suppose that the following conditions are satisfied: (i)  $\Gamma' \Gamma / \underline{\lambda}_{1,1} = I_K$  and  $\bar{\Sigma}_F = \underline{\Lambda} / \underline{\lambda}_{1,1}$ , where  $\bar{\Sigma}_F = T_h^{-1} \sum_{t=1}^{T_h} E[(F_t - \mu_F)(F_t - \mu_F)']$ ,  $\underline{\Lambda} = \text{diag}(\underline{\lambda}_{1,1}, \dots, \underline{\lambda}_{1,\kappa}, \underline{\lambda}_{2,1}, \dots, \underline{\lambda}_{2,K-\kappa})$ , and where for all  $t$ ,  $E[u_t] = 0$  and  $\Sigma_{u,t} = E[u_t u_t']$ , and assume that there exists some constant  $C^* > 0$ , such that  $\sup_t \lambda_{\max}(\Sigma_{u,t}) \leq C^* < \infty$ ;<sup>14</sup> (ii) let  $\beta = \beta_{\text{env}} / \sqrt{\underline{\lambda}_{1,1}}$ , where  $\beta_{\text{env}} = L_1 \varphi_1$ , with  $L_1 = \Gamma_1 / \sqrt{\underline{\lambda}_{1,1}}$ . Then,

$$(X_T - \mu_X)' \beta = (F_{1,T} - \mu_{F,1})' \varphi_1 + O_p\left(\frac{1}{\sqrt{\underline{\lambda}_{1,1}}}\right) = (F_{1,T} - \mu_{F,1})' \varphi_1 + O_p\left(\frac{1}{\sqrt{\lambda_1}}\right),$$

where  $\lambda_1 = \lambda_{\max}(\bar{\Sigma}_X)$ .

---

<sup>14</sup>Note again that the assumptions  $\Gamma' \Gamma / \underline{\lambda}_{1,1} = I_K$  and  $\bar{\Sigma}_F = \underline{\Lambda} / \underline{\lambda}_{1,1}$  are not restrictive since we can always transform the model so that these conditions will hold. This is true even though we divide by  $\underline{\lambda}_{1,1}$  which is not guaranteed under Scenario 1 to be the largest eigenvalue of the common component covariance matrix  $\Gamma \bar{\Sigma}_F \Gamma'$ . As we show in Lemma D-1 in Appendix D of the Online Appendix, it turns out that we can transform the diffusion index model based on division by any of the top  $K$  eigenvalues to obtain representations under which conditions similar to the one given above will hold, as long as the eigenvalue with which one standardizes has the same order of magnitude as the largest eigenvalue of  $\Gamma \bar{\Sigma}_F \Gamma'$ . See Lemma D-1 of the Online Appendix for details and a proof of this result.

Our next result shows that the conditional mean function of a diffusion index type forecast equation can be approximated by the conditional mean function of a many-predictors regression with the population PLS estimator  $\bar{\beta}$ , i.e. expression (26), plugged in for  $\beta$ .

**Theorem 4:** Consider the diffusion index forecast model:

$$\begin{aligned} y_{t+h} &= \alpha + (F_t - \mu_F)' \varphi + \varepsilon_{t+h}, \\ X_t &= \mu_X + \Gamma (F_t - \mu_F) + u_t, \end{aligned} \quad (31)$$

where  $\mu_F = E[F_t]$  and  $\mu_X = E[X_t]$ , and where  $t \in \mathbb{Z}$ . Assume that the following conditions hold: (i)  $\Gamma' \Gamma / \underline{\lambda}_{1,1} = I_K$  and  $\bar{\Sigma}_F = \underline{\Lambda} / \underline{\lambda}_{1,1}$ , where  $\bar{\Sigma}_F = T_h^{-1} \sum_{t=1}^{T_h} E[(F_t - \mu_F)(F_t - \mu_F)']$  and where  $\underline{\Lambda}$  is a diagonal matrix whose diagonal elements are the nonzero eigenvalues of  $\Gamma \bar{\Sigma}_F \Gamma'$ ; (ii) Suppose that  $E[u_t] = 0$  and  $E[(F_t - \mu_F)u_t] = 0$ , for all  $t$ . Also, suppose that there exists some constant  $C^\dagger > 0$  such that  $\sup_t \lambda_{\max}(\Sigma_{u,t}) \leq C^\dagger < \infty$  and  $E\{(F_{T,g} - \mu_{F,g})^4\} \leq C^\dagger < \infty$ , for all  $g \in \{1, \dots, K\}$ , where  $\Sigma_{u,t} = E[u_t u_t']$  and where  $F_{T,g}$  and  $\mu_{F,g}$  denote the  $g^{th}$  component of  $F_T$  and  $\mu_F$ , respectively; (iii)  $E[\varepsilon_t] = 0$  for all  $t$ . In addition, let the forecast horizon  $h$  be a finite positive integer, and assume that  $E[(F_t - \mu_F)\varepsilon_{t+h}] = 0$  for all  $t$  and  $E[u_{it}\varepsilon_{t+h}] = 0$  for all  $i$  and  $t$ . Finally, let  $\bar{\beta} = G(G'\bar{\Sigma}_X G)^{-1} G' \bar{\sigma}_{Xy}$  be the population PLS estimator defined in (26).

(a) Consider Scenario 1. Suppose that Assumptions S1-1 and S1-2 hold in addition to conditions (i)-(iii). Then,

$$(X_T - \mu_X)' \bar{\beta} = (F_{1,T} - \mu_{F,1})' \varphi_1 + O_p\left(\frac{1}{\sqrt{\underline{\lambda}_{1,1}}}\right) = (F_{1,T} - \mu_{F,1})' \varphi_1 + O_p\left(\frac{1}{\sqrt{\lambda_1}}\right), \quad (32)$$

where  $\underline{\lambda}_{1,1} = \lambda_{\max}(\underline{\Lambda}_1)$ ,  $\lambda_1 = \lambda_{\max}(\bar{\Sigma}_X)$  and  $F_{1,T}$  is a  $\kappa \times 1$  vector containing the relevant factors.

(b) Consider Scenario 2. Suppose that Assumptions S2-1 and S2-2 hold in addition to conditions (i)-(iii). Then,

$$(X_T - \mu_X)' \bar{\beta} = (F_T - \mu_F)' \varphi + O_p\left(\frac{1}{\sqrt{\underline{\lambda}_{1,1}}}\right) = (F_T - \mu_F)' \varphi + O_p\left(\frac{1}{\sqrt{\lambda_1}}\right), \quad (33)$$

where  $F_T$  is a  $K \times 1$  vector, since all the available factors are relevant for forecasting  $y_{t+h}$ .

**Remark 5.2:** (a) Although  $\bar{\beta}$  is not a feasible estimator, the result given in Theorem 4 is encouraging because it suggests a strategy for estimating the conditional mean function of a many-predictors regression which can also allow one to asymptotically recover the conditional mean function of the diffusion index forecast equation in the case where the latter is the true model. This can be done by replacing  $\bar{\beta}$  with a feasible PLS-type estimator which consistently estimates  $\bar{\beta}$ . For feasible estimation, in the next section we introduce regularized versions of the PLS estimator which are more adaptive to alternative statistical environments than the original unregularized PLS estimator. In addition, we show explicitly in Section 6 that

point forecasts constructed from our regularized PLS estimators can consistently estimate the conditional mean function of the true forecast equation, regardless of whether the true model takes the form of a many-predictors regression with an envelope structure or that of a diffusion index model.

(b) Importantly, note that given  $\bar{\beta} = G (G' \bar{\Sigma}_X G)^{-1} G' \bar{\sigma}_{Xy}$ , where  $G = \begin{pmatrix} \bar{\sigma}_{Xy} & \cdots & \bar{\Sigma}_X^{\kappa-1} \bar{\sigma}_{Xy} \end{pmatrix}$ , constructing a sample version of  $\bar{\beta}$  does not require us to identify and estimate the basis matrix  $L_1$ , so that we do not need to pre-test to find out which eigenvector or factor is useful for predicting  $y_{t+h}$  and which is not. In fact, to obtain a feasible version of  $\bar{\beta}$ , we only need to estimate  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$  and to correctly specify the envelope dimension parameter  $\kappa$ . We will say more about the correct specification of  $\kappa$  in the next section of the paper. See, in particular, part (c) of Remark 6.1.

## 6 Regularized PLS Estimators

As shown in Theorem 2, the population version of the PLS estimator,

$$\bar{\beta} = G (G' \bar{\Sigma}_X G)^{-1} G' \bar{\sigma}_{Xy}, \text{ where } G = \begin{pmatrix} \bar{\sigma}_{Xy} & \bar{\Sigma}_X \bar{\sigma}_{Xy} & \cdots & \bar{\Sigma}_X^{\kappa-1} \bar{\sigma}_{Xy} \end{pmatrix}, \quad (34)$$

converges asymptotically to the coefficient vector  $\beta$  of the many-predictors forecast equation (15). Given this result, a natural approach to estimating  $\beta$  is to replace  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$  in expression (34) with their sample analogues. Indeed, as discussed previously, replacing  $\bar{\Sigma}_X$  with the sample covariance matrix of  $X_t$  and replacing  $\bar{\sigma}_{Xy}$  with the sample covariance vector of  $X_t$  and  $y_{t+h}$  results in the original (unregularized) PLS estimator. Below, we introduce new versions of the PLS estimator, where  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$  are not replaced by their usual sample analogues but by certain regularized covariance estimators. The procedures that we introduce have the nice feature that they not only can take advantage of dimension reduction that is available due to the presence of a factor structure but they can also make use of any parameter sparsity that may exist in the underlying data mechanism. To this end, we estimate  $\bar{\Sigma}_X$  using the generalized thresholding estimator for high-dimensional covariance matrices proposed by Rothman, Levina, and Zhu (2009). We also estimate  $\bar{\sigma}_{Xy}$  using a generalized thresholding estimator. More precisely, let  $\hat{\Sigma}_X = T_h^{-1} \sum_{t=1}^{T_h} (X_t - \bar{X}_{T_h}) (X_t - \bar{X}_{T_h})'$  and  $\hat{\sigma}_{Xy} = T_h^{-1} \sum_{t=1}^{T_h} [X_t - \bar{X}_{T_h}] y_{t+h}$  be, respectively, the usual sample analogue of  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$ , and let  $\hat{\Sigma}_X^S$  and  $\hat{\sigma}_{Xy}^S$  denote, respectively, the generalized thresholding estimators of  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$ . Then, the  $(i, j)^{th}$  element of  $\hat{\Sigma}_X^S$  and the  $i^{th}$  element of  $\hat{\sigma}_{Xy}^S$  are defined as follows:

$$\hat{\Sigma}_{X,(i,j)}^S = s_{\eta_1} \left( \hat{\Sigma}_{X,(i,j)} \right) \text{ and } \hat{\sigma}_{Xy,i}^S = s_{\eta_2} \left( \hat{\sigma}_{Xy,i} \right), \quad (35)$$

where  $\hat{\Sigma}_{X,(i,j)}$  and  $\hat{\sigma}_{Xy,i}$  denote, respectively, the  $(i, j)^{th}$  element of  $\hat{\Sigma}_X$  and the  $i^{th}$  element of  $\hat{\sigma}_{Xy}$  and where  $\eta_1, \eta_2 \geq 0$  are the specified thresholds. Note that we allow  $\hat{\Sigma}_{X,(i,j)}^S$  and

$\hat{\sigma}_{xy,i}^S$  to possibly have different values of the threshold function. The following conditions are imposed on the function  $s_\eta(z)$ , where the more general notation  $\eta$  can denote either  $\eta_1$  or  $\eta_2$ .

**Assumption 6-1:** Let  $s_\eta(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  satisfy the following three conditions:

$$(i) \quad |s_\eta(z)| \leq |z|. \quad (36)$$

$$(ii) \quad s_\eta(z) = 0 \text{ for } |z| \leq \eta. \quad (37)$$

$$(iii) \quad |s_\eta(z) - z| \leq \eta. \quad (38)$$

As discussed in Rothman, Levina, and Zhu (2009), Assumption 6-1 allows the generalized threshold estimator to combine shrinkage with hard thresholding. In particular, condition (i) imposes shrinkage, condition (ii) allows for hard thresholding, and condition (iii) constrains the amount of shrinkage to no more than  $\eta$ . Some leading examples of estimators which satisfy these three conditions include:

**1. Hard Thresholding:**

$$s_\eta^{\text{Hard}}(z) = z \mathbb{I}\{|z| > \eta\},$$

where  $\mathbb{I}(\cdot)$  here denotes the indicator function.

**2. Soft Thresholding:**

$$s_\eta^{\text{Soft}}(z) = \text{sgn}(z) (|z| - \eta)_+,$$

where  $\text{sgn}(\cdot)$  denotes the sign function and where

$$(x)_+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}.$$

**3. Smoothly Clipped Absolute Deviation (SCAD) Proposed by Fan and Li (2001):** For  $a > 2$ ,

$$s_\eta^{\text{SCAD}}(z) = \begin{cases} \text{sgn}(z) (|z| - \eta)_+ & \text{if } |z| \leq 2\eta \\ [(a-1)|z| - \text{sgn}(z)a\eta] / (a-2) & \text{if } 2\eta < |z| \leq a\eta \\ |z| & \text{if } |z| > a\eta \end{cases}. \quad (39)$$

**4. Adaptive LASSO Proposed by Zou (2006):**

$$s_\eta^{\text{AL}}(z) = \text{sgn}(z) \left( |z| - \eta^{\zeta+1} |z|^{-\zeta} \right)_+, \text{ where } \zeta > 0. \quad (40)$$

Making use of the generalized thresholding estimator given above, we propose two classes of regularized PLS estimators, one of which we shall refer to as the Doubly Regularized PLS (or DRPLS) estimator and the other we shall call the Generalized-inverse-based Regularized PLS (or GRPLS) estimator. More precisely, let

$$\hat{G}_S = \begin{pmatrix} \hat{\sigma}_{Xy}^S & \hat{\Sigma}_X^S \hat{\sigma}_{Xy}^S & \cdots & [\hat{\Sigma}_X^S]^{\kappa-1} \hat{\sigma}_{Xy}^S \end{pmatrix} \text{ and } \hat{D}_{\lambda,\kappa} = \begin{pmatrix} \hat{\lambda}_1^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \hat{\lambda}_1^{\frac{3}{2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \hat{\lambda}_1^{\kappa-\frac{1}{2}} \end{pmatrix},$$

where  $\hat{\Sigma}_X^S$  and  $\hat{\sigma}_{Xy}^S$  are as previously defined in expression (35) and satisfy conditions (i)-(iii) of Assumption 6-1, where  $\kappa$  is the dimension of the approximate envelope model, and where  $\hat{\lambda}_1 = \|\hat{\Sigma}_X^S\|_2 = \sqrt{\lambda_{\max}(\hat{\Sigma}_X^S \hat{\Sigma}_X^S)} = \lambda_{\max}(\hat{\Sigma}_X^S)$ , so that  $\hat{\lambda}_1$  is the largest eigenvalue of  $\hat{\Sigma}_X^S$  or the spectral norm of  $\hat{\Sigma}_X^S$ . Applying this notation, our regularized PLS estimators can be written in terms of the following closed form representations:

1. **DRPLS:**

$$\hat{\beta}_{\text{drpls}} = \hat{G}_S \left[ \hat{G}_S' \hat{\Sigma}_X^S \hat{G}_S + \frac{1}{\hat{\lambda}_1} \hat{D}_{\lambda,\kappa}^2 \right]^{-1} \hat{G}_S' \hat{\sigma}_{Xy}^S, \quad (41)$$

where  $v > 0$  is a tuning parameter that can be either pre-specified or estimated.

2. **GRPLS:**

$$\hat{\beta}_{\text{grpls}} = \hat{G}_S \hat{D}_{\lambda,\kappa}^{-1} \left[ \hat{D}_{\lambda,\kappa}^{-1} \hat{G}_S' \hat{\Sigma}_X^S \hat{G}_S \hat{D}_{\lambda,\kappa}^{-1} \right]^+ \hat{D}_{\lambda,\kappa}^{-1} \hat{G}_S' \hat{\sigma}_{Xy}^S, \quad (42)$$

where  $A^+$  denotes the Moore-Penrose generalized inverse of the matrix  $A$ .

**Remark 6.1:** (a) Note that in empirical implementation, the general formulae given in expressions (41) and (42) lead to different versions of the DRPLS estimator and of the GRPLS estimator depending on which special case of the generalized thresholding estimator is used (i.e., hard thresholding, soft thresholding, SCAD, adaptive Lasso, or others that satisfy conditions (i)-(iii) above).

(b) It should also be noted that in an earlier version of this paper we proposed a regularized PLS estimator of the form  $\hat{\beta}_{\text{rpls}} = \hat{G}_S \left[ \hat{G}_S' \hat{\Sigma}_X^S \hat{G}_S \right]^{-1} \hat{G}_S' \hat{\sigma}_{Xy}^S$   
 $= \hat{G}_S \hat{D}_{\lambda,\kappa}^{-1} \left[ \hat{D}_{\lambda,\kappa}^{-1} \hat{G}_S' \hat{\Sigma}_X^S \hat{G}_S \hat{D}_{\lambda,\kappa}^{-1} \right]^{-1} \hat{D}_{\lambda,\kappa}^{-1} \hat{G}_S' \hat{\sigma}_{Xy}^S$ , which is clearly very similar to the GRPLS estimator given in expression (42) above, except that  $\hat{\beta}_{\text{rpls}}$  is constructed using an ordinary inverse whereas GRPLS uses a Moore-Penrose generalized inverse. Moreover, note that the DRPLS estimator  $\hat{\beta}_{\text{drpls}}$  has a form that resembles a ridge regression estimator. The reason why we now recommend DRPLS and GRPLS instead of a regularized PLS estimator based

on the ordinary inverse is because to apply these estimators to real world data such as the FRED-MD dataset, we must first perform preliminary training of our procedure to estimate the envelope dimension  $\kappa$ . Such a training exercise requires us to calculate any specified PLS estimator for a range of possible values of the envelope dimension including those values that are greater than the true value. However, by following arguments similar to those given for the last result in part (c) of Lemma B-3 of the Online Appendix<sup>15</sup>, we can show that,

$$\left\| \frac{D_{\lambda,k}^{-1} G' \bar{\Sigma}_X G D_{\lambda,k}^{-1}}{\lambda_1} - \underline{V}(\kappa, k)' \underline{D}_{\phi,\lambda}(\kappa) \underline{V}(\kappa, k) \right\|_2 = o(1), \quad (43)$$

where  $\underline{V}(\kappa, k)$  denotes a (possibly rectangular) Vandermonde-type matrix

$$\underline{V}_{\kappa \times k}(\kappa, k) = \begin{pmatrix} 1 & \underline{\lambda}_{1,1}/\lambda_1 & \underline{\lambda}_{1,1}^2/\lambda_1^2 & \cdots & \underline{\lambda}_{1,1}^{k-1}/\lambda_1^{k-1} \\ 1 & \underline{\lambda}_{1,2}/\lambda_1 & \underline{\lambda}_{1,2}^2/\lambda_1^2 & \cdots & \underline{\lambda}_{1,2}^{k-1}/\lambda_1^{k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \underline{\lambda}_{1,\kappa}/\lambda_1 & \underline{\lambda}_{1,\kappa}^2/\lambda_1^2 & \cdots & \underline{\lambda}_{1,\kappa}^{k-1}/\lambda_1^{k-1} \end{pmatrix}, \quad (44)$$

and where  $\underline{D}_{\phi,\lambda}(\kappa) = \text{diag}(\underline{\lambda}_{1,1}^3 \phi_1^2 / [\lambda_1^2 \underline{\lambda}_{1,1}], \underline{\lambda}_{1,2}^3 \phi_2^2 / [\lambda_1^2 \underline{\lambda}_{1,1}], \dots, \underline{\lambda}_{1,\kappa}^3 \phi_\kappa^2 / [\lambda_1^2 \underline{\lambda}_{1,1}])$ , and  $D_{\lambda,k} = \text{diag}(\lambda_1^{1/2}, \lambda_1^{3/2}, \dots, \lambda_1^{k-1/2})$ , with  $\lambda_1 = \lambda_{\max}(\bar{\Sigma}_X)$  and with  $\underline{\lambda}_{1,1}$  as previously defined in Assumption 4-2. This result shows that  $D_{\lambda,k}^{-1} G' \bar{\Sigma}_X G D_{\lambda,k}^{-1} / \lambda_1$ , the properly standardized version of the matrix  $G' \bar{\Sigma}_X G$ , converges to  $\underline{V}(\kappa, k)' \underline{D}_{\phi,\lambda}(\kappa) \underline{V}(\kappa, k)$ . Moreover, it is easy to see that if  $k > \kappa$ , that is if one overspecifies the envelope dimension, then  $\underline{V}(\kappa, k)$  is a rectangular matrix where the number of columns exceeds the number of rows, so that  $\underline{V}(\kappa, k)' \underline{D}_{\phi,\lambda}(\kappa) \underline{V}(\kappa, k)$  must be singular. It follows that no matter what type of model selection procedure we use to get an estimate of the envelope dimension (i.e., information criteria, cross validation, etc.), we must confront the issue where we need to calculate the specified PLS estimator (regularized or not) for certain cases where the specified envelope dimension may be greater than the true envelope dimension so that the matrix  $G' \bar{\Sigma}_X G$ , when appropriately standardized, is singular asymptotically. It follows that in these cases the sample versions of this matrix are also likely to be singular or at least have a large condition number in finite samples.<sup>16</sup> Thus, an added value of DRPLS and GRPLS is that they can help to facilitate the estimation of  $\kappa$ , by allowing us to circumvent this problem.

(c) Note also that the need to correctly specify  $\kappa$  is analogous to preliminary model selection performed when estimating the diffusion index model using principal component analysis (or PCA). Consider the situation described in expressions (27) and (28) previously, i.e., the case where the forecast and factor equations are given by

<sup>15</sup>Both Lemma B-3 and its proof can be found in Appendix B of the Online Appendix.

<sup>16</sup>Note that this issue has long been recognized in the PLS literature, although previous authors have not framed the discussion within an envelope model framework. See, for exmple, the discussion given in Stoica and Söderström (1998). However, to the best of our knowledge, we are the first to propose a way of finessing this problem using modified PLS estimators of the form DRPLS and GRPLS.

$$\begin{aligned}
y_{t+h} &= \alpha + (F_{1,t} - \mu_{F,1})' \varphi_1 + \varepsilon_{t+h} \\
X_t &= \mu_X + \Gamma_1 (F_{1,t} - \mu_{F,1}) + \Gamma_2 (F_{2,t} - \mu_{F,2}) + u_t.
\end{aligned}$$

$\kappa \times 1$ 
 $(K-\kappa) \times 1$

In this case, there are  $K$  total factors but only  $\kappa$  ( $< K$ ) of them are relevant for forecasting purposes so that to specify the most parsimonious diffusion index model for estimation by PCA, one needs to not only estimate the number of factors  $K$  but also  $\kappa$  the number of relevant factors that enter into the forecast equation. Interestingly, PLS-type procedures work directly with the many-predictors forecast equation so that only  $\kappa$  needs to be estimated. Given a correctly specified  $\kappa$ , we will show below that both  $\hat{\beta}_{\text{drpls}}$  and  $\hat{\beta}_{\text{grpls}}$  are consistent. Although not reported here, we have obtained some preliminary results showing that a certain information criterion based on a version of the DRPLS estimator can be used to obtain a consistent estimator of  $\kappa$ . We plan to write a follow-up paper on this issue.<sup>17</sup>

To help establish the consistency of  $\hat{\beta}_{\text{drpls}}$  and  $\hat{\beta}_{\text{grpls}}$  and characterize their rates of convergence, we make the following two additional assumptions.

**Assumption 6-2:** (i) For some  $q$  such that  $0 \leq q < 1$  and for some positive constant  $\mathbb{M}$ , consider the set of symmetric, positive semi-definite matrices given by

$$\mathcal{U}_{\Sigma}(q, c_0(N), \mathbb{M}) = \left\{ \Sigma_{N \times N} : \max_{1 \leq i \leq N} \Sigma_{(i,i)} \leq \mathbb{M}, \max_{1 \leq i \leq N} \sum_{j=1}^N |\Sigma_{(i,j)}|^q \leq c_0(N) \right\}; \quad (45)$$

i.e.,  $\mathcal{U}_{\Sigma}(q, c_0(N), \mathbb{M})$  is the set of symmetric, positive semi-definite matrices satisfying the two conditions given in expression (45). Suppose that  $\bar{\Sigma}_X \in \mathcal{U}_{\Sigma}(q, c_0(N), \mathbb{M})$ . (ii) Again, for some  $q$  such that  $0 \leq q < 1$  and for some positive constant  $\mathbb{M}$ , let

$$\mathcal{U}_{\sigma}(q, a(N), \mathbb{M}) = \left\{ \sigma_{N \times 1} : \max_{1 \leq i \leq N} |\sigma_i| \leq \mathbb{M}, \sum_{i=1}^N |\sigma_i|^q \leq a(N) \right\}; \quad (46)$$

i.e.,  $\mathcal{U}_{\sigma}(q, a(N), \mathbb{M})$  is the set of  $N \times 1$  covariance vectors satisfying the two conditions given in expression (46). Suppose that  $\bar{\sigma}_{X_y} \in \mathcal{U}_{\sigma}(q, a(N), \mathbb{M})$ .

**Assumption 6-3:** Let  $c_0(N)$ ,  $a(N)$ , and  $q \in [0, 1)$  be as defined in Assumption 6-2, and suppose that, as  $N, T, \lambda_1 \rightarrow \infty$ :

$$\frac{c_0(N)}{\lambda_1} \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}} = o(1) \text{ and } \sqrt{\frac{a(N)}{\lambda_1}} \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})} = o(1),$$

---

<sup>17</sup>It should be noted that in practical applications  $\kappa$  has often been estimated successfully using information criteria or cross-validation methods (see Cook (2018) Chapter 1), although there seems to be a dearth of theoretical results which provide provable guarantees for the use of these methods in envelope dimension estimation. That, we hope, will be the contribution of our follow-up paper.

where  $\lambda_1$  denotes the largest eigenvalue of the population covariance matrix,  $\bar{\Sigma}_X$ .

**Remark 6.2:** (a) Assumption 6-2 characterizes the parameter space for  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$ . The constraint set  $\mathcal{U}_\Sigma(q, c_0(N), \mathbb{M})$  has been called the uniform class of “approximately sparse” covariance matrices in the literature on high-dimensional covariance matrix estimation. It was first introduced by Bickel and Levina (2008). It should be noted that, depending on the specification of  $c_0(N)$  and  $a(N)$ , the parameter sets  $\mathcal{U}_\Sigma(q, c_0(N), \mathbb{M})$  and  $\mathcal{U}_\sigma(q, a(N), \mathbb{M})$  do not necessarily impose any sparsity on the underlying problem, since we can always choose to impose no sparsity at all, in which case both  $c_0(N)$  and  $a(N)$  will be specified to be of order  $N$ . On the other hand, the notation introduced in Assumption 6-2 will be helpful to us in presenting the asymptotic results given in Theorems 5, 6, and 7 below, as it enables us to more precisely characterize the effect that different degrees of sparsity, including a lack of sparsity, may have on the rate of convergence of our regularized PLS estimators and also of the point forecast constructed from these estimators. Note also that in the case where  $q = 0$ , equations (45) and (46) specialize to the following two expressions:

$$\mathcal{U}_\Sigma(0, c_0(N), \mathbb{M}) = \left\{ \Sigma_{N \times N} : \max_{1 \leq i \leq N} \Sigma_{(i,i)} \leq \mathbb{M}, \max_{1 \leq i \leq N} \sum_{j=1}^N \mathbb{I}\{|\Sigma_{(i,j)}| \neq 0\} \leq c_0(N) \right\} \quad (47)$$

$$\mathcal{U}_\sigma(0, a(N), \mathbb{M}) = \left\{ \sigma_{N \times 1} : \max_{1 \leq i \leq N} |\sigma_i| \leq \mathbb{M}, \sum_{i=1}^N \mathbb{I}\{|\sigma_i| \neq 0\} \leq a(N) \right\}. \quad (48)$$

(b) The class of constraints specified in expressions (45) and (46) corresponds to different constraint sets constructed using different  $\ell_q$  quasi-norms. The  $q = 0$  case, i.e., the constraint set given by expressions (47) and (48) above are most useful for tracking what is sometimes called exact sparsity, i.e., cases where there may be elements of  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$  which are exactly zero. Even in cases where there is no exact sparsity, however, there could be approximate sparsity where a significant number of the elements of  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$  may not be exactly zero but may be relatively small in magnitude.  $\ell_q$  quasi-norms for  $0 < q < 1$ , in contrast to the  $\ell_0$  quasi-norm, can be more useful for capturing these types of approximate sparsity which, in turn, can also be viewed as a type of dense model in the language introduced by the well-known paper due to Giannone, Lenza, and Primiceri (2021).

(c) Assumption 6-3 gives rate conditions that are imposed in conducting our asymptotic analysis. These conditions can accommodate a wide variety of alternative scenarios with different signal strengths, as characterized by the order of magnitude of  $\lambda_1$ ; different degrees of sparsity, as characterized by the order of magnitude of  $c_0(N)$  and  $a(N)$ ; and different rates at which  $N$  and  $T$  diverge to infinity. In the particular case where  $X_t$  is generated by a factor model, it is easy to see that Assumption 6-3 is satisfied when the usual factor pervasiveness condition is assumed, even if there is no sparsity. This is because in this case  $\lambda_1 \asymp N$ ,  $c_0(N) \asymp N$ ,  $a(N) \asymp N$ , and  $q \in [0, 1]$ ; so that

$$\frac{c_0(N)}{\lambda_1} \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}} \asymp \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}} = o(1),$$

$$\sqrt{\frac{a(N)}{\lambda_1}} \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})} \asymp \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})} = o(1),$$

given the condition  $\ln N/T \rightarrow 0$ <sup>18</sup>. In addition, Assumption 6-3 is also satisfied under various weak factor situations where  $\lambda_1$  grows much slower than  $N$ . This is clearly true in cases where there are some parameter sparsity so that the constraints given can be satisfied even when both  $c_0(N)$  and  $a(N)$  have order of magnitude less than  $N$ , i.e.,  $c_0(N) \ll N$  and  $a(N) \ll N$ . However, even in many situations where factors may be weak and there is no sparsity, Assumption 6-3 may still be satisfied. For example, consider the case where  $c_0(N) \asymp N$ ,  $a(N) \asymp N$ ,  $N \asymp T$ , and  $q = 0$ . In this case, we will have

$$\begin{aligned} \frac{c_0(N)}{\lambda_1} \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}} &\asymp \frac{N}{\lambda_1} \sqrt{\frac{\ln N}{N}} = \frac{\sqrt{N \ln N}}{\lambda_1} = o(1), \\ \sqrt{\frac{a(N)}{\lambda_1}} \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})} &\asymp \sqrt{\frac{N}{\lambda_1}} \sqrt{\frac{\ln N}{N}} = \sqrt{\frac{\ln N}{\lambda_1}} = \sqrt{\frac{\lambda_1}{N}} \frac{\sqrt{N \ln N}}{\lambda_1} = o(1), \end{aligned}$$

even if  $\lambda_1 \ll N$  as long as  $\sqrt{N \ln N} \ll \lambda_1$  so that  $\lambda_1$  may diverge at a rate slower than  $N$ . In fact, from this example, it is easy to see that if  $T$  diverges at a rate faster than  $N$ , then Assumption 6-3 can be satisfied for  $\lambda_1$  having an even smaller order of magnitude than that for the case where  $N \asymp T$ . Hence, Assumption 6-3 allows for a wide variety of signal-to-noise type trade-offs between the order of magnitude of the leading eigenvalue  $\lambda_1$ , the degree of sparsity in  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$ , and the rates of divergence of  $N$  and  $T$ .

Theorem 5 below provides consistency and convergence rate results for both the DRPLS estimator  $\hat{\beta}_{\text{drpls}}$  and the GRPLS estimator  $\hat{\beta}_{\text{grpls}}$ . Theorem 6 provides consistency and convergence rate results for point forecasts obtained by estimating the conditional mean function of the many-predictors forecast equation using  $\hat{\beta}_{\text{drpls}}$  and also using  $\hat{\beta}_{\text{grpls}}$ .

**Theorem 5:** Suppose that Assumptions 4-1, 4-2, 4-3, 4-4, 6-1, 6-2, and 6-3 hold.

- (a) Let  $\hat{\beta}_{\text{drpls}} = \hat{G}_S \left[ \hat{G}_S' \hat{\Sigma}_X^S \hat{G}_S + \hat{\lambda}_1^{-v} \hat{D}_{\lambda, \kappa}^2 \right]^{-1} \hat{G}_S' \hat{\sigma}_{Xy}^S$  be the DRPLS estimator previously defined in expression (41) above. Suppose further that  $(\ln N)^{\frac{2}{\gamma_*}-1} = o(T)$ , where  $\gamma_* = \gamma_1 \gamma_2 / (\gamma_1 + 2\gamma_2)$  and where  $\gamma_1$  and  $\gamma_2$  are the positive constants defined in Assumption 4-3. Under these conditions, let  $\eta_1 = B_1 \sqrt{\ln N/T}$  and  $\eta_2 = B_2 \sqrt{\ln N/T}$  for positive constants  $B_1$  and  $B_2$  sufficiently large. Then,

$$\left\| \hat{\beta}_{\text{drpls}} - \beta \right\|_2 = O_p \left( \max \left\{ \frac{c_0(N)}{\lambda_1^{\frac{3}{2}}} \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}}, \frac{\sqrt{a(N)}}{\lambda_1} \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})}, \frac{1}{\lambda_1^{\frac{3}{2}}} \right\} \right) = o_p(1).$$

---

<sup>18</sup> $\ln N/T \rightarrow 0$  is a mild condition that is usually assumed in high-dimensional regression problems. In Theorems 5, 6, and 7 below, we explicitly assume that there exists positive constant  $\gamma_* \in (0, 1)$  such that  $(\ln N)^{\frac{2}{\gamma_*}-1} = o(T)$ , where  $\gamma_* = \gamma_1 \gamma_2 / (\gamma_1 + 2\gamma_2)$  and where  $\gamma_1$  and  $\gamma_2$  are the positive constants defined in Assumption 4-3 so that  $0 < \gamma_* < 1$ . It is easy to see that this latter condition implies the condition that  $\ln N/T \rightarrow 0$ .

(b) Let  $\hat{\beta}_{\text{grpls}} = \hat{G}_S \hat{D}_{\lambda, \kappa}^{-1} \left[ \hat{D}_{\lambda, \kappa}^{-1} \hat{G}'_S \hat{\Sigma}_X^S \hat{G}_S \hat{D}_{\lambda, \kappa}^{-1} \right]^+ \hat{D}_{\lambda, \kappa}^{-1} \hat{G}'_S \hat{\sigma}_{Xy}^S$  be the GRPLS estimator previously defined in expression (42) above. Suppose further that  $(\ln N)^{\frac{2}{\gamma_*}-1} = o(T)$ , where  $\gamma_* = \gamma_1 \gamma_2 / (\gamma_1 + 2\gamma_2)$  and where  $\gamma_1$  and  $\gamma_2$  are the positive constants defined in Assumption 4-3. Under these conditions, let  $\eta_1 = B_1 \sqrt{\ln N/T}$  and  $\eta_2 = B_2 \sqrt{\ln N/T}$  for positive constants  $B_1$  and  $B_2$  sufficiently large. Then,

$$\left\| \hat{\beta}_{\text{grpls}} - \beta \right\|_2 = O_p \left( \max \left\{ \frac{c_0(N)}{\lambda_1^{\frac{3}{2}}} \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}}, \frac{\sqrt{a(N)}}{\lambda_1} \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})}, \frac{1}{\lambda_1^{\frac{3}{2}}} \right\} \right) = o_p(1).$$

Note that after estimating  $\beta$  using data in the training sample  $\{y_{t+h}, X'_t\}_{t=1}^{T_h}$ , we can perform an  $h$ -period(s) ahead forecast of  $y_{T+h}$  based on either the point forecast  $\hat{y}_{T+h}^{\text{drpls}} = \bar{y}_{T_h} + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{drpls}}$  or the point forecast  $\hat{y}_{T+h}^{\text{grpls}} = \bar{y}_{T_h} + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{grpls}}$ , where  $T_h^{-1} \sum_{t=1}^{T_h} y_{t+h}$ . Theorem 6 below provides some asymptotic properties for both of these point forecasts. To obtain these results we need to impose the following condition on the covariance matrix of  $X_T$ .

**Assumption 6-4:** Let  $\Sigma_{W,T} = E[W_T W_T']$ . There exists a constant  $C > 0$  such that, for all  $N$  and  $T$ ,  $\frac{\text{tr}\{\Sigma_{W,T}\}}{N} \leq C < \infty$ .

**Remark 6.3:** To help interpret Assumption 6-4, note first that under the near-reduced-rank envelope structure specified in Assumption 4-2, it is easy to see that there exists a constant,  $C^\dagger > 0$ , such that:  $\frac{\text{tr}\{\bar{\Sigma}_W\}}{N} \leq C^\dagger < \infty$ , where  $\bar{\Sigma}_W = T_h^{-1} \sum_{t=1}^{T_h} E[W_t W_t']$ , as previously defined. Hence, Assumption 6-4 simply requires this one feature of our training model to also hold for  $\Sigma_{W,T} = E[W_T W_T']$  without imposing any stronger stationarity-type assumptions. This is a rather mild and sensible condition, since if no feature of our model is expected to extend beyond the training period then we cannot hope to have reasonable forecasting performance.

**Theorem 6:** Let  $\hat{y}_{T+h}^{\text{drpls}} = \bar{y}_{T_h} + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{drpls}}$  and  $\hat{y}_{T+h}^{\text{grpls}} = \bar{y}_{T_h} + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{grpls}}$ , where  $\bar{y}_{T_h} = T_h^{-1} \sum_{t=1}^{T_h} y_{t+h}$  and  $T_h = T - h$ . Suppose that the conditions of Theorem 5 all hold, and suppose in addition that Assumption 6-4 also holds. Then,

(a)

$$\begin{aligned} & \hat{y}_{T+h}^{\text{drpls}} - (\alpha + [X_T - \mu_X]' \beta) \\ &= O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} \right\} \right) + O_p \left( \frac{1}{\sqrt{T}} \right). \end{aligned}$$

(b)

$$\begin{aligned} & \widehat{y}_{T+h}^{\text{grpls}} - (\alpha + [X_T - \mu_X]' \beta) \\ = & O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} \right\} \right) + O_p \left( \frac{1}{\sqrt{T}} \right). \end{aligned}$$

**Remark 6.4:** (a) To help interpret Theorem 5, it is useful to specialize our framework to the case where  $X_t$  is generated by a factor model. In this setting, consider the special case where  $\lambda_1 \asymp N$ , which corresponds to the situation where the usual factor pervasiveness condition holds. Within this setting, consider also the case where:

$$\overline{\Sigma}_X \in \mathcal{U}_\Sigma(0, N, \mathbb{M}) = \left\{ \Sigma_{N \times N} : \max_{1 \leq i \leq N} \Sigma_{(i,i)} \leq \mathbb{M}, \max_{1 \leq i \leq N} \sum_{j=1}^N \mathbb{I}\{|\Sigma_{(i,j)}| \neq 0\} \leq N \right\} \quad (49)$$

$$\overline{\sigma}_{Xy} \in \mathcal{U}_\sigma(0, N, \mathbb{M}) = \left\{ \sigma_{N \times 1} : \max_{1 \leq i \leq N} |\sigma_i| \leq \mathbb{M}, \sum_{i=1}^N \mathbb{I}\{|\sigma_i| \neq 0\} \leq N \right\}, \quad (50)$$

Note that the constraints given by expressions (49) and (50) do not require that there be any sparsity so that no (exact) sparsity is assumed in this case. Moreover, suppose that either  $T \ll N$  or  $T \asymp N$ . In this situation, given that  $c_0(N) \asymp a(N) \asymp N$ , it is easily seen by setting  $q = 0$  that:

$$\begin{aligned} \frac{c_0(N)}{\lambda_1^{\frac{3}{2}}} \left[ \frac{\ln N}{T} \right]^{\frac{1-q}{2}} &= \frac{c_0(N)}{\lambda_1^{\frac{3}{2}}} \sqrt{\frac{\ln N}{T}} \asymp \sqrt{\frac{\ln N}{NT}}, \\ \frac{\sqrt{a(N)}}{\lambda_1} \left[ \frac{\ln N}{T} \right]^{\frac{1}{2}(1-\frac{q}{2})} &= \frac{\sqrt{a(N)}}{\lambda_1} \sqrt{\frac{\ln N}{T}} \asymp \sqrt{\frac{\ln N}{NT}}, \\ \frac{1}{\lambda_1^{\frac{3}{2}}} &\asymp \frac{1}{N^{\frac{3}{2}}} \lesssim \sqrt{\frac{\ln N}{NT}}, \text{ given that } T \lesssim N. \end{aligned}$$

It follows that, in this case,

$$\left\| \widehat{\beta}_{\text{drpls}} - \beta \right\|_2 = O_p \left( \sqrt{\frac{\ln N}{NT}} \right) \text{ and } \left\| \widehat{\beta}_{\text{grpls}} - \beta \right\|_2 = O_p \left( \sqrt{\frac{\ln N}{NT}} \right),$$

so that having an envelope structure helps to achieve a rate of convergence that is faster than the usual near oracle convergence rate of  $O_p \left( \sqrt{\ln N/T} \right)$  for high-dimensional problems.

(b) We can also consider Theorem 6 in the setting given in part (a) of this remark, where  $\lambda_1 \asymp N$  and  $T \lesssim N$  and where the constraints given in expressions (49) and (50) above are

assumed to be satisfied so that, in particular, no sparsity is assumed. In this case, since

$$\begin{aligned} \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}} &= \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \sqrt{\frac{\ln N}{T}} \asymp \sqrt{\frac{\ln N}{T}}, \\ \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})} &= \frac{\sqrt{a(N)N}}{\lambda_1} \sqrt{\frac{\ln N}{T}} \asymp \sqrt{\frac{\ln N}{T}}, \\ \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} &\asymp \frac{1}{N} \lesssim \sqrt{\frac{\ln N}{T}} \text{ (given that } T \lesssim N), \text{ and } \frac{1}{\sqrt{T}} \lesssim \sqrt{\frac{\ln N}{T}}, \end{aligned}$$

it is easy to see that

$$\begin{aligned} \hat{y}_{T+h}^{\text{drpls}} - (\alpha + [X_T - \mu_X]' \beta) &= O_p \left( \sqrt{\frac{\ln N}{T}} \right) \text{ and} \\ \hat{y}_{T+h}^{\text{grpls}} - (\alpha + [X_T - \mu_X]' \beta) &= O_p \left( \sqrt{\frac{\ln N}{T}} \right). \end{aligned}$$

(c) Theorems 5 and 6 show, on the other hand, that when there is sparsity, i.e., in cases where the constraints are fulfilled even when the order of magnitude of  $c_0(N)$  and/or  $a(N)$  is less than  $N$ , then, our procedures can potentially accommodate cases with weaker signal, i.e., cases where  $\lambda_1$  diverges at rates much less than  $N$ , and still achieve fairly fast rates of convergence. Hence, one of the useful features of the asymptotic results given in Theorems 5 and 6 is that they show the interactions between the different aspects of the problem, including the degree of sparsity, the strength of the signal as given by the order of magnitude of  $\lambda_1$ , and the orders of magnitude of  $N$  and  $T$ . In particular, our results shed light on how these different attributes contribute to the rate of convergence of both  $\hat{\beta}_{\text{drpls}}$  and  $\hat{\beta}_{\text{grpls}}$  and also to the rate of convergence of the point forecasts constructed from  $\hat{\beta}_{\text{drpls}}$  and  $\hat{\beta}_{\text{grpls}}$ .

(d) It is also of interest to compare our regularized PLS procedure with the principal component (PC) method in the case where  $X_t$  is generated by a factor model with weak loadings. In an interesting recent paper, Bai and Ng (2023) provided weaker conditions under which factors can be estimated consistently, up to a rotation matrix, by the PC method. Two critical conditions (given, respectively, as part (ii) of Assumption A2 and Assumption A4 in their paper) are:

$$(1) \lim_{N \rightarrow \infty} \frac{\Lambda^{0'} \Lambda^0}{N^\alpha} = \Sigma_\Lambda > 0, \text{ for some } \alpha \in (0, 1], \text{ and } (2) \frac{N}{N^\alpha T} \rightarrow 0, \quad (51)$$

where  $\Lambda^0$  is their notation for the (true) loading matrix. Note that condition (1) in expression (51) implies that the top eigenvalue of the covariance matrix of  $X_t$  will grow at the rate  $N^\alpha$ .

Condition (2) in expression (51) requires that  $N/T \ll N^\alpha$ . These two conditions together thus imply that the top eigenvalue must grow faster than  $N/T$ . Now, consider the scenario where  $N^{\frac{1}{3}} \ll \lambda_1 \ll \sqrt{N}$  and  $T \asymp \sqrt{N}$ . In this setting, it is easily seen that the conditions given in Bai and Ng (2023) are not fulfilled since  $\lambda_1 \ll \sqrt{N} = \frac{N}{\sqrt{N}} \asymp \frac{N}{T}$ . Given that the conditions provided in Bai and Ng (2023) are likely to be the weakest possible for the consistent estimation of factors using the PC method, it is not clear that in this setting one will be able to use the PC method to obtain consistent estimates of the factors and of the conditional mean function of the associated factor-augmented forecast equation. On the other hand, it is easily seen that in this setting

$$\hat{y}_{T+h}^{\text{drpls}} - (\alpha + [X_T - \mu_X]' \beta) = o_p(1) \text{ and } \hat{y}_{T+h}^{\text{grpls}} - (\alpha + [X_T - \mu_X]' \beta) = o_p(1),$$

provided that:

$$c_0(N) \ll \frac{\lambda_1^{\frac{3}{2}}}{\sqrt{N}} \left( \frac{T}{\ln N} \right)^{\frac{1-q}{2}} \text{ and } a(N) \ll \frac{\lambda_1^2}{N} \left( \frac{T}{\ln N} \right)^{(1-\frac{q}{2})}. \quad (52)$$

In addition, since  $N^{\frac{1}{3}} \ll \lambda_1 \ll \sqrt{N}$  and  $T \asymp \sqrt{N}$ , it follows by straightforward calculations that:

$$\begin{aligned} \frac{\lambda_1^{\frac{3}{2}}}{\sqrt{N}} \left( \frac{T}{\ln N} \right)^{\frac{1-q}{2}} &\gg \left( \frac{T}{\ln N} \right)^{\frac{1-q}{2}} \rightarrow \infty, \text{ for all } q \in [0, 1), \text{ and} \\ \frac{\lambda_1^2}{N} \left( \frac{T}{\ln N} \right)^{(1-\frac{q}{2})} &\gg N^{\frac{2-3q}{12}} \left( \frac{1}{\ln N} \right)^{(1-\frac{q}{2})} \rightarrow \infty, \text{ for } 0 \leq q < \frac{2}{3}, \end{aligned}$$

so that the conditions given in expression (52) can be satisfied even if both  $c_0(N)$  and  $a(N)$  diverge as  $N \rightarrow \infty$ , as long as they grow at rates slower than  $\frac{\lambda_1^{\frac{3}{2}}}{\sqrt{N}} \left( \frac{T}{\ln N} \right)^{\frac{1-q}{2}}$  and  $\frac{\lambda_1^2}{N} \left( \frac{T}{\ln N} \right)^{(1-\frac{q}{2})}$ , respectively. Hence, although achieving consistent estimation of the conditional mean function of the forecast equation in this case requires some parameter restrictions on  $\bar{\Sigma}_X$  and  $\bar{\sigma}_{Xy}$ , the parameter restrictions needed are not so severe that they require  $c_0(N)$  and  $a(N)$  to be fixed constants that do not diverge, as  $N \rightarrow \infty$ . This example illustrates how the regularized procedures we have proposed may be quite useful in certain situations where the underlying signal is quite weak, since in such cases consistent estimation of the conditional mean function of the forecast equation may still be achievable under our regularized PLS procedures due to their ability to take advantage of possible sparsity constraints on the elements of  $\bar{\Sigma}_X$  and/or  $\bar{\sigma}_{Xy}$ .

So far, we have shown in Theorem 6 that if the forecast equation is a many-predictors regression with an approximate envelope structure, then we can consistently estimate its conditional mean function using the point forecasts constructed from regularized PLS estimators. Now, we want to show that these point forecasts can also consistently estimate the conditional mean function of a diffusion index forecast equation if that happens to be the true model. However, before we proceed, we first point out some minor differences between

these two cases. Suppose first that the true model takes the form of a diffusion index model, i.e.,

$$y_{t+h} = \alpha + (F_t - \mu_F)' \varphi + \varepsilon_{t+h} \quad (53)$$

$$X_t = \mu_X + \Gamma (F_t - \mu_F) + u_t. \quad (54)$$

In this case, it is easy to see that we can transform the first equation (i.e., the forecast equation) into a form similar to that of an envelope-type many-predictors regression as follows:

$$\begin{aligned} y_{t+h} &= \alpha + (F_t - \mu_F)' \varphi + \varepsilon_{t+h} \\ &= \alpha + (F_t - \mu_F)' \frac{\Gamma' \Gamma}{\Delta_{1,1}} \varphi + \varepsilon_{t+h} \quad \left( \text{since } \frac{\Gamma' \Gamma}{\Delta_{1,1}} = I_K \right) \\ &= \alpha + (F_t - \mu_F)' \Gamma' \frac{\Gamma}{\sqrt{\Delta_{1,1}}} \frac{\varphi}{\sqrt{\Delta_{1,1}}} + \varepsilon_{t+h} \\ &= \alpha + (F_t - \mu_F)' \Gamma' \beta + \varepsilon_{t+h} \\ &= \alpha + [(X_t - \mu_X)' - u_t'] \beta + \varepsilon_{t+h} \\ &= \alpha + (X_t - \mu_X)' \beta + \xi_{t+h}, \end{aligned} \quad (55)$$

where  $\xi_{t+h} = -u_t' \beta + \varepsilon_{t+h}$  and where  $\beta = L_1 \varphi / \sqrt{\Delta_{1,1}}$  with  $L_1 = \Gamma / \sqrt{\Delta_{1,1}}$ . Notice that although the right-hand side of the last line of expression (55) has the form of a many-predictors forecast equation, one subtle difference between expression (55) and the many-predictors forecast equation discussed in Section 4 is that the error  $\xi_{t+h}$  in this case is actually correlated with  $X_t$  since both  $X_t$  and  $\xi_{t+h}$  includes the error vector  $u_t$ . However, it turns out that the point forecasts constructed using our regularized PLS estimators (i.e.,  $\hat{y}_{T+h}^{\text{drpls}} = \bar{y}_T + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{drpls}}$  and  $\hat{y}_{T+h}^{\text{grpls}} = \bar{y}_T + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{grpls}}$ ) are still consistent, as we will show in Theorem 7 below. One way to understand why this is so is to note that, given the result provided in Theorem 4, the key step to consistent estimation of the conditional mean function of the diffusion index forecast equation is to show that we can construct PLS-type estimators which will consistently estimate, at a fast enough rate of convergence, the population analogue of the PLS estimator, i.e.,  $\bar{\beta} = G (G' \bar{\Sigma}_X G)^{-1} G' \bar{\sigma}_{Xy}$ , where  $G = \begin{pmatrix} \bar{\sigma}_{Xy} & \bar{\Sigma}_X \bar{\sigma}_{Xy} & \cdots & \bar{\Sigma}_X^{\kappa-1} \bar{\sigma}_{Xy} \end{pmatrix}$ . Now, note that  $\bar{\beta}$  is a function only of  $\bar{\sigma}_{Xy} = T_h^{-1} \sum_{t=1}^{T_h} E \{(X_t - \mu_X)(y_{t+h} - \alpha)\}$  and  $\bar{\Sigma}_X = T_h^{-1} \sum_{t=1}^{T_h} E [(X_t - \mu_X)(X_t - \mu_X)']$ . Hence, unlike estimation by OLS, consistent estimation via PLS-type procedures does not require that we are able to rewrite the model in terms of a linear representation where the regressors are not correlated with the error term. In fact, we show in Appendix E of the Online Appendix that even when the true model is the diffusion index specification given in equations (53) and (54) as opposed to the many-predictors regression model, we can still

prove the following uniform (across indices) convergence results:

$$\max_{1 \leq i \leq N} |\hat{\sigma}_{Xy,i} - \bar{\sigma}_{Xy,i}| = O_p \left( \sqrt{\frac{\ln N}{T}} \right) \quad \text{and} \quad \max_{1 \leq i,j \leq N} |\hat{\Sigma}_{X,(i,j)} - \bar{\Sigma}_{X,(i,j)}| = O_p \left( \sqrt{\frac{\ln N}{T}} \right).$$

See, in particular, Lemmas E-10 and E-14 in Appendix E of the Online Appendix. This, in turn, allows us to construct regularized PLS estimators based on the regularized estimators  $\hat{\sigma}_{Xy}^S$  and  $\hat{\Sigma}_X^S$ , which can potentially improve upon the performance of the unregularized PLS estimator when the strength of the underlying factors is weak.

To provide a formal result, we first state some assumptions, which are imposed on the diffusion index model specified in expressions (53)-(54).

**Assumption (i):**  $\Gamma' \Gamma / \lambda_{1,1} = I_K$  and  $\bar{\Sigma}_F = \underline{\Lambda} / \lambda_{1,1}$ , where  $\bar{\Sigma}_F = T_h^{-1} \sum_{t=1}^{T_h} E[(F_t - \mu_F)(F_t - \mu_F)']$  is the factor covariance matrix and where  $\underline{\Lambda}$  is a diagonal matrix whose diagonal elements are the nonzero eigenvalues of the common component covariance matrix  $\Gamma' \bar{\Sigma}_F \Gamma$ .

**Assumption (ii):** (a) Suppose that  $E[u_t] = 0$  and  $E[(F_t - \mu_F)u_t] = 0$ , for all  $t \in \mathbb{Z}$ . (b) Suppose that there exists some constant  $C^\dagger > 0$  such that  $\sup_t \lambda_{\max}(\Sigma_{u,t}) \leq C^\dagger < \infty$  and  $E\left\{(F_{t,g} - \mu_{F,g})^4\right\} \leq C^\dagger < \infty$ , for all  $g \in \{1, \dots, K\}$  and for all  $t \in \mathbb{Z}$ , where  $\Sigma_{u,t} = E[u_t u_t']$  and where  $F_{t,g}$  and  $\mu_{F,g}$  denote the  $g^{th}$  components of  $F_t$  and  $\mu_F$ , respectively.

**Assumption (iii):**  $E[\varepsilon_t] = 0$  for all  $t$ . In addition, let the forecast horizon  $h$  be a finite positive integer, and assume that  $E[(F_t - \mu_F)\varepsilon_{t+h}] = 0$  for all  $t$  and  $E[u_{it}\varepsilon_{t+h}] = 0$  for all  $i$  and  $t$ .

**Assumption (iv):** Let  $\Gamma_{(i,g)}$  denote the  $(i, g)^{th}$  element of  $\Gamma$ , and suppose that there exists some constant  $c^\ddagger > 1$  such that  $\max_{1 \leq i \leq N, 1 \leq g \leq K} |\Gamma_{(i,g)}| \leq c^\ddagger < \infty$  for all  $N$ .

**Assumption (v):** Let  $\gamma$ ,  $\gamma_1$ , and  $\gamma_2$  be positive constants satisfying the relationship  $\frac{1}{\gamma} = \frac{1}{\gamma_1} + \frac{1}{\gamma_2}$  or  $\gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}$ , such that  $0 < \gamma < 1$ . Suppose that the following conditions hold: (a) There exists a constant  $b_1 > 0$ , such that for any  $s > 0$ , for any  $\delta \in \mathbb{R}^N$ , with  $\|\delta\|_2 = 1$ , and for  $g \in \{1, \dots, K\}$ :

$$\begin{aligned} \sup_{t \in \mathbb{Z}} P(|u'_t \delta| > s) &\leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \\ \sup_{t \in \mathbb{Z}} P(|F_{t,g} - \mu_{F,g}| > s) &\leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \\ \sup_{t \in \mathbb{Z}} P(|\varepsilon_t| > s) &\leq \exp \left\{ 1 - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\} \leq \bar{C} \exp \left\{ - \left( \frac{s}{b_1} \right)^{\gamma_1} \right\}, \end{aligned}$$

where  $\bar{C}$  is a positive constant such that  $\bar{C} \geq \exp\{1\}$ , where  $\mu_F = E[F_t]$ , and where  $F_{t,g}$  and  $\mu_{F,g}$  are the  $g^{th}$  element of  $F_t$  and  $\mu_F$ , respectively. (b) Let  $\underline{V}_t = (\varepsilon_t, F'_t, u'_t)'$ . Define  $\underline{\mathcal{F}}_{-\infty}^t = \sigma(\dots, \underline{V}_{t-2}, \underline{V}_{t-1}, \underline{V}_t)$  and  $\underline{\mathcal{F}}_{t+m}^\infty = \sigma(\underline{V}_{t+m}, \underline{V}_{t+m+1}, \underline{V}_{t+m+2}, \dots)$ . Suppose that

there exist positive constants  $\gamma_2$ ,  $a$ , and  $c$ , such that:

$$\alpha(m) = \sup_t \alpha(\underline{\mathcal{F}}_{-\infty}^t, \underline{\mathcal{F}}_{t+m}^\infty) \leq a \exp(-cm^{\gamma_2}) \text{ for all } m \geq 1,$$

where

$$\alpha(\underline{\mathcal{F}}_{-\infty}^t, \underline{\mathcal{F}}_{t+m}^\infty) = \sup_{G \in \underline{\mathcal{F}}_{-\infty}^t, H \in \underline{\mathcal{F}}_{t+m}^\infty} |P(G \cap H) - P(G)P(H)|.$$

**Remark 6.5:** Most of the assumptions given above are similar in nature to the assumptions stated in Section 4, except that they are more specific to the diffusion index case.

We also consider the two scenarios previously analyzed in Section 5.

**Scenario 1** ( $1 \leq \kappa < K$  **Case**). Assume that the following conditions hold:

**Assumption S1-1:** There exists a constant  $\bar{c} > 1$  such that  $0 < \frac{1}{\bar{c}} \leq \min_{1 \leq g \leq \kappa} |\varphi_{1,g}| \leq \max_{1 \leq g \leq \kappa} |\varphi_{1,g}| \leq \bar{c} < \infty$ , whereas  $\varphi_2 = (\varphi_{2,1}, \dots, \varphi_{2,K-\kappa})' = 0$ .

**Assumption S1-2:** (a) Suppose that  $\underline{\lambda}_{1,g} \rightarrow \infty$  and  $\underline{\lambda}_{2,j} \rightarrow \infty$ , as  $N, T \rightarrow \infty$  (for  $g = 1, \dots, \kappa$  and  $j = 1, \dots, K - \kappa$ ) such that  $\underline{\lambda}_{1,1} \asymp \dots \asymp \underline{\lambda}_{1,\kappa} \asymp \underline{\lambda}_{2,1} \asymp \dots \asymp \underline{\lambda}_{2,K-\kappa}$ . (b) Suppose also that there exists a constant,  $\underline{c}$ , with  $0 < \underline{c} < 1$ , such that  $(\underline{\lambda}_{1,g} - \underline{\lambda}_{1,k}) / \underline{\lambda}_{1,1} \geq \underline{c}$ , for all  $g < k$ , with  $g \in \{1, \dots, \kappa - 1\}$  and  $k \in \{2, \dots, \kappa\}$ , for all  $N$  and  $T$  sufficiently large.

**Scenario 2** ( $\kappa = K$  **Case**). Assume that the following conditions hold:

**Assumption S2-1:** There exists a constant  $\bar{c} > 1$  such that  $0 < \frac{1}{\bar{c}} \leq \min_{1 \leq g \leq K} |\varphi_{1,g}| \leq \max_{1 \leq g \leq K} |\varphi_{1,g}| \leq \bar{c} < \infty$ .

**Assumption S2-2** (a) Suppose that  $\underline{\lambda}_{1,g} \rightarrow \infty$  as  $N, T \rightarrow \infty$  (for  $g = 1, \dots, K$ ) such that  $\underline{\lambda}_{1,1} \asymp \dots \asymp \underline{\lambda}_{1,K}$ . (b) Suppose also that there exists a constant,  $\underline{c}$ , with  $0 < \underline{c} < 1$ , such that  $(\underline{\lambda}_{1,g} - \underline{\lambda}_{1,k}) / \underline{\lambda}_{1,1} \geq \underline{c}$ , for all  $g < k$ , with  $g \in \{1, \dots, K - 1\}$  and  $k \in \{2, \dots, K\}$ , for all  $N$  and  $T$  sufficiently large.

We now present our results showing that point forecasts constructed from our regularized PLS estimators can consistently estimate the conditional mean function of the diffusion index forecast equation.

**Theorem 7:** Let  $\hat{y}_{T+h}^{\text{drpls}} = \bar{y}_{T_h} + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{drpls}}$  and  $\hat{y}_{T+h}^{\text{grpls}} = \bar{y}_{T_h} + [X_T - \bar{X}_{T_h}]' \hat{\beta}_{\text{grpls}}$ , where  $\hat{\beta}_{\text{drpls}}$  and  $\hat{\beta}_{\text{grpls}}$  are as previously defined in Theorem 5 and where  $\bar{y}_{T_h} = T_h^{-1} \sum_{t=1}^{T_h} y_{t+h}$ .

Suppose that Assumptions (i)-(v) are satisfied, and suppose that  $(\ln N)^{\frac{2}{\gamma_*}-1} = o(T)$ , where  $\gamma_* = \gamma_1 \gamma_2 / (\gamma_1 + 2\gamma_2)$  and where  $\gamma_1$  and  $\gamma_2$  are the positive constants defined in Assumption (v) above. In addition, suppose that Assumptions 6-1, 6-2, and 6-3 hold.

(a) Consider Scenario 1 where  $1 \leq \kappa < K$ . Suppose that Assumptions S1-1 and S1-2 hold.

Then,

$$\begin{aligned}
& \hat{y}_{T+h}^{\text{drpls}} - \left( \alpha + [F_{1,T} - \mu_{F,1}]' \varphi_1 \right) \\
&= O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} \right\} \right) \\
&+ O_p \left( \max \left\{ \frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{T}} \right\} \right).
\end{aligned}$$

Consider Scenario 2 where  $\kappa = K$ . Suppose that Assumptions S2-1 and S2-2 hold. Then,

$$\begin{aligned}
& \hat{y}_{T+h}^{\text{drpls}} - \left( \alpha + [F_T - \mu_F]' \varphi \right) \\
&= O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} \right\} \right) \\
&+ O_p \left( \max \left\{ \frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{T}} \right\} \right).
\end{aligned}$$

(b) Consider Scenario 1 where  $1 \leq \kappa < K$ . Suppose that Assumptions S1-1 and S1-2 hold. Then,

$$\begin{aligned}
& \hat{y}_{T+h}^{\text{grpls}} - \left( \alpha + [F_{1,T} - \mu_{F,1}]' \varphi_1 \right) \\
&= O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} \right\} \right) \\
&+ O_p \left( \max \left\{ \frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{T}} \right\} \right).
\end{aligned}$$

Consider Scenario 2 where  $\kappa = K$ . Suppose that Assumptions S2-1 and S2-2 hold. Then,

$$\begin{aligned}
& \hat{y}_{T+h}^{\text{grpls}} - \left( \alpha + [F_T - \mu_F]' \varphi \right) \\
&= O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} \right\} \right) \\
&+ O_p \left( \max \left\{ \frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{T}} \right\} \right),
\end{aligned}$$

**Remark 6.6:** In an interesting paper, De Mol, Giannone, and Reichlin (2008) study many-predictors forecasting within a Bayesian framework using a diffusion index-type-model given

by the equations  $y_{t+h} = \gamma F_t + v_{t+h}$  and  $X_t = \Lambda F_t + \xi_t$ , where  $y_{t+h}$  is the target variable of interest,  $F_t$  is a  $r \times 1$  vector of latent factors,  $\Lambda$  is an  $N \times r$  matrix of factor loading, and  $X_t$  is an  $N \times 1$  vector of predictor variables. De Mol et al. (2008) define  $\beta = \Sigma_x^{-1} \Sigma_{xy}$ , where  $\Sigma_x = E[X_t X_t']$  and  $\Sigma_{xy} = E[X_t y_{t+h}]$ . Under the assumption of strong factors as given in Assumption D of their paper, i.e.,

$$0 < \liminf_{N \rightarrow \infty} \frac{1}{N} \lambda_{\min}(\Lambda' \Lambda) \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \lambda_{\max}(\Lambda' \Lambda) < \infty,$$

and under a Gaussian prior specification,  $\beta \sim N(0, \Phi_0)$ , where  $\|\Phi_0\|^{-1} = cNT^{\frac{1}{2}+\delta}$ , with  $0 < \delta < \frac{1}{2}$  and  $c$  being some positive constant, the authors provide the following consistency result:

$$\begin{aligned} X_T' \hat{\beta}^{\text{bay}} - y_{T+h|T}^* &= O_p(N^{-\frac{\alpha}{2}}) + O_p\left(\frac{1}{NT\|\Phi_0\|}\right) + O_p(N\sqrt{T}\|\Phi_0\|) \\ &= O_p(N^{-\frac{\alpha}{2}}) + O_p\left(\frac{1}{T^{\frac{1}{2}-\delta}}\right) + O_p\left(\frac{1}{T^\delta}\right) \text{ as } N, T \rightarrow \infty, \end{aligned}$$

where  $0 < \alpha \leq 1$  and where  $y_{T+h|T}^* = \gamma F_T$ . See Corollary 2 of their paper. Given that the fastest rate of convergence obtained in De Mol et al. (2008) for their procedure occurred under Assumption D, we shall compare their results with ours, focusing on the case where  $\lambda_1 \asymp N$ . In addition, we shall conduct our comparison without imposing any sparsity restrictions on the underlying model, assuming that the  $q = 0$  constraint sets given in expressions (49) and (50) are satisfied<sup>19</sup>. This is in accordance with their assumptions, as they do not impose sparsity. In this case,  $c_0(N) \asymp a(N) \asymp N$ , so we have

$$\begin{aligned} \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left(\frac{\ln N}{T}\right)^{\frac{1-q}{2}} &= \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \sqrt{\frac{\ln N}{T}} \asymp \sqrt{\frac{\ln N}{T}}, \\ \frac{\sqrt{a(N)N}}{\lambda_1} \left(\frac{\ln N}{T}\right)^{\frac{1}{2}(1-\frac{q}{2})} &= \frac{\sqrt{a(N)N}}{\lambda_1} \sqrt{\frac{\ln N}{T}} \asymp \sqrt{\frac{\ln N}{T}}, \\ \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}} &\asymp \frac{1}{N} \ll \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{\lambda_1}} \asymp \frac{1}{\sqrt{N}}, \text{ and } \frac{1}{\sqrt{T}} \ll \sqrt{\frac{\ln N}{T}}, \end{aligned}$$

---

<sup>19</sup>Note that it is appropriate here to focus on the  $q = 0$  case. This is because the actual rate of convergence of our procedure is given by the fastest rate of convergence across all different  $\ell_q$  constraint sets that are satisfied by the data generating process. In the case where we set  $c_0(N)$  and  $a(N)$  to be on the order of  $N$ , then the different constraint sets based on the different  $\ell_q$  quasi norms, for  $0 \leq q < 1$ , will all be satisfied even when there is no sparsity (exact or approximate). Therefore, in this case, the rate of convergence is given by that of the  $q = 0$  case since the  $q = 0$  case yields the most stringent constraint set, so that, when satisfied, it also provides the sharpest estimate of the actual rate of convergence.

from which it follows, by first focusing on the DRPLS estimator, that

$$\begin{aligned}
& \widehat{y}_{T+h}^{\text{drpls}} - \left( \alpha + [F_{1,T} - \mu_{F,1}]' \varphi_1 \right) \\
&= O_p \left( \max \left\{ \frac{c_0(N) \sqrt{N}}{\lambda_1^{\frac{3}{2}}} \left( \frac{\ln N}{T} \right)^{\frac{1-q}{2}}, \frac{\sqrt{a(N)N}}{\lambda_1} \left( \frac{\ln N}{T} \right)^{\frac{1}{2}(1-\frac{q}{2})}, \frac{\sqrt{N}}{\lambda_1^{\frac{3}{2}}}, \frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{T}} \right\} \right) \\
&= O_p \left( \max \left\{ \sqrt{\frac{\ln N}{T}}, \frac{1}{\sqrt{N}} \right\} \right).
\end{aligned}$$

On the other hand, since  $0 < \alpha \leq 1$  and since  $\|\Phi_0\|^{-1} = cNT^{\frac{1}{2}+\delta}$ , with  $0 < \delta < \frac{1}{2}$ , and with  $c$  being some positive constant, it is easy to see that one can optimize the rate of convergence of the Bayesian regression specified in De Mol et al. (2008) by taking  $\delta = 1/4$  and  $\alpha = 1$  to obtain

$$\begin{aligned}
X'_T \widehat{\beta}^{\text{bay}} - y_{T+h|T}^* &= O_p(N^{-\frac{\alpha}{2}}) + O_p\left(\frac{1}{NT\|\Phi_0\|}\right) + O_p(N\sqrt{T}\|\Phi_0\|) \\
&= O_p(N^{-\frac{\alpha}{2}}) + O_p\left(\frac{1}{T^{\frac{1}{2}-\delta}}\right) + O_p\left(\frac{1}{T^\delta}\right) \text{ as } N, T \rightarrow \infty, \\
&= O_p\left(\max\left\{\frac{1}{\sqrt{N}}, \frac{1}{T^{\frac{1}{4}}}\right\}\right).
\end{aligned}$$

Within this context, first consider the case where  $T \lesssim N$ . In this case

$$\begin{aligned}
\widehat{y}_{T+h}^{\text{drpls}} - (\alpha + [F_T - \mu_F]' \varphi) &= O_p\left(\sqrt{\frac{\ln N}{T}}\right), \\
X'_T \widehat{\beta}^{\text{bay}} - y_{T+h|T}^* &= O_p\left(\frac{1}{T^{\frac{1}{4}}}\right),
\end{aligned}$$

so that the point forecast based on the DRPLS estimator will have a faster rate of convergence than the point forecast based on the Bayesian regression as long as  $\sqrt{\ln N}/T^{\frac{1}{4}} \rightarrow 0$  or if  $N \ll \exp\{\sqrt{T}\}$ . Hence, when  $T \lesssim N$ , the DRPLS procedure will have a faster rate of convergence except in very extreme cases where  $N$  is exponentially larger than  $T$ .

Next, consider the case where  $N \ll T$ . This case can be divided into three subcases. First, consider the subcase where  $\sqrt{T} \ll \sqrt{N \ln N} \ll N \ll T$ . In this subcase, it is easily seen that

$$\begin{aligned}
\widehat{y}_{T+h}^{\text{drpls}} - (\alpha + [F_T - \mu_F]' \varphi) &= O_p\left(\max\left\{\sqrt{\frac{\ln N}{T}}, \frac{1}{\sqrt{N}}\right\}\right) = O_p\left(\sqrt{\frac{\ln N}{T}}\right) \\
X'_T \widehat{\beta}^{\text{bay}} - y_{T+h|T}^* &= O_p\left(\max\left\{\frac{1}{\sqrt{N}}, \frac{1}{T^{\frac{1}{4}}}\right\}\right) = O_p\left(\frac{1}{T^{\frac{1}{4}}}\right).
\end{aligned}$$

Moreover, in this case given that  $N \ll T$ , we have  $\sqrt{\ln N/T} = T^{-\frac{1}{4}} \left( \sqrt{\ln N}/T^{\frac{1}{4}} \right) \ll T^{-\frac{1}{4}}$ , so that the point forecast based on the DRPLS estimator has a faster rate of convergence than that based on the Bayesian estimator specified in De Mol et al. (2008).

Second, consider the subcase where  $\sqrt{N \ln N} \lesssim \sqrt{T} \ll N \ll T$ . In this subcase,

$$\begin{aligned} \hat{y}_{T+h}^{\text{drpls}} - (\alpha + [F_T - \mu_F]' \varphi) &= O_p \left( \max \left\{ \sqrt{\frac{\ln N}{T}}, \frac{1}{\sqrt{N}} \right\} \right) = O_p \left( \frac{1}{\sqrt{N}} \right) \\ X_T' \hat{\beta}^{\text{bay}} - y_{T+h|T}^* &= O_p \left( \max \left\{ \frac{1}{\sqrt{N}}, \frac{1}{T^{\frac{1}{4}}} \right\} \right) = O_p \left( \frac{1}{T^{\frac{1}{4}}} \right), \end{aligned}$$

and given that  $\sqrt{T} \ll N$ , we have  $1/\sqrt{N} \ll 1/T^{\frac{1}{4}}$ , so that again the point forecast based on the DRPLS estimator converges faster than that based on the Bayesian estimator specified in De Mol et al (2008).

Finally, consider the subcase where  $N \rightarrow \infty$ , such that  $N \lesssim \sqrt{T}$ . In this subcase,

$$\begin{aligned} \hat{y}_{T+h}^{\text{drpls}} - (\alpha + [F_T - \mu_F]' \varphi) &= O_p \left( \max \left\{ \sqrt{\frac{\ln N}{T}}, \frac{1}{\sqrt{N}} \right\} \right) = O_p \left( \frac{1}{\sqrt{N}} \right) \\ X_T' \hat{\beta}^{\text{bay}} - y_{T+h|T}^* &= O_p \left( \max \left\{ \frac{1}{\sqrt{N}}, \frac{1}{T^{\frac{1}{4}}} \right\} \right) = O_p \left( \frac{1}{\sqrt{N}} \right). \end{aligned}$$

Thus, for this case, both the point forecast based on the DRPLS estimator and that based on the Bayesian estimator have the same rate of convergence.

In summary, point forecasts based on the DRPLS estimator has a faster rate of convergence than those based on the Bayesian estimator proposed in De Mol et al. (2008), whenever  $\sqrt{T} \ll N \ll \exp \left\{ \sqrt{T} \right\}$ . On the other hand, point forecasts based on the DRPLS estimator will have same rate of convergence as those based on the Bayesian estimator when  $N \asymp \exp \left\{ \sqrt{T} \right\}$ , or in the cases where  $N \lesssim \sqrt{T}$ , so that only when  $N \gg \exp \left\{ \sqrt{T} \right\}$  will point forecasts based on the Bayesian estimator have faster rates of convergence than those based on our DRPLS estimator. Moreover, it is easy to see that using the same calculations, we can obtain the same set of results for the rates of convergence of point forecasts constructed using our GRPLS estimator.

## 7 Empirical Illustration

In this section, we report the results of a forecasting comparison exercise analyzing the monthly real-time macroeconomic FRED-MD dataset maintained by the St. Louis Federal Reserve Bank<sup>20</sup>. In our implementation, we follow the variable transformation recommendations outlined on the FRED-MD website, as well as removing all discontinued series, leaving

---

<sup>20</sup>See <https://www.stlouisfed.org/research/economists/mccracken/fred-databases>

us with a dataset containing 96 variables for the period spanning March 1973 to June 2024. The full list of all variables and their transformations is available upon request from the authors. It should be noted that the FRED-MD database is truly a real-time dataset. To provide an example illustrating what this means, consider the value of industrial production for January 2000 which is first reported by the government in February of 2000. Subsequently, the government updates its estimate of industrial production that was reported in February, so that in March, a “second release” of the January datum occurs. This process continues as we move forward in time. It follows that a real-time dataset contains a series of vintages where each vintage consists of all of the historical data that are available as of the calendar date of the vintage. Thus, a vintage of data mirrors what one would obtain by extracting a single vector time series at a specific point in time from the internet, say. If the same calendar-dated dataset were extracted at a later point in time, various data in the time-indexed vector might be revised, relative to the earlier vintage. The use of this type of data allows researchers to truly simulate a real time forecasting environment in which models are updated at each point in time using data that were actually available at the time. The dataset we use has 295 vintages with the first vintage having data spanning the period March 1973 to December 1999 and with the last vintage having (updated) data from March 1973 to June 2024. Further discussion about the structure of real-time datasets, as well as methods for real-time forecasting, can be found in Swanson (1996), Swanson and van Dijk (2006), and Kim and Swanson (2018).

Making use of this real-time dataset, we compare the forecast performance of the following models and methods:

1. **Autoregression (AR):** This serves as the benchmark model in our forecasting comparison. The number of lags is estimated using BIC or the Schwarz information criterion.
2. **Principal Components (or PC):** For this method, we specify a diffusion index model and estimate factors using conventional principal components analysis with the number of factors estimated using the  $PC_{p2}$  criterion proposed by Bai and Ng (2002). Forecasts are generated by plugging the estimated factors into the diffusion index forecast equation and estimating the forecast equation using OLS.
3. **PLS (unregularized):** Forecasts are generated by applying the original (unregularized) PLS procedure used for estimating a many-predictors forecast regression.<sup>21</sup>
4. **Regularized PLS:** As discussed previously, this paper presents eight different regularized PLS estimators, each of which we use to estimate a many-predictors forecast regression and generate point forecasts. We shall use the following acronyms to refer to the different versions of regularized PLS.

(a) **DRPLS1:** hard thresholding DRPLS

---

<sup>21</sup>Due to the same numerical issue as explained in part (b) of Remark 6.1, our (unregularized) PLS procedure is also calculated using a Moore-Penrose generalized inverse instead of an ordinary inverse.

- (b) **DRPLS2:** soft thresholding DRPLS
- (c) **DRPLS3:** SCAD DRPLS
- (d) **DRPLS4:** adaptive Lasso DRPLS
- (e) **GRPLS1:** hard thresholding GRPLS
- (f) **GRPLS2:** soft thresholding GRPLS
- (g) **GRPLS3:** SCAD GRPLS
- (h) **GRPLS4:** adaptive Lasso GRPLS

For each of the eleven methods listed above, we carry out  $h = 1$ -month ahead real-time forecasting, which requires us to retrain or re-estimate the forecast model using the most recent vintage of the data that is available at the time when each point forecast is made. We compare the different methods based on the accuracy of their one-month ahead forecasts of the following six target variables.

**Table 1: FRED-MD Target Variables Used in Forecasting Illustration**

| Target Variable                        | FRED-MD Abbreviation |
|----------------------------------------|----------------------|
| Civilian Employment                    | CE16OV               |
| Capacity Utilization: Manufacturing    | CUMFNS               |
| Housing Starts (new, privately owned)  | HOUST                |
| Industrial Production                  | INDPRO               |
| Housing Permits (new, privately owned) | PERMIT               |
| Civilian Unemployment Rate             | UNRATE               |

Note that in order to implement our regularized PLS procedures we need to estimate a number of tuning parameters. In particular, we select the value of the threshold parameter  $\eta_1$  for estimating  $\bar{\Sigma}_X$  and the value of the threshold parameter  $\eta_2$  for estimating  $\bar{\sigma}_{Xy}^{22}$ . Also, for DRPLS procedures, we select the value for the tuning parameter  $v$  (see expression (41)). Finally, for both DRPLS and GRPLS, we estimate the value of the tuning parameter  $a$  for the SCAD type estimator and the value of  $\zeta$  for the adaptive Lasso type estimator (see expressions (39) and (40)). For all PLS-type procedures (regularized or not), we need to estimate the envelope parameter  $\kappa$ . Our forecasting procedure estimates all of these tuning parameters jointly based on the most updated vintage of the data, prior to the construction of each new forecast. For example, suppose we wish to forecast Industrial Production for January 2000. In this case, we train our regularized PLS procedures using the most updated data vintage, which contains monthly data spanning the period from March of 1973 to December 1999. Now, for each variable in our dataset, this initial training+validation dataset has 322 data points, which we further partition into a training sample containing 222 data points with which we do initial estimation and a validation sample of 100 data points with which we carry out our main prediction experiment. We then choose our vector of tuning

---

<sup>22</sup>As noted before, we allow for the possibility that  $\hat{\Sigma}_{X,(i,j)}^S$  and  $\hat{\sigma}_{Xy,i}^S$  may have different values of the threshold function.

parameter values to be that vector value which leads to the smallest mean square forecast error, MSFE, over the validation sample.<sup>23</sup> Moving forward to forecast Industrial Production for February 2000, we then update our training+validation set using the vintage of data that is available one month forward and so on, keeping the size of the validation sample fixed at 100 observations. Finally, we evaluate forecast performance based on relative MSFE, i.e., the ratio  $MSFE/MSFE_{AR}$ , where  $MSFE_{AR}$  denotes the MSFE of the benchmark AR model. The numerator of this statistic is the MSFE of either PC, PLS (unregularized), or one of our regularized PLS procedures.

Our empirical findings are reported in Tables 2a and 2b below. Inspection of the results across the six target variables indicates that regularized PLS procedures finish in first place in terms of having the lowest relative MSFE in five out of the six cases, while all second place finishers are regularized PLS procedures. The only exception amongst first and second place finishers is the original (unregularized) PLS procedure which comes in first place for forecasting housing starts (HOUST). Drilling down a bit further, the soft thresholding versions of regularized PLS (DRPLS2 and GRPLS2) seem to do well as a group, since together they finish in first place twice and in second place four times. A bit behind them are the hard thresholding versions of regularized PLS (DRPLS1 and GRPLS1), which as a group, finish in first place twice and in second place once. It should also be noted that across all of our results, GRPLS3 (a SCAD version of our regularized PLS procedure using the generalized inverse) achieves the smallest relative MSFE when forecasting housing permits (PERMIT). Finally, note that no method uniformly beats all competing methods across all target variables studied here. This should not be surprising when one applies econometric methods to forecast real world data. In summary, we provide evidence that the regularized PLS procedures proposed in the paper can be useful for FRED-MD forecasting and, hence, should be added to the “toolbox” of methods that practitioners use when making time series forecasts.

---

<sup>23</sup>More precisely, this model selection step involves minimizing MSFE with respect to the parameters  $\eta_1$ ,  $\eta_2$ ,  $v$ ,  $a$ ,  $\zeta$ , and  $\kappa$  over the following range of possible values for the tuning and envelope parameters:

$$\begin{aligned}
\Theta_{\eta_1} &= \{0.00001, 0.00002, 0.00004, 0.00006, 0.00008, 0.0001, 0.0002, 0.0004, 0.0006, 0.0008, \\
&\quad 0.001, 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8\}, \\
\Theta_{\eta_2} &= \{0.00001, 0.00002, 0.00004, 0.00006, 0.00008, 0.0001, 0.0002, 0.0004, 0.0006, 0.0008, \\
&\quad 0.001, 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8\}, \\
\Theta_v &= \{5, 10, 15\}, \\
\Theta_a &= \{2.2, 2.7, 3.2, 3.7, 4.2, 4.7, 5.2\}, \\
\Theta_\zeta &= \{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}, \\
\Theta_\kappa &= \{1, 2, 3, 4, 5\}.
\end{aligned}$$

**Table 2a: Forecast Mean-Square Errors for Select FRED-MD Variables**

|        | MSFE/MSFE <sub>AR</sub> | $h = 1$   |           |
|--------|-------------------------|-----------|-----------|
|        | CE16OV                  | CUMFNS    | HOUST     |
| PC     | 1.0238777               | 1.0221386 | 0.9060513 |
| PLS    | 1.0324039               | 1.0075549 | 0.8199532 |
| DRPLS1 | 0.9853878               | 0.9643379 | 0.8556645 |
| DRPLS2 | 1.0100628               | 1.0229169 | 0.8360654 |
| DRPLS3 | 1.0350322               | 0.9999052 | 0.8371145 |
| DRPLS4 | 1.0197774               | 0.9580890 | 0.8639672 |
| GRPLS1 | 0.9854965               | 0.9463483 | 0.8557394 |
| GRPLS2 | 1.0316492               | 1.1634302 | 0.8361134 |
| GRPLS3 | 1.0023393               | 1.2370992 | 0.8490366 |
| GRPLS4 | 1.0032172               | 1.2524952 | 0.8509814 |

**Table 2b: Forecast Mean-Square Errors for Select FRED-MD Variables**

|        | MSFE/MSFE <sub>AR</sub> | $h = 1$   |           |
|--------|-------------------------|-----------|-----------|
|        | INDPRO                  | PERMIT    | UNRATE    |
| PC     | 0.9916734               | 0.8968817 | 0.9459886 |
| PLS    | 1.0510325               | 0.8069158 | 0.8645720 |
| DRPLS1 | 0.9986302               | 0.8106202 | 0.8590605 |
| DRPLS2 | 0.9188996               | 0.8064706 | 0.8440758 |
| DRPLS3 | 0.9499646               | 0.8672492 | 0.9541491 |
| DRPLS4 | 0.9637669               | 0.8134856 | 0.8769876 |
| GRPLS1 | 1.0087073               | 0.8106235 | 0.8606016 |
| GRPLS2 | 0.9258872               | 0.8064655 | 0.8440506 |
| GRPLS3 | 1.0365507               | 0.7933287 | 0.9019084 |
| GRPLS4 | 1.0653416               | 0.8153741 | 5.3099634 |

## 8 Conclusion

In this paper, we consider time series forecasting using a many-predictors regression model that has an approximate envelope structure. Our framework allows us to perform dimension reduction in a supervised or targeted manner, taking into consideration possible relationships between linear combinations of the predictor variables and the target variable of interest. Our approximate envelope framework generalizes the exact envelope framework of Cook, Li, and Chiaromonte (2010) and Cook, Helland, and Su (2013) to allow for situations where the vector of predictor variables  $X_t$  is generated by an approximate factor model whose error component exhibits general cross-sectional dependence. This generalization is of practical importance for applications in economics since an exact factor model whose error term does not exhibit cross-sectional dependence is usually considered to be too restrictive for applications in macroeconomic and financial data analysis. Moreover, our framework allows the

data to exhibit general temporal dependence and heterogeneity in the second moments, both of which are often important in the modeling of economic time series. Given this approximate envelope framework, we establish an interesting connection between the diffusion index model and a many-predictors regression model. In particular, we show that the conditional mean function of a diffusion index model can be well approximated asymptotically by the conditional mean function of a many-predictors regression with an approximate envelope structure, given the additional assumption that coefficient vector of the many-predictors forecasting regression is modeled as being in a  $1/\sqrt{\lambda_{1,1}}$  neighborhood of zero. This and other results given in the paper help to provide insights as to why PLS-type procedures can be effective in recovering the conditional mean function of a diffusion index type model, even though, being a least squares type procedure, one might à priori expect PLS to only be useful for estimating regression models for which the right-hand side variables are observed.

Motivated by these results, we introduce regularized versions of the PLS estimator. More specifically, we propose two different classes of regularized PLS estimators: one is referred to as the class of doubly regularized PLS (or DRPLS) estimators and the other is called the class of generalized-inverse-based PLS (or GRPLS) estimators. Both DRPLS and GRPLS are easy to apply to real world data as they address problems associated with singularity or high condition number that can arise when PLS-type estimators based on the ordinary inverse are used in a model selection step to estimate the envelope parameter  $\kappa$ . More importantly, our regularized PLS estimators are adaptive to alternative statistical environments because they are able to not only take advantage of the possibility for dimension reduction due to the presence of a factor structure but can also make use of any availability of parameter sparsity that may be present in the underlying data structure. We prove the consistency of both the DRPLS estimator and the GRPLS estimator and also show that they have the same rate of convergence. In addition, we show that point forecasts constructed from either the DRPLS or the GRPLS estimator can consistently estimate the conditional mean function of the underlying forecast equation regardless of whether the true forecast equation happens to be a many-predictors regression with an envelope structure or a diffusion index model. In an empirical illustration of our methodology using the well-known FRED-MD dataset, we provide some encouraging results indicating good forecast performance when our regularized PLS procedures are compared with PCA, (unregularized) PLS, and a benchmark AR model.

## References

- [1] Ahn, S. C. and J. Bae (2022): “Forecasting with Partial Least Squares When a Large Number of Predictors Are Available,” Working Paper, Arizona State University and University of Glasgow.
- [2] Bai, J. and S. Ng (2023): “Approximate Factor Models with Weaker Loading,” *Journal of Econometrics*, 235, 1893-1916.

- [3] Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.
- [4] Bickel, P. J. and E. Levina (2008): "Covariance Regularization by Thresholding," *Annals of Statistics*, 36, 2577-2604.
- [5] Carrasco, M. and B. Rossi (2016): "In-Sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business & Economic Statistics*, 34, 313-338.
- [6] Cook, R.D. (2018): *An Introduction to Envelopes: Dimension Reduction to Efficient Estimation in Multivariate Statistics*. John Wiley & Sons, Hoboken, NJ.
- [7] Cook, R.D. and L. Forzani (2019): "Partial Least Squares Prediction in High-Dimensional Regression," *Annals of Statistics*, 47, 884-908.
- [8] Cook, R.D., I.S. Helland, and Z. Su (2013): "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, Series B*, 75, 851-877.
- [9] Cook, R.D., B. Li, and F. Chiaromonte (2010): "Envelope Models for Parsimonious and Efficient Multivariate Linear Regression," *Statistica Sinica*, 20, 927-1010.
- [10] De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?" *Journal of Econometrics*, 146, 318-328.
- [11] Fagiolo, G., M. Napoletano, and A. Roventini (2008): "Are Output Growth-Rate Distributions Fat-Tailed? Some Evidence from OECD Countries," *Journal of Applied Econometrics*, 23, 639-669.
- [12] Fan, J. and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.
- [13] Fan, J., Y. Liao, and M. Mincheva (2011): "High-dimensional Covariance Matrix Estimation in Approximate Factor Models," *Annals of Statistics*, 39, 3320-3356.
- [14] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004): "The Generalized Dynamic Factor Model, Consistency and Rates," *Journal of Econometrics*, 119, 231-255.
- [15] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): "The Generalized Dynamic Factor Model, One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830-840.
- [16] Giannone, D., M. Lenza, and G.E. Primiceri (2021): "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409-2437.
- [17] Groen, J. and G. Kapetanios (2016): "Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting," *Computational Statistics & Data Analysis*, 100, 221-239.

- [18] Helland, I. S. (1988): "On the Structure of Partial Least Squares Regression," *Communications in Statistics - Simulation and Computation*, 17, 581-607.
- [19] Helland, I. S. (1990): "Partial Least Squares Regression and Statistical Models," *Scandinavian Journal of Statistics*, 17, 97-114.
- [20] Merlevède, F., M. Peligrad, and E. Rio (2011): "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Processes," *Probability Theory and Related Fields*, 151, 435-474.
- [21] Rothman, A. J., E. Levina, J. Zhu (2009): "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177-186.
- [22] Samadi, S. Y. and H. M. W. B. Herath (2024): "Reduced-Rank Envelope Vector Autoregressive Model," *Journal of Business and Economic Statistics*, 42, 918-932.
- [23] Samadi, S. Y. and T. P. De Alwis (2025): "Envelope Matrix Autoregressive Models," *Journal of Business and Economic Statistics*, 44, 397-412.
- [24] Stock, J. H. and M. W. Watson (1998): "Diffusion Indexes," Working Paper 6702, NBER.
- [25] Stock, J. H. and M. W. Watson (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.
- [26] Stoica, P. and T. Söderström (1998): "Partial Least Squares: A First-order Analysis," *Scandinavian Journal of Statistics*, 25, 17-24.
- [27] Wold, H. (1982): "Soft Modeling, the Basic Design and Some Extensions." In *Systems Under Indirect Observation: Causality-Structure-Prediction, Part II* (H. Wold and K.-G. Jöreskog, eds.). North-Holland, Amsterdam, 1-54.
- [28] Wold, H. (1985): "Partial Least Squares." In *Encyclopedia of Statistical Sciences, Vol. 6* (S. Kotz and N. L. Johnson, eds.). Wiley, New York, 581-591.
- [29] Wold, S., H. Martens, and H. Wold (1983): "The Multivariate Calibration Problem in Chemistry Solved by the PLS Method." In *Proceedings of the Conference on Matrix Pencils* (A. Ruhe and B. Kagström, eds.). *Lecture Notes in Mathematics*, Springer Verlag, Heidelberg, 286-293.
- [30] Zou, H. (2006): "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418-1429.