Hybrid Regularisation of Functional Linear Models

ANIRVAN CHAKRABORTY and VICTOR M. PANARETOS

Institut de Mathématiques,
École Polytechnique Fédérale de Lausanne
E-mail: anirvan.chakraborty@epfl.ch; victor.panaretos@epfl.ch

We consider the problem of estimating the slope function in a functional regression with a scalar response and a functional covariate. This central problem of functional data analysis is well known to be ill-posed, thus requiring a regularised estimation procedure. The two most commonly used approaches are based on spectral truncation or Tikhonov regularisation of the empirical covariance operator. In principle, Tikhonov regularisation is the more canonical choice. Compared to spectral truncation, it is robust to eigenvalue ties, while it attains the optimal minimax rate of convergence in the mean squared sense, and not just in a concentration probability sense. In this paper, we show that, surprisingly, one can strictly improve upon the performance of the Tikhonov estimator in finite samples by means of a linear estimator, while retaining its stability and asymptotic properties by combining it with a form of spectral truncation. Specifically, we construct an estimator that additively decomposes the functional covariate by projecting it onto two orthogonal subspaces defined via functional PCA; it then applies Tikhonov regularisation to the one component, while leaving the other component unregularised. We prove that when the covariate is Gaussian, this hybrid estimator uniformly improves upon the MSE of the Tikhonov estimator in a non-asymptotic sense, effectively rendering it inadmissible. This domination is shown to also persist under discrete observation of the covariate function. The hybrid estimator is linear, straightforward to construct in practice, and with no computational overhead relative to the standard regularisation methods. By means of simulation, it is shown to furnish sizeable gains even for modest sample sizes.

Keywords: admissibility, condition index, functional data analysis, ill-posed problem, mean integrated squared error, principal component analysis, rate of convergence, ridge regression, spectral truncation, Tikhonov regularisation.

1. Introduction

1.1. Functional Linear Models and their regularisation

For a real-valued response y, and a random functional covariate X taking values in a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, the functional linear regression model with scalar response is given by

$$y = \alpha + \langle X, \beta \rangle + \epsilon, \tag{1.1}$$

where ϵ is a scalar random measurement error term that is assumed to be zero mean and independent of the covariate X (Ramsay and Silverman (2005), Horváth and Kokoszka (2012), Hsing and Eubank (2015)). The so-called slope parameter $\beta \in \mathcal{H}$ is typically the object of primary importance. The statistical task is then to estimate β on the basis of an i.i.d. sample of pairs $\{(y_i, X_i)\}_{i=1}^n$ generated according to the model (1.1). The classical least squares approach of estimating β results in the normal equation

$$\hat{\mathscr{K}}\, \beta = \hat{C}, \tag{1.2}$$

where $\hat{\mathscr{K}}$ is the empirical covariance operator of the $\{X_i\}$ and $\hat{C}$ is the empirical cross-covariance of the $\{y_i\}$ and the $\{X_i\}$. Since the population operator $\mathscr{K}$ is a trace-class operator, its empirical version $\hat{\mathscr{K}}$ is so too, for all n. Its failure to be boundedly invertible gives rise to an ill-posed inverse problem, which is

usually solved by regularising the inverse of $\hat{\mathscr{K}}$. The regularisation strategies employed in the functional data analysis literature can be broadly categorised into two classes[1]: *sieve methods* and *penalised methods*.

In the method of sieves (Grenander (1981)), one selects an orthonormal basis $\{\varphi_k\}$ of $\mathcal{H}$, and projects the covariates $\{X_i\}$ and the slope function β onto the subspace spanned by the first r basis elements. If the basis $\{\varphi_k\}$ and truncation level r are selected judiciously, one obtains a stable multivariate regression problem, with a small amount of bias. In terms of asymptotics, one must let $K \to \infty$ but regulate its growth as a function of n, in order to guarantee that the regression problem remain stable for each n. The challenge here is to determine a "good" basis $\{\varphi_k\}$ whose first r elements provide a parsimonious representation simultaneously for X *and* β – but of course β is unknown, and worse still, does not have any intrinsic relationship to X that might prove the existence of such a basis. The typical choice is to rely on the Karhunen-Loève expansion of X and to choose $\{\varphi_k\}$ to be the basis of eigenfunctions of $\mathscr{K}$, and has evolved in the most popular choice in practice. This approach, known as *spectral truncation* or *PCA regression*, uses a sieve that is optimally adapted to X, but makes no reference to β, thus potentially not providing a good approximation of β. The thought is, however, that those characteristics of β that do not correlate well with X (and thus are not well expressed in the Karhunen-Loève basis) are worth sacrificing, as they cannot be well-recovered through the model (1.1) anyway. In practice, the Karhunen-Loève basis is estimated from the data, using a functional principal component analysis (Ramsay and Silverman (2005, Chapter 10); Ferraty and Vieu (2000); Cuevas, Febrero and Fraiman (2002); Cardot and Sarda (2006); Yao, Muller and Wang (2005)).

Penalised methods, on the other hand, regularise the problem by placing restrictions directly on β, most often by penalising the degree of roughness of β by means of a suitable norm. They lead to constrained least squares problems, instead of the unconstrained problem (1.2), that are well posed. Estimation procedures following this paradigm have been studied, for instance, by Ramsay and Dalzell (1991), Marx and Eilers (1999), Cardot, Ferraty and Sarda (2003), Li and Hsing (2007) and Crambes, Kneip and Sarda (2009) to name only a few (see also Ramsay and Silverman (2005)). Depending on the nature of the penalty, the estimator can be represented in a finite a basis $\{\varphi_k\}$, typically a spline basis corresponding to a curvature penalty, and the functional regression translates to a multivariate ridge regression problem. The approach can be elegantly formulated within a reproducing kernel Hilbert space framework, which directly translates the infinite dimensional and ill-posed problem into a finite dimensional and well-posed one (Yuan and Cai (2010)). A general description of penalised methods can be viewed as instances of *Tikhonov regularisation* (Tikhonov and Arsenin), where the sum of squares objective function $SS(\beta) = \sum_{i=1}^{n}(y_i - \bar{y} - \langle X_i - \bar{X}, \beta\rangle)^2$ is penalised by the addition of a multiple of some norm of β.

1.2. Tikhonov vs Spectral regularisation and Our Contributions

For a regularisation parameter $\rho > 0$, the Tikhonov regularised estimator is defined as

$$\hat{\beta}_{TR} = \hat{\mathscr{K}}_{\rho}^{-1}\hat{C}$$

where $\hat{\mathscr{K}}_{\rho} = \hat{\mathscr{K}} + \rho\mathscr{I}$ with $\mathscr{I}$ the identity operator on $\mathcal{H}$. This estimator is the direct analogue of the ridge estimator (Hoerl and Kennard (1970)) in classical multivariate linear regression with correlated regressors and is the minimiser of the penalized least squares problem

$$\min_{\beta\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y} - \langle X_i - \bar{X}, \beta\rangle)^2 + \rho\|\beta\|^2\right\}. \tag{1.3}$$

[1]Though there exist even more general descriptions that include the two categories as special cases, see Cardot, Mas and Sarda (2007).

On the other hand, given $r \in \mathbb{N}$, the spectral truncation estimator is defined as

$$\hat{\beta}_{ST} = \left(\sum_{j=1}^{r} \hat{\lambda}_j^{-1} \hat{\phi}_j \otimes \hat{\phi}_j \right) \hat{C},$$

where $\{\hat{\lambda}_j, \hat{\phi}_j\}$ is the spectrum of the empirical covariance $\hat{\mathscr{K}}$. In one of the landmark papers on functional regression, Hall and Horowitz (2007) established general error properties of both spectral truncation and of Tikhonov regularisation. Their results indicate the latter approach to be a more canonical avenue mainly due to two reasons, namely its minimaxity and its stability. More specifically, Hall and Horowitz (2007) established minimax optimal rates of convergence for the Tikhonov estimator in the mean squared error (MSE) sense, but only obtained minimax rates for spectral truncation estimator in the weaker concentration probability sense[2] (see also Remark 4). Secondly, Hall and Horowitz (2007) demonstrated that the spectral truncation approach can suffer from instabilities when the eigenvalues of $\mathscr{K}$ are not well-spaced, while Tikhonov regularisation is immune to such effects (see also Yuan and Cai (2010) for similar arguments). In the well-spaced regime, neither approach is seen to dominate the other, except in the rather special circumstance where:

(i) the leading eigenvalues of $\mathscr{K}$ are well-conditioned,
and
(ii) β mostly contained in the span of the leading eigenfunctions of $\mathscr{K}$.

When (i) and (ii) occur simultaneously, spectral truncation prevails, as the problem essentially reduces to a well-conditioned multivariate regression, in no need of regularisation.

The question this paper considers is the following: is it possible to leverage this last observation in order to improve upon the more canonical Tikhonov approach, by combining it in part with the projection rationale of spectral truncation? The answer is an unequivocal *yes*, and surprisingly the improvement is realised by a *linear* estimator: a simple combination of the two approaches yields a hybrid estimator that remains linear, straightforward to compute, and provably strictly improves upon the Tikhonov estimator in a non-asymptotic sense (i.e. not a rate but an exact MSE sense). We note in passing that while adaptive estimators of the slope function have been considered (see e.g. Cardot and Johannes (2010), Comte and Johannes (2012)), they typically introduce a thresholding of the spectral estimator, thus becoming non-linear (and, there has not been any theoretical comparison of the non-asymptotic MSEs of these estimators to those of the Tikhonov or spectral truncation estimator).

We are not aware of any other work in the functional data analysis literature where a regularized estimator has been shown to be inadmissible by virtue of the existence of another regularized estimator (indeed, strict inadmissibility results are quite rare even in the broader statistical literature). In the high dimensional multivariate setting, one may draw a parallel with the inadmissibility of the James-Stein regularized positive part estimator by the regularized estimator proposed by Shao and Strawderman (1994). The domination of the Tikhonov estimator by the linear hybrid estimator is also surprising since in the multivariate setup, it is well known that the classical Tikhonov (ridge regression) estimator is in fact a generalized Bayes estimator under squared error loss with Gaussian errors and an appropriate Gaussian prior for the slope vector; it would therefore seem that the result is an intrinsically functional effect.

[2] In an earlier paper, Hall and Hosseini-Nasab (2006) proved that a *modified, non-linear (thresholded)* version of the spectral truncation estimator *can* attain the minimax MSE rate. The modification is done to ensure that the resulting estimator does not take very large values (see Theorem 5 in Appendix A.2 in Hall and Hosseini-Nasab (2006), pp. 116–117 in that paper, and the discussion after Theorem 1 in Hall and Horowitz (2007)). Unfortunately, this modified estimator depends on arbitrary constants whose choices are subjective and so the estimator is not practically feasible. In fact, it remains unknown whether the original spectral truncation estimator attains the minimax rate of convergence in the mean squared sense at all. The non-linear modification appears to be necessary for the proof techniques of Hall and Horowitz (2007) to work.

The hybrid estimator we introduce (defined rigorously in Section 3) projects onto a finite dimensional subspace $\mathcal{H}_r$ (as would a sieve estimator), but rather than discard the residual component (i.e. the projection onto the orthogonal complement $\mathcal{H}_r^{\perp}$), it retains it, and applies a ridge regularisation to that and only that. The dimension $r < \infty$ of $\mathcal{H}_r$ does not grow with respect to n, and only the ridge parameter ρ is sample-size dependent. We demonstrate in Section 3 that choosing $\mathcal{H}_r$ to be *any* eigenspace of $\mathscr{K}$ of dimension $r < n$ yields an estimator that attains the minimax MSE rate but in fact strictly improves upon the Tikhonov approach for large enough samples, uniformly over β (Theorem 1 and Corollary 1). Section 4.1 exploits this observation in order to construct a practically feasible hybrid estimator, by empirical construction of $\mathcal{H}_r$. Section 4.2 establishes that the empirically constructed estimator also yields the same strict improvement and rates. The practicalities of constructing the estimator are discussed in detail Subsection 4.3, where recommendations are given on how to best choose the dimension r of $\mathcal{H}_r$ as well as the ridge parameter ρ. The key message is that one ought to select r so that the first r eigenvalues yield a mild condition index (λ_1/λ_r). Section 5 then treats the case where the covariates are functions observed discretely on a grid, and shows that even in this case, the hybrid estimator still enjoys the same properties and improvement in mean squared error over the Tikhonov estimator in this setup. In Section 6, we conduct a simulation study that illustrates that one can make considerable performance gains in practice, even for moderate sample sizes. The proofs of our formal results are collected in Section 7. First, though, Section 2 introduces some notation that will be employed throughout of the paper.

2. Preliminaries

As mentioned in the introduction, $\mathcal{H}$ will be a real separable Hilbert space, assumed infinite dimensional, with inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, and induced norm $\| \cdot \| : \mathcal{H} \to [0, \infty)$. Given a linear operator $\mathscr{A} : \mathcal{H} \to \mathcal{H}$, we will denote its adjoint operator by $\mathscr{A}^*$, its Moore-Penrose generalised inverse by $\mathscr{A}^-$, and its inverse by $\mathscr{A}^{-1}$, provided the latter is well-defined. The operator, Hilbert-Schmidt, and nuclear norms will respectively be

$$|||\mathscr{A}|||_\infty = \sup_{\|h\|=1} \|\mathscr{A}h\|, \quad |||\mathscr{A}||| = \sqrt{\operatorname{trace}(\mathscr{A}^*\mathscr{A})}, \quad |||\mathscr{A}|||_1 = \operatorname{trace}\left(\sqrt{\mathscr{A}^*\mathscr{A}}\right).$$

It is well-known that

$$|||\mathscr{A}|||_\infty \leq |||\mathscr{A}||| \leq |||\mathscr{A}|||_1$$

for any bounded linear operator satisfying $|||\mathscr{A}|||_1 < \infty$. The identity operator on $\mathcal{H}$ will be denoted by $\mathscr{I}$. For a pair of elements $f, g \in \mathcal{H}$, the tensor product $f \otimes g : \mathcal{H} \to \mathcal{H}$ will be defined as the linear operator

$$(f \otimes g)u = \langle g, u \rangle f, \qquad u \in \mathcal{H}.$$

The same notation will be used to denote the tensor product between two operators, so that for operators $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{G}$, one has

$$(\mathscr{A} \otimes \mathscr{B})\mathscr{G} = \operatorname{trace}(\mathscr{B}^*\mathscr{G})\,\mathscr{A}.$$

Given an estimator δ of the slope parameter $\beta \in \mathcal{H}$, we define the *Mean Square Error (MSE)* in order to probe the performance of δ,

$$\mathrm{MSE}(\delta) = \mathbb{E}\|\delta - \beta\|^2 = |||\mathbb{E}\{(\delta - \beta) \otimes (\delta - \beta)\}|||_1.$$

In the usual setting of $\mathcal{H} = L_2[0,1]$, this risk function reduces to the so-called *Mean Integrated Squared Error*,

$$\mathbb{E}\left\{\int_0^1 (\delta(x) - \beta(x))^2 \mathrm{d}x\right\},$$

but of course our results will be valid for any separable Hilbert space $\mathcal{H}$. We also note that all our results hold verbatim if instead of the MSE, we consider the (weaker) Hilbert-Schmidt norm of $\mathbb{E}\{(\delta-\beta)\otimes(\delta-\beta)\}$ as the risk function.

3. Motivation: Multivariate Plus Functional Regressors

To motivate our hybrid estimator, let $X = Y + Z$, where:

1. The covariance operator $\mathscr{K}_1$ of Y is of finite rank r.
2. The random elements Y and Z are uncorrelated.
3. The eigenspaces of $\mathscr{K}_1$ are orthogonal to those of the covariance $\mathscr{K}_2$ of Z.

Observe that such a decomposition always exists by the Karhunen-Loève theorem. The heuristic now is that if Y and Z were observable, we would have a model with two orthogonal and uncorrelated regressors, one multivariate, and one functional,

$$y = \langle X, \beta\rangle + \epsilon = \langle Y, \beta_1\rangle + \langle Z, \beta_2\rangle + \epsilon,$$

with β_1 and β_2 being the projections of β on the (orthogonal) ranges of $\mathscr{K}_1$ and $\mathscr{K}_2$. So, if Y has a well-conditioned covariance $\mathscr{K}_1$, then instead of regularising the entire spectrum of the covariance operator $\mathscr{K}$ of X, one should carry out two separate regressions: a multivariate one, without regularisation, corresponding to the well-conditioned $\mathscr{K}_1$; and a functional one, with Tikhonov regularisation, corresponding to the ill-conditioned $\mathscr{K}_2$. The point here is that functional regression is not ill-conditioned as a result of poor design (as in the multivariate case when covariates may be correlated); it is ill-conditioned by the mere fact that it is infinite dimensional. But, in general, we should be able to extract a subspace on which it is well-conditioned.

We now turn to transforming our heuristic to a concrete result. Write the spectra of the two covariance operators $\mathscr{K}_1$ and $\mathscr{K}_2$ as

$$\mathscr{K}_1 = \sum_{j=1}^{r} \lambda_{j1}\phi_{j1}\otimes\phi_{j1} \qquad \& \qquad \mathscr{K}_2 = \sum_{j\geq 1} \lambda_{j2}\phi_{j2}\otimes\phi_{j2}.$$

Define

$$\beta_1 = \left(\sum_{j=1}^{r} \phi_{j1}\otimes\phi_{j1}\right)\beta = \mathscr{P}_1\beta \qquad \& \qquad \beta_2 = \left(\sum_{j\geq 1} \phi_{j2}\otimes\phi_{j2}\right)\beta = \mathscr{P}_2\beta$$

to be the projections of β into the eigenspaces of $\mathscr{K}_1$ and $\mathscr{K}_2$. Note that we must have $\beta = \beta_1 + \beta_2$ for identifiablity so we henceforth assume that range$(K) = \mathcal{H}$. Thus, $\mathscr{P}_1 + \mathscr{P}_2 = \mathscr{I}$, where $\mathscr{I}$ is the identity operator on $\mathcal{H}$, and indeed $\langle X, \beta\rangle = \langle Y, \beta_1\rangle + \langle Z, \beta_2\rangle$. Now consider the following modification of the population version of the Tikhonov penalised least squares problem

$$\min_{\beta_1,\beta_2\in\mathcal{H}} \left\{\mathbb{E}\big[y - \mathbb{E}[y] - \langle Y - \mathbb{E}[Y], \beta_1\rangle - \langle Z - \mathbb{E}[Z], \beta_2\rangle\big]^2 + \rho\|\beta_2\|^2\right\}, \tag{3.1}$$

where we only penalize the part of the norm of β that corresponds to β_2. Direct calculation in the above minimisation problem yields the unique minimiser

$$\beta_{min} = \mathscr{K}_1^{-}C_1 + \mathscr{K}_{\rho,2}^{-}C_2,$$

where

$$C_1 = \mathbb{E}[yY] - \mathbb{E}[y]\mathbb{E}[Y], \quad C_2 = \mathbb{E}[yZ] - \mathbb{E}[y]\mathbb{E}[Z], \quad \mathscr{K}_{\rho,2} = \mathscr{K}_2 + \rho\mathscr{P}_2.$$

The form of the minimiser β_{min} motivates the following definition of a hybrid regularised estimator of β in the oracle case. Assume that a sample (y_i, X_i) is available, and the oracle reveals the decompositions $X_i = Y_i + Z_i$ into uncorrelated orthogonal components, as well as their respective covariances $(\mathscr{K}_1, \mathscr{K}_2)$. Define a hybrid estimator as

$$\widetilde{\beta}_{HR} = \mathscr{K}_1^{-}\widetilde{C}_1 + \mathscr{K}_{\rho,2}^{-}\widetilde{C}_2, \tag{3.2}$$

where

$$\begin{aligned}\widetilde{C}_1 &= n^{-1}\sum_{i=1}^{n}(y_i - \overline{y})(Y_i - \overline{Y}), \qquad \text{with } \overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i \\ \widetilde{C}_2 &= n^{-1}\sum_{i=1}^{n}(y_i - \overline{y})(Z_i - \overline{Z}), \qquad \text{with } \overline{Z} = n^{-1}\sum_{i=1}^{n} Z_i.\end{aligned}$$

On the other hand, the oracle version of the Tikhonov estimator is

$$\widetilde{\beta}_{TR} = \mathscr{K}_\rho^{-1}\hat{C}, \tag{3.3}$$

where $\mathscr{K}_\rho = \mathscr{K} + \rho\mathscr{I}$ and $\hat{C} = n^{-1}\sum_{i=1}^{n}(y_i - \overline{y})(X_i - \overline{X})$. Our first theorem shows that, since the hybrid estimator makes explicit use of the additional information (the decomposition (Z_i, Y_i) instead of just X_i), it *improves* upon the Tikhonov estimator.

Theorem 1. *Let $X = Y + Z$, where Y and Z are uncorrelated random elements with $\mathbb{E}(\|Y\|^4) < \infty$ and $\mathbb{E}(\|Z\|^4) < \infty$. Assume that the eigenspaces of the respective covariances $\mathscr{K}_1$ and $\mathscr{K}_2$ of Y and Z are orthogonal. Further, assume that the $\langle X, \phi_j\rangle$'s are independent, where ϕ_j's are the eigenfunctions of $\mathscr{K}$. Then,*

(a) For any fixed $\rho > 0$, $\mathrm{MSE}(\widetilde{\beta}_{TR}) > \mathrm{MSE}(\widetilde{\beta}_{HR})$ for all sufficiently large n.
(b) If we choose $\rho = \rho(n) \sim cn^{-\gamma}$ for some $\gamma \in (0, 1/2]$ and a constant $c > 0$, we have

$$n^{2\gamma}\{\mathrm{MSE}(\widetilde{\beta}_{TR}) - \mathrm{MSE}(\widetilde{\beta}_{HR})\} > B(n) + o(1).$$

Here, $B(n)$ converges to a positive constant if at least one of $\langle \beta, \phi_{j1}\rangle$, $j = 1, 2, \ldots, r$ is non-zero, else it converges to zero as $n \to \infty$.

The independence assumption in the above theorem obviously holds for Gaussian processes, and for any process whose Karhunen-Lòeve expansion has independent coefficients. It can be relaxed to requiring that $\mathbb{E}(\prod_{u=1}^{4}\langle X, \phi_{j_u}\rangle^{l_u}) = \prod_{u=1}^{4}\mathbb{E}(\langle X, \phi_{j_u}\rangle^{l_u})$ for l_u's satisfying $1 \le l_u \le 4$ and $\sum_{u=1}^{4} l_u \le 4$. This can be viewed as a "pseudo-independence" condition, and similar assumptions have been considered for analysis of high-dimensional data (see, e.g., Sec. 3 in Chen and Qin (2010), Sec. 4 in Bai and Saranadasa (1996)). As a direct consequence of part (b) of the above theorem, we have the corollary:

Corollary 1. *Under the conditions of Theorem 1 and in the setup of part (b) of that theorem, if at least one of $\langle \beta, \phi_{j1}\rangle$, $j = 1, 2, \ldots, r$ is non-zero, there exists a constant $c_0 > 0$ such that*

$$\mathrm{MSE}(\widetilde{\beta}_{TR}) - \mathrm{MSE}(\widetilde{\beta}_{HR}) > c_0 n^{-2\gamma}$$

for all sufficiently large n. If $\langle \beta, \phi_{j1}\rangle$ is uniformly zero for $1 \le j \le r$, the two MSE norms are asymptotically equal.

Thus, in the oracle case, as long as β is at least partially expressed by the principal components of Y, then the hybrid regularisation estimator will improve on the Tikhonov estimator – whether ρ is held fixed, or allowed to decay polynomially in n, as one usually does. The next section deals with carrying over this improvement to an empirically feasible estimator.

4. The Hybrid Estimator

In practice, the components Y and Z are unobservable, and their covariance operators $\mathscr{K}_1$ and $\mathscr{K}_2$ are unknown. Still, we can replace them by their empirical versions, and consider whether we can still improve upon the Tikhonov estimator by the hybrid approach when doing so. We will focus on the case where Y is the projection of X onto its first r principal components, and $Z = X - Y$, since this case admits straightforward empirical versions of all the quantities involved. We first define the empirical version of the hybrid estimator (Subsection 4.1); next we establish its superiority to Tikhonov regularisation (Subsection 4.2); and then, we discuss its (straightforward) practical implementation (Subsection 4.3).

4.1. Definition

Given an i.i.d. sample $X_1, \dots, X_n$ distributed as X, denote their empirical covariance and its spectrum as

$$\hat{\mathscr{K}} = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i = \sum_{i=1}^{n} \hat{\lambda}_j \hat{\phi}_j \otimes \hat{\phi}_j.$$

Now define

$$\hat{Y}_i = \sum_{j=1}^{r} \langle X_i, \hat{\phi}_j \rangle \hat{\phi}_j = \hat{\mathscr{P}}_1 X_i \quad \text{and} \quad \hat{Z}_i = X_i - \hat{Y}_i = \hat{\mathscr{P}}_2 X_i, \qquad i = 1, \dots, n,$$

with $\hat{\mathscr{P}}_1 = \sum_{i=1}^{r} \hat{\phi}_i \otimes \hat{\phi}_i$ the projection onto $\operatorname{span}\{\hat{\phi}_1, \dots, \hat{\phi}_r\}$ and $\hat{\mathscr{P}}_2 = \mathscr{I} - \hat{\mathscr{P}}_1$. Let us denote the sample covariance operators of the $\hat{Y}_i$'s and the $\hat{Z}_i$'s as

$$\hat{\mathscr{K}}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \otimes \hat{Y}_i \quad \& \quad \hat{\mathscr{K}}_2 = \frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i \otimes \hat{Z}_i,$$

respectively. Finally, let $\hat{C}_1$ and $\hat{C}_2$ be the empirical covariances between the proxy regressors and the responses,

$$\widehat{C}_1 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y}) \left(Y_i - \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \right), \quad \widehat{C}_2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y}) \left(Z_i - \frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i \right).$$

From (3.2), it is clear that a natural definition of the hybrid regularisation estimator of β is:

Definition 1 (Hybrid Regularisation Estimator)**.** *The hybrid regularisation estimator $\hat{\beta}_{HR}$ is defined as the solution to the penalised least squares problem*

$$\min_{\beta_1, \beta_2 \in \mathcal{H}} n^{-1} \sum_{i=1}^{n} (y_i - \bar{y} - \langle \hat{Y}_i - \bar{Y}, \beta_1 \rangle - \langle \hat{Z}_i - \bar{Z}, \beta_2 \rangle)^2 + \rho \|\beta_2\|^2,$$

where $\beta = \beta_1 + \beta_2$ with $\beta_1 = \hat{\mathscr{P}}_1 \beta$, $\beta_2 = \hat{\mathscr{P}}_2 \beta$, $\bar{Y} = n^{-1}\sum_{i=1}^{n} \hat{Y}_i$ and $\bar{Z} = n^{-1}\sum_{i=1}^{n} \hat{Z}_i$. It is explicitly given by

$$\hat{\beta}_{HR} = \hat{\mathscr{K}}_1^{-} \hat{C}_1 + \hat{\mathscr{K}}_{\rho,2}^{-} \hat{C}_2, \tag{4.1}$$

where $\hat{\mathscr{K}}_{\rho,2} = \hat{\mathscr{K}}_2 + \rho \hat{\mathscr{P}}_2$.

Remark 1. *A priori, there is no reason why one should choose $\hat{Y}$ to be the projection of X onto the* ***first*** *r eigenfunctions:* ***any*** *collection of r eigenfunctions could be considered. In principle, we choose r eigenfunctions of $\hat{K}$ that: (1) yield a component $\hat{Y}$ with a well-conditioned covariance operator $\hat{K}_1$; (2) and capture a large part of the norm of β. Since β is unknown in practice, (2) is impossible to control. Still, the whole point of fitting a functional linear model is the understanding that β correlates well with the signal rather than the noise in X, and thus this correlation is expected to be carried by the leading principal components of X, explaining our choice of selecting the first r components, subject to a well-conditioning restriction.*

4.2. Theoretical Properties

We now turn to prove that both the gain in efficiency and the minimaxity observed in the oracle setup also carry over to the practically feasible hybrid estimator. We will make use of the following assumptions.

(A1) X is a centered Gaussian process.
(A2) The eigenvalues $\lambda_1 > \lambda_2 > \ldots$ of $\mathscr{K}$ are all positive. Also, for constants $\alpha > 1$, $0 < c < C$ and $j_0 \geq 1$, we have $cj^{-\alpha} \leq \lambda_j \leq Cj^{-\alpha}$ for all $j \geq j_0$.
(A3) For constants $d > 0$, $\eta > 1/2$ and $j_0 \geq 1$, we have $|\langle \beta, \phi_j \rangle| \leq dj^{-\eta}$ for all $j \geq j_0$.

The distributional assumption in Condition (A1) is only for simplicity of presentation and can be relaxed to accommodate other distributions. In that case, we would need to assume that $\mathbb{E}(\|X\|^{16}) < \infty$, that $\mathbb{E}(\langle X, \phi_j \rangle^4) \leq C\lambda_j^2$ for all $j \geq 1$ and a constant $C > 0$, and the pseudo-independence condition similar to that mentioned earlier, i.e. that $\mathbb{E}(\prod_{u=1}^4 \langle X, \phi_{j_u} \rangle^{l_u}) = \prod_{u=1}^4 \mathbb{E}(\langle X, \phi_{j_u} \rangle^{l_u})$ for l_u's satisfying $1 \leq l_u \leq 4$ and $\sum_{u=1}^4 l_u \leq 16$. These in particular hold if X has the representation $X = \sum_{j=1}^\infty \lambda_j^{1/2} V_j \phi_j$, where the V_j's are i.i.d. zero mean random variables with finite moments (cf. the discussion before Corollary 1).

Conditions (A2) and (A3) have been used by Hall and Horowitz (2007) to obtain the rate of convergence of the Tikhonov regularisation estimator in terms of its integrated mean squared error. The eigenvalue decay regime considered in Condition (A2) corresponds to the so-called mildly ill-posed case. In the inverse problems and FDA literature, typically two different eigenvalue regimes are considered – (a) the mildly ill-posed case, when the eigenvalues decay at a polynomial rate, and (b) the severely ill-posed case, when the eigenvalues decay exponentially (also known as the *super-smooth* case). It is well known that the asymptotic properties of statistical procedures in inverse problems (e.g., deconvolution, estimation of a slope parameter, etc.) depend very heavily on the particular eigenvalue regime under consideration (see, e.g., Meister (2009), Ch. 1 in Alquier, Gautier and Stoltz (2011) etc. for the general inverse problem and deconvolution literature). In functional data analysis, in particular, while prediction is easier if the eigenvalues have faster decay, the converse is true for estimation (see, e.g., Cai and Hall (2006), Hall and Horowitz (2007), Cardot, Mas and Sarda (2007)). The mildly ill-posed regime considered here is the standard setting in the inverse problem literature – indeed this is also the regime that has exclusively been considered for estimation in functional regression.

The interplay between α and η determines the degree of difficulty of estimating β. Clearly, the larger the value of η, the easier is the estimation problem. If α is large, then the distribution of X becomes almost finite dimensional. In that case, if η is small, then the estimation problem is difficult if there are important components of β, namely, $\langle \beta, \phi_j \rangle$ corresponding to small values of λ_j. This is because there is very little information on X in those directions, and thus those components of β will be difficult to estimate. We will later see exactly how α and η determine the precision in estimating β. Condition (A2) is sufficient to ensure that $\mathbb{E}(\|X\|^2) = \sum_{j=1}^\infty \lambda_j < \infty$, which in turn implies that X is a (tight) random element in $\mathcal{H}$. Condition (A3) ensures that $\|\beta\|^2 < \infty$.

Our first result compares the efficiency of the oracle version of the hybrid and Tikhonov estimator to that of their empirical version.

Theorem 2. *Suppose that conditions (A1)–(A3) hold, and $\alpha < 2\eta$. Then,*

$$\begin{aligned} &|\mathrm{MSE}(\hat{\beta}_{HR}) - \mathrm{MSE}(\widetilde{\beta}_{HR})| \\ &= O(1)\left[\left\{\frac{1}{n\rho^{1+\frac{1}{\alpha}}} + \rho^m\right\}^{1/2}\left(\frac{1}{n\rho^{1+\frac{1}{\alpha}}}\right)^{1/2} + \frac{1}{n\rho^{1+\frac{1}{\alpha}}}\right] \end{aligned} \tag{4.2}$$

for any sequence $\rho \to 0$ satisfying $n\rho^2 \to \infty$ as $n \to \infty$. Further,

$$\mathrm{MSE}(\hat{\beta}_{HR}) = O(1)\left\{\frac{1}{n\rho^{1+\frac{1}{\alpha}}} + \rho^m\right\}$$

as $n \to \infty$. Here $m = (2\eta - 1)/\alpha$ or $m = 2$ according as $\alpha > \eta - 1/2$ or $\alpha < \eta - 1/2$. Moreover, analogous rates of convergence also hold for $|\mathrm{MSE}(\hat{\beta}_{TR}) - \mathrm{MSE}(\widetilde{\beta}_{TR})|$ and $\mathrm{MSE}(\hat{\beta}_{TR})$.

The terms $n^{-1}\rho^{-1-1/\alpha}$ and ρ^m in the expression of $\mathrm{MSE}(\hat{\beta}_{HR})$ given in the above theorem clearly show the effects of the variance and the bias terms, respectively, in the estimation of β. It also reveals that only the bias is affected by the rate of decay of the $\langle \beta, \phi_j \rangle$'s but not the variance. This is expected because the variability in the estimation of β should purely depend on the fluctuations in X, which depends on the rate of decay of the eigenvalues of the covariance operator $\mathscr{K}$ of X.

As a corollary, we can obtain rates of convergence when the ridge parameter ρ decays with n in specific manners:

Corollary 2. *Consider the setup of Theorem 2. Let $c > 0$ be a fixed constant. Then,*

$$\begin{aligned} &\mathrm{MSE}(\hat{\beta}_{HR}) \\ &= \begin{cases} O(n^{-(2\eta-1)/(\alpha+2\eta)}) & \text{if } \eta - 1/2 < \alpha < 2\eta \text{ and } \rho \sim cn^{-\frac{\alpha}{\alpha+2\eta}} \\ O(n^{-2\alpha/(3\alpha+1)}) & \text{if } \alpha < \eta - 1/2 \text{ and } \rho \sim cn^{-\frac{\alpha}{3\alpha+1}} \end{cases} \end{aligned} \tag{4.3}$$

as $n \to \infty$. Further, the same rates of convergence also hold for $\mathrm{MSE}(\hat{\beta}_{TR})$.

The above corollary gives the rates of convergences of the hybrid regularisation estimator in terms of its mean squared error under different regimes determined by α and η. These regimes correspond to the different degrees of difficulty of the estimation problem in the functional linear regression setting. The rates of decay of ρ to zero are chosen so as to optimize the rates of convergence of the MSEs.

Remark 2. *Note that the asymptotic rate of convergence of $\mathrm{MSE}(\hat{\beta}_{TR})$ was proved in Theorem 2 in Hall and Horowitz (2007) under the restriction that $\alpha < \eta + 1/2$ and $\rho \sim cn^{-\alpha/(\alpha+2\eta)}$. The above corollary reveals that this upper bound on the values of the decay rate of the eigenvalues of X can be relaxed. Further, the same rate of convergence is in fact true for a wider class of values of α and η so long as $\eta - 1/2 < \alpha < 2\eta$. Note that Hall and Horowitz (2007) did not require $\alpha > \eta - 1/2$.*

Remark 3. *Hall and Horowitz (2007) showed that the rate of convergence of $\mathrm{MSE}(\hat{\beta}_{TR})$ is optimal in a minimax sense under the conditions of Theorem 2 when $1 < \alpha < \eta + 1/2$ and $\rho \sim cn^{-\alpha/(\alpha+2\eta)}$. From Corollary 2, Remark 2 and the proof of equation (3.11) in Hall and Horowitz (2007), it follows that the hybrid regularisation estimator also enjoys the same minimax optimal rate of convergence for the same choice of regularisation parameter in the regime $\max(1, \eta - 1/2) < \alpha < 2\eta$.*

Remark 4. *The spectral truncation estimator $\widehat{\beta}_{ST}$ studied by Hall and Horowitz (2007) is known to satisfy* $\mathrm{MSE}(\widehat{\beta}_{ST}) > \delta n^{-(2\eta-1)/(\alpha+2\eta)}$ *for some* $\delta > 0$ *and sufficiently large* n *and that* $\mathrm{ISE}(\widehat{\beta}_{ST}) := ||\widehat{\beta}_{ST} - \beta||^2 = O_p(n^{-(2\eta-1)/(\alpha+2\eta)})$ *under appropriate conditions including* $1 < \alpha < 2\eta - 2$ *(see Theorem 1 in Hall and Horowitz (2007)). This rate is also the minimax rate of convergence in a concentration probability sense. Now, it follows from Corollary 2 that when* $\eta - 1/2 < \alpha < 2\eta$, *we have* $\mathrm{ISE}(\widehat{\beta}_{HR}) := ||\widehat{\beta}_{HR} - \beta||^2 = O_p(n^{-(2\eta-1)/(\alpha+2\eta)})$. *In particular, when* $\lambda_j \sim cj^{-\alpha}$ *for all large* j *(so that both condition (A2) in our paper and condition (3.2) in Hall and Horowitz (2007) are satisfied) and when* $\max(1, \eta - 1/2) < \alpha < 2\eta - 2$, *it follows that both of these two estimators have the same minimax rate of convergence in the concentration probability sense. Note that it is unknown whether the spectral truncation estimator will attain the minimax rate of convergence in the MSE sense like the hybrid and the Tikhonov estimators discussed in Remark 3.*

Theorem 2 and Corollary 1 set the stage for our main result, showing that the hybrid estimator can improve upon the Tikhonov estimator in a non-asymptotic sense, even in the empirical case:

Theorem 3. *Suppose that the conditions of Theorem 2 hold. Let* $c > 0$ *be a fixed constant and* $\rho \sim cn^{-\varepsilon}$ *for some* $\varepsilon > 0$. *Also assume that at least one of* $\langle \beta, \phi_j \rangle$, $j = 1, 2, \ldots, r$, *is non-zero. Then, there exists a constant* $\kappa_0 > 0$ *such that*

$$\mathrm{MSE}(\hat{\beta}_{TR}) - \mathrm{MSE}(\hat{\beta}_{HR}) > \kappa_0 n^{-2\varepsilon}$$

for all sufficiently large n *if* $\varepsilon < \alpha/(5\alpha - 2\eta + 2)$ *in case* $\eta - 1/2 < \alpha < 2\eta$ *or if* $\varepsilon < \alpha/(3\alpha + 1)$ *in case* $\alpha < \eta - 1/2$.

Although the hybrid estimator and the Tikhonov estimator enjoy the same rate of convergence by Theorem 2, the latter is effectively rendered *inadmissible* by the hybrid estimator for a broad range of choices of ρ, including choices arbitrarily close to the optimal one (as in Corollary 2) – and this is true for all sample sizes above a threshold. It is illustrated in the simulations study in Section 6, that this improvement can be sizeable, even for modest sample sizes. Moreover, it is interesting to note that we can attain this improvement regardless of the choice of r may be, even for $r = 1$ (provided, of course, that $\langle \beta, \phi_1 \rangle \neq 0$ as the theorem requires). The domination result in Theorem 3 can be viewed in the same light as the result on the domination of the ordinary least squrares estimator by the ridge regression estimator in the multivariate setting, which holds for a whole range (depending on unknown population parameters) of values of the regularization parameter (see Thm. 2 in Theobald (1974)).

The proof of the Theorem reveals that the determining factor in the inadmissibility of the Tikhonov estimator is the larger bias component compared to the hybrid estimator (see also equation (1.4) in the proof of the oracle case provided in the Supplementary Material (Chakraborty and Panaretos, 2016)). An important requirement is that the choice of Y to be such that β is at least partially expressed by the eigenfunctions of Y. Of course, how large the sample size has to be will depend on r and also depend on the condition number of the covariance operator of Y. The latter is determined by the relative magnitudes of the eigenvalues of Y, equivalently, the relative importance of the associated eigenfunctions in explaining the variation in X.

Remark 5. *Since in practice the regularization parameters are typically chosen in a data-dependent manner (e.g. using either generalized cross-validation (GCV) or other methods, cf. subsection 4.3), it may be asked whether the hybrid estimator strictly dominates the Tikhonov estimator (i.e. in a non-asymptotic sense as in Theorem 3) with a data-dependent choice of the tuning parameter. However, any result related to tuning parameters would most likely need to be asymptotic (it would rely on establishing the asymptotic rate of the empirically chosen tuning parameter), and therefore in the very best case it seems that all that one could establish is that both Tikhonov and hybrid regularisation achieve the same rate asymptotically under cross-validation. Such results (including convergence rate results as in Corollary 2 with estimated*

Algorithm 1 Construction of the Hybrid Estimator

(Step 0) Determine the eigenvalues $\widehat{\lambda}_j$ and eigenfunctions $\widehat{\phi}_j$ of $\widehat{\mathscr{K}}$.

(Step 1) Fix a condition number L and choose r using the eigenvalues of $\widehat{\mathscr{K}}$ as

$$r = \sup\left\{j \geq 1 : \left(\hat{\lambda}_1/\hat{\lambda}_j\right)^{1/2} \leq L\right\}.$$

(Step 2) Set $\hat{Y}_i = \sum_{j=1}^{r}\langle X_i, \hat{\phi}_j\rangle\hat{\phi}_j$ and $\widehat{Z}_i = X - \hat{Y}_i$, and compute

$$\widehat{C}_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})\left(Y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i\right) \quad \& \quad \widehat{C}_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})\left(Z_i - \frac{1}{n}\sum_{i=1}^{n}\hat{Z}_i\right).$$

(Step 3) For the chosen r, choose ρ by generalized cross-validation.

(Step 4) Use this value of ρ and the value of r obtained in Step 1 to compute $\widehat{\beta}_{HR}$ using (4.1),

$$\widehat{\beta}_{HR} = \sum_{j\leq r}\widehat{\lambda}_j^{-1}\langle\widehat{C}_1, \widehat{\phi}_j\rangle\widehat{\phi}_j + \sum_{j>r}(\widehat{\lambda}_j + \rho)^{-1}\langle\widehat{C}_2, \widehat{\phi}_j\rangle\widehat{\phi}_j.$$

regularisation parameters) seem to require substantially novel analysis and beyond the scope of this paper. A potentially more accessible result would be to show that the GCV choice of the tuning parameter yields the minimax rate for our estimator. Still, such questions appear to remain open even for classical multivariate Tikhonov regularisation. Results involving asymptotics for a GCV choice of regularization parameter ρ (see, e.g., Utreras (1987), Wahba (1990), Lukas (1993, 2006)) involve inverse problems with regression function $L_i f = Kf(x_i)$, where the design points x_i's are fixed (i.e., deterministic), f is a smooth function, and K is an appropriate compact operator or the identity operator in case of spline smoothing. Even then, the tuning considered is with respect to the expected GCV statistic rather than the GCV statistic itself, yielding a deterministic choice that depends on unknown population parameters. In our case, a closed form of this expected value appears out of reach due to the complicated involvement of estimators of eigenvalues and eigenfunctions in the expression, unlike the fixed design case (indeed cross-validation asymptotics for the random design case do not appear to exist even in the classical multivariate regression setting). In the functional case, there do not appear to be any such results as of yet, whatever the regularization method may be. Convergence rates are typically obtained for some range of tuning parameters, and then the practical implementation uses some form of data-driven choice.

4.3. Computational Aspects

Algorithm 1 provides a the step-by-step construction of the hybrid estimator. In summary, our recommendation is to fix an r by the condition index approach discussed in in Remark 1, and to then choose ρ by generalized cross-validation (as in standard Tikhonov regularisation, see e.g. Yuan and Cai (2010)). Going through the steps in Algorithm 1, one case see that there is no computational overhead or added complexity relative to the construction of a spectral truncation or Tikhonov estimator. An alternate, slightly more complex albeit fully automated procedure would be to use a double generalized cross-validation for choosing both r and ρ.

It is worth remarking that one of the widely-used methods for choosing finitely many principal components of X is to select the number required to capture the bulk of the trace of the empirical covariance operator – a typical choice of threshold is that of 85% of the total variation in X (see Jolliffe (2002) and Ramsay and Silverman (2005)). We shall later see in the simulation studies in Section 6 that this choice is far from optimal, as it makes no reference to the condition number of the resulting multivariate regression.

Our counterproposal on choosing r guarantees that the covariance operator of $\hat{Y}$ is *well-conditioned* – the whole point of the hybrid estimator is to extract a component of the regression that does not need regularisation, after all, and such components are in no way connected with the cumulative variance explained. Condition indices and their maximum, which is called the condition number, are well-known in the classical multivariate regression setup as indicators of the degree of collinearity among the covariates, and more generally in numerical analysis as a measure of the instability of a linear problem. A rule-of-thumb is that a condition number ≤ 30 indicates well-posedness (see, e.g., Hocking (2003)). An alternative way to choose L could be to consider a plot of the empirical condition indices and look for the "elbow". With a pre-fixed L, it is obvious why the choice of r should be large when the eigenvalues decay slowly, and why it should be more conservative when they decay fast. Furthermore, since $\sup_{j\geq 1}|\hat{\lambda}_j - \lambda_j| \to 0$ in probability as $n \to \infty$, it follows that $\sup\{j \geq 1 : [\hat{\lambda}_1/\hat{\lambda}_j]^{1/2} \leq L\} \to \sup\{j \geq 1 : [\lambda_1/\lambda_j]^{1/2} \leq L\}$ in probability as $n \to \infty$, i.e., r is chosen consistently by this procedure. We shall see later in the simulations that in some cases, this choice yields a better estimator in the MSE sense.

5. The Case of Discretely Observed Functions

For data in a function space, say, $L_2[0,1]$, it may happen that instead of observing the entire curve X, one can only observe it on a grid, say,

$$0 \leq t_1 < t_2 < \ldots < t_m \leq 1.$$

Thus, the regressor at hand is an m-dimensional vector

$$X^{(m)} = (X(t_1), X(t_2), \ldots, X(t_m))'.$$

In this setup, an approximation of the functional linear model considered in (1.1) is

$$y = \alpha + m^{-1}\sum_{p=1}^{m} X(t_p)\beta(t_p) + \epsilon. \tag{5.1}$$

We define $\beta^{(m)} = (\beta(t_1), \beta(t_2), \ldots, \beta(t_m))'$. This setup of discretely observed data is closely related to the *time-sampling model* considered by Amini and Wainwright (2012) or the *common design model* considered by Cai and Yuan (2011) with the difference that we do not consider measurement errors in the discrete observations of the X_i. We do not consider the case of irregular grids since this would require pre-smoothing of the individual observations, and consequently, the asymptotic convergence rate results would depend on the particular smoothing procedure used. The common grid approach allows us to study the performance of the two regularisation procedures without relying on a specific smoothing method, and hence independently of any external arbitrary choice.

In the discretely sampled setup considered above, the oracle and the empirical hybrid regularisation estimators of $\beta^{(m)}$ are defined analogously and are denoted by $\widetilde{\beta}^{(m)}_{HR}$ and $\widehat{\beta}^{(m)}_{HR}$, respectively. Similarly, the oracle and the empirical Tikhonov estimators are denoted by $\widetilde{\beta}^{(m)}_{TR}$ and $\widehat{\beta}^{(m)}_{TR}$, respectively.

In order to state results analogous to Theorems 2 and 3, we need to assume the following modifications of assumptions (A2) and (A3). We denote the eigenvalue-eigenvector pairs of the covariance matrix of $X^{(m)}/\sqrt{m}$ by $(\lambda^{(m)}_j, \phi^{(m)}_j)$ for $j = 1, 2, \ldots, m$.

(A2') Suppose that $\lambda^{(m)}_1 > \ldots > \lambda^{(m)}_m > 0$. Also, for constants $\alpha > 1$, $0 < c' < C'$ and $j'_0 \geq 1$, we have $c'j^{-\alpha} \leq \lambda^{(m)}_j \leq C'j^{-\alpha}$ for all $j'_0 \leq j \leq m$ when m is sufficiently large.

(A3') For constants $d' > 0$, $\eta' > 1/2$ and $j_0' \geq 1$, we have $m^{-1/2}|\langle \beta^{(m)}, \phi_j^{(m)} \rangle| \leq d'\{j^{-\eta'} + m^{-1}\}$ for all $j_0' \leq j \leq m$ when m is sufficiently large.

In assumption (A3'), the parameter η' is some function of the parameters α and η that appear in assumptions (A2) and (A3) earlier. The two components in the inequality in assumption (A3') may be respectively interpreted as the contribution at the functional level and the error due to discretization. For instance, when X is a standard Brownian motion, and if β lies in its RKHS and satisfies assumption (A3), then $\eta' = \alpha$ for $\alpha \leq \eta$ and $\eta' = \eta$ for $\eta < \alpha < 2\eta - 1$ (see the Supplementary Material (Chakraborty and Panaretos, 2016) for a proof of this fact). Note that the condition $\alpha < 2\eta - 1$ is needed to ensure that β lies in the RKHS of the standard Brownian motion. Also, using the arguments in Amini and Wainwright (2012), it can be shown that condition (A2') holds in this case (see Appendix A in the Supplementary Material of Amini and Wainwright (2012)).

Theorem 4 now shows that, even when we have discretely observed data, the hybrid and the Tikhonov estimators enjoy the same properties as their fully functional counterparts provided that the grid size grows to infinity sufficiently fast. We focus on the m-dimensional version of the slope parameter since in the discrete observation setting, it is this m-dimensional estimator that is typically used in practice. One could of course smooth it using any smoothing technique, but we prefer to give a result that is independent of external arbitrary choices.

Theorem 4. *Suppose that conditions (A1), (A2') and (A3') hold, and $\alpha < 2\eta'$. Also assume that $m > \rho^{-2}$. Then,*

$$\begin{aligned} & m^{-1}|\mathrm{MSE}(\hat{\beta}_{HR}^{(m)}) - \mathrm{MSE}(\widetilde{\beta}_{HR}^{(m)})| \\ & = O(1)\left[\left\{\frac{1}{n\rho^{1+\frac{1}{\alpha}}} + \rho^M\right\}^{1/2}\left(\frac{1}{n\rho^{1+\frac{1}{\alpha}}}\right)^{1/2} + \frac{1}{n\rho^{1+\frac{1}{\alpha}}}\right] \end{aligned} \tag{5.2}$$

for any sequence $\rho \to 0$ satisfying $n\rho^2 \to \infty$ as $n \to \infty$. Further,

$$m^{-1}\mathrm{MSE}(\hat{\beta}_{HR}^{(m)}) = O(1)\left\{\frac{1}{n\rho^{1+\frac{1}{\alpha}}} + \rho^M\right\}$$

as $n \to \infty$. Here $M = (2\eta' - 1)/\alpha$ or $M = 2$ according as $\alpha > \eta' - 1/2$ or $\alpha < \eta' - 1/2$. Moreover, analogous rates of convergence also hold for $m^{-1}|\mathrm{MSE}(\hat{\beta}_{TR}^{(m)}) - \mathrm{MSE}(\widetilde{\beta}_{TR}^{(m)})|$ and $m^{-1}\mathrm{MSE}(\hat{\beta}_{TR}^{(m)})$. Thus,

$$\begin{aligned} & m^{-1}\mathrm{MSE}(\hat{\beta}_{HR}^{(m)}) \\ & = \begin{cases} O(n^{-(2\eta'-1)/(\alpha+2\eta')}) & \text{if } \eta' - 1/2 < \alpha < 2\eta' \text{ and } \rho \sim cn^{-\frac{\alpha}{\alpha+2\eta'}} \\ O(n^{-2\alpha/(3\alpha+1)}) & \text{if } \alpha < \eta' - 1/2 \text{ and } \rho \sim cn^{-\frac{\alpha}{3\alpha+1}} \end{cases} \end{aligned}$$

as $n \to \infty$. Further, the same rates of convergence also hold for $m^{-1}\mathrm{MSE}(\hat{\beta}_{TR}^{(m)})$.

Note that in the above theorem, the condition $m > \rho^{-2}$ implicitly specifies a rate of growth of m with the sample size n. Indeed, since ρ depends on n, the previous condition puts a restriction on the sampling rate m/n, which has to be greater than $n^{-1}\rho^{-2}$. This rate of course depends on the rate of decay of ρ, the amount of regularisation involved.

Finally, our last result shows that, similar to the case of perfect functional observations, the hybrid estimator outperforms the Tikhonov estimator for sufficiently large sample sizes and suitably chosen regularisation even when observations are discrete.

Theorem 5. *Suppose that the conditions of Theorem 4 hold. Let $c > 0$ be a fixed constant and $\rho \sim cn^{-\varepsilon}$ for some $\varepsilon > 0$. Also assume that at least one of $\langle \beta, \phi_j \rangle$, $j = 1, 2, \ldots, r$, is non-zero. Then, there exists a constant $\theta_0 > 0$ such that*

$$m^{-1}\{\mathrm{MSE}(\hat{\beta}_{TR}^{(m)}) - \mathrm{MSE}(\hat{\beta}_{HR}^{(m)})\} > \theta_0 n^{-2\varepsilon}$$

for all sufficiently large n if $\varepsilon < \alpha/(5\alpha - 2\eta' + 2)$ in case $\eta' - 1/2 < \alpha < 2\eta'$ or if $\varepsilon < \alpha/(3\alpha + 1)$ in case $\alpha < \eta' - 1/2$.

The proof of Theorem 5 can be developed in the same way as that of the proof of Theorem 3 and is thus omitted.

6. Simulation Study

We now turn to the assessment of the practical performance of the hybrid regularisation estimator relative to the Tikhonov estimator by means of a simulation study. To this aim, we shall consider the same simulation framework considered in Hall and Horowitz (2007) and Yuan and Cai (2010). Take $\mathcal{H} = L_2[0, 1]$, the space of square-integrable real functions on the interval $[0, 1]$, with the usual inner product. Let X be defined via its Karhunen-Loève expansion as

$$X = \sum_{j=1}^{50} \gamma_j Z_j \phi_j,$$

with the Z_j's being i.i.d. uniform random variables on $[-3^{1/2}, 3^{1/2}]$, $\phi_1(t) = 1$ and $\phi_j(t) = 2^{1/2}\cos(j\pi t)$ for $t \in [0, 1]$. Further, $\gamma_j = (-1)^{j+1} j^{-\alpha/2}$ for $j \geq 1$, and we choose α to either be equal to 1.1 or 2. These two values of α correspond to slow and fast decays of the eigenvalues of X. Let $b_1 = 1$ and $b_j = 4(-1)^{j+1} j^{-2}$ for $j = 2, 3, \ldots, 50$. We have chosen three different kinds of slope function: (a) $\beta = \beta_1 = \sum_{j=1}^{50} b_j \phi_j$, (b) $\beta = \beta_2 = \sum_{j=1}^{5} b_j \phi_j$, and (c) $\beta = \beta_3 = \sum_{j=6}^{50} b_j \phi_j$. Note that in cases (b) and (c) above, β is expressed by two mutually orthogonal subcollections of eigenfunctions of X. We have considered these two choices of β to study how the parsimony of β in fewer or more eigenfunctions of X influences the performance of the hybrid estimator. The sample size is $n = 100$. The distribution of the error variable ϵ in the functional regression model is standard Gaussian. The X's are evaluated at 50 equispaced grid points in $[0, 1]$. All the estimated mean squared errors are averaged over 1000 Monte-Carlo replications.

6.1. Comparison of MSEs of hybrid and Tikhonov estimator over various choices of tuning parameters

Figure 1 gives the plots of the MSEs of the hybrid and the Tikhonov estimators for different choices of ρ. In each plot, we have considered the mean squared errors of the hybrid estimator for every $r = 1, 2 \ldots, 5$. For the Tikhonov estimator, the smallest value of the mean squared is designated by a triangle. For the hybrid estimator, the smallest value of the mean squared error for each choice of r is marked by a circle. We also point out the smallest mean squared error across the different choices of r by a star. In all of the above cases, the optimal values of ρ and r can be read from the plot and we do not mark them to avoid clutter.

The top four plots in Figure 1 show that the optimal value of the mean squared error is markedly smaller for the hybrid estimator than for the Tikhonov estimator with the ratio between the two mean squared errors being about 2 and 1.4 for $\alpha = 1.1$ and 2, respectively for both $\beta = \beta_1$ and β_2 (see Table 1). It can also be remarked that the optimal mean squared error corresponding to the hybrid estimator can also improve upon the optimal Tikhonov mean squared error even for some values of r that are suboptimal. In fact, the

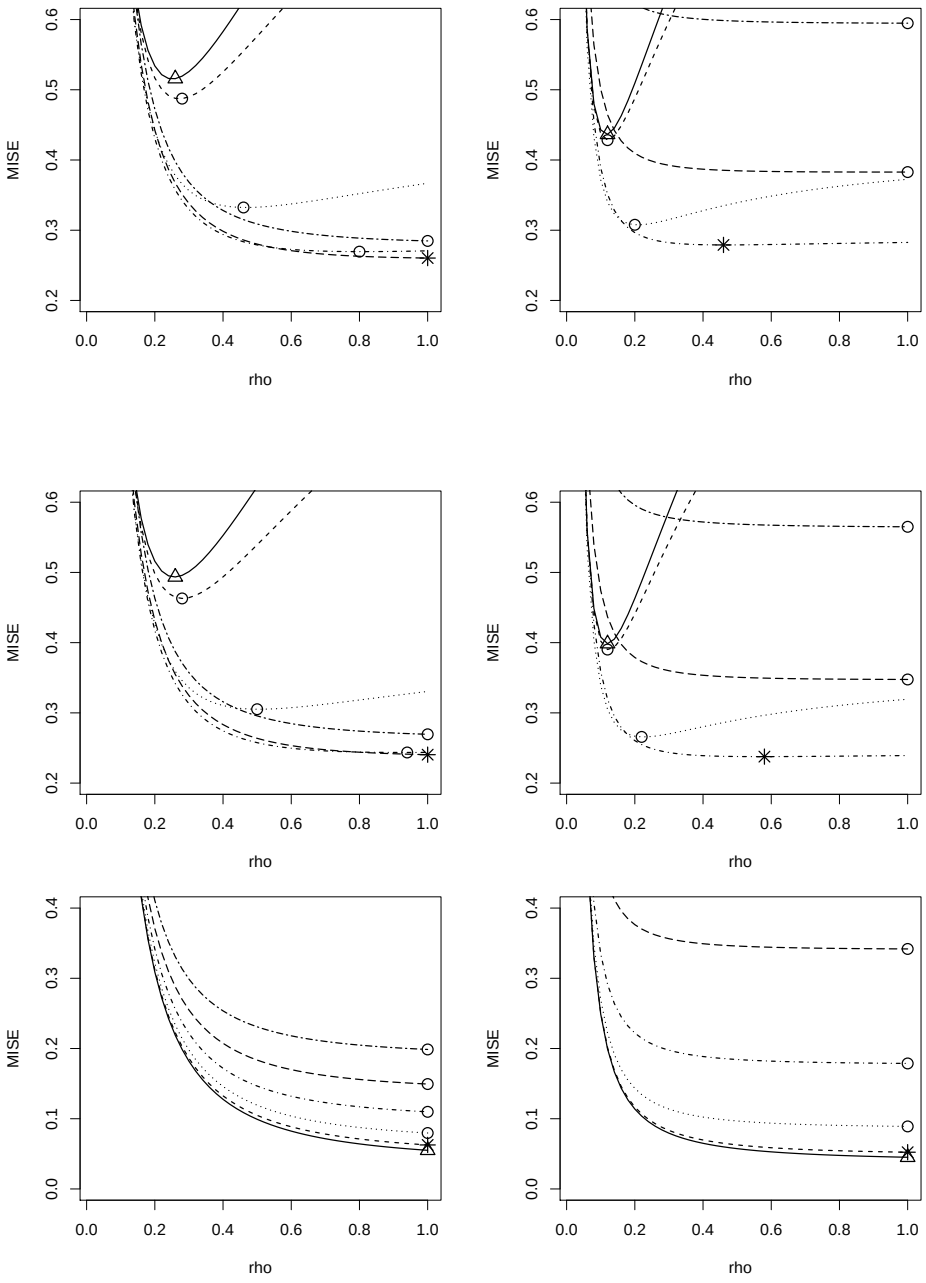

Figure 1. Plots of the MSEs of the Tikhonov estimator (solid curves) and the hybrid regularisation estimator for $r = 1$ (dashed curves), $r = 2$ (dotted curves), $r = 3$ (dot-dashed curves), $r = 4$ (long-dashed curves) and $r = 5$ (two-dashed curves). The plots in the left and the right columns correspond to $\alpha = 1.1$ and 2, respectively. The plots in the top, middle and bottom rows correspond to β_1, β_2 and β_3, respectively.

difference in each case is statistically significant in the following sense – the two mean squared errors, which are averages of independent Monte Carlo iterations, are significantly different, when a large sample test of difference of two means is applied. These observations lend support to Theorem 3. In the plots in the last row in Figure 1, where $\beta = \beta_3$, the minimum mean squared errors of the two estimators are not significantly different. This does not contradict Theorem 3 since β_3 does not satisfy the assumption in that theorem for any $r = 1, 2, \ldots, 5$.

The first two choices of β are at least partially expressed by the eigenfunctions associated with the five largest eigenvalues of X. Note that these eigenvalues explain only about 56% of the total variation of X if $\alpha = 1.1$, while this percentage is about 90% if $\alpha = 2$. Thus, the performance of the hybrid estimator does not seem to depend much on whether or not the eigenvalues associated with the eigenfunctions expressing β explain a large amount of the total variation of X. It is also observed that if we had chosen r by the "85%-rule", then one would end up choosing more principal components compared to the optimal value of r found in the simulation studies for each value of α and $\beta = \beta_1$ or β_2. Further, the optimal number of principal components is 1 for both values of α when $\beta = \beta_3$. These findings indicate that one should generally *not* use the "85%-rule" for choosing r in the construction of the hybrid estimator. Further, the simulation studies also confirm that when the eigenvalues of X decay slowly (well-conditioned regime), it is better to choose a higher value of r. By doing so, we make substantial gains if β is at least partially expressed by those r eigenfunctions, and we will only perhaps lose out slightly otherwise. On the other hand, if the eigenvalues of X decay fast (ill-conditioned regime), then a more conservative choice should be used (see Figure 1). This is consistent with the choice of r using condition numbers that was recommended in subsection 4.3.

We also observe that the mean squared error of the hybrid estimator for appropriately chosen r is significantly smaller than that of the Tikhonov estimator for all values of ρ greater than the optimal one for the latter estimator, which is small except when $\beta = \beta_3$. In that case, for all $\rho > 0.2$, the mean squared errors of the two estimators are almost coincident for an appropriately chosen r. From the simulation studies, it seems that the hybrid estimator acts as a safeguard against over-estimation. This is in contrast to the Tikhonov estimator which is found to be much more sensitive to choice of large values of ρ when $\beta = \beta_1$ or β_2 (see Figure 1).

6.2. Comparison of MSEs of hybrid, Tikhonov and spectral truncation estimator using specific choices of tuning parameters

We next compare the MSEs of the hybrid regularisation estimator with that of the Tikhonov regularisation estimator as well as the spectral truncation estimator when the regularization parameters r and ρ in the hybrid estimator are chosen using the fully automated double cross-validation technique discussed in subsection 4.3 and the regularisation parameter involved in each of the other two estimators is also chosen using cross-validation. Later in this subsection, we will also provide a similar comparison when the parameter r in the hybrid estimator is chosen using a user-specified condition number L as outlined in Algorithm 1 in subsection 4.3.

The sample sizes used for the above studies are $n = 50, 100$ and 300. Also, we have included another choice of γ_j's in addition to that used in the previous subsection: $\gamma_1 = 1$, $\gamma_j = 0.2(-1)^{j+1}(1 - 0.0001j)$ if $2 \leq j \leq 4$ and $\gamma_{5j+k} = 0.2(-1)^{5j+k+1}\{(5j)^{-\alpha/2} - 0.0001k\}$ for $j \geq 1$ and $0 \leq k \leq 4$. As in section 6, we have chosen $\alpha = 1.1$ or 2. This new set of γ_j's generate "closely-spaced" eigenvalues and was also considered by Hall and Horowitz (2007) and Yuan and Cai (2010). The choice of the γ_j's considered towards the beginning of this section leads to "well-spaced" eigenvalues. It is known that the spectral truncation estimator has better (worse) performance compared to the Tikhonov regularisation estimator in the "well-spaced" ("closely-spaced") scenario (see Hall and Horowitz (2007)).

Table 1. MSEs of the hybrid regularisation, the Tikhonov regularisation and the spectral truncation estimators when $n = 100$.

well-spaced							
β	α	MSE_{ST}^{GCV}	MSE_{ST}^{true}	MSE_{TR}^{GCV}	MSE_{TR}^{true}	MSE_{HR}^{GCV}	MSE_{HR}^{true}
β_1	1.1	0.285	0.272	0.773	0.516	0.346	0.26
	2	0.296	0.286	0.608	0.445	0.311	0.274
β_2	1.1	0.271	0.247	0.763	0.494	0.357	0.24
	2	0.252	0.241	0.689	0.409	0.284	0.234
β_3	1.1	0.052	0.05	0.057	0.055	0.066	0.063
	2	0.05	0.049	0.046	0.045	0.052	0.051
closely-spaced							
β	α	MSE_{ST}^{GCV}	MSE_{ST}^{true}	MSE_{TR}^{GCV}	MSE_{TR}^{true}	MSE_{HR}^{GCV}	MSE_{HR}^{true}
β_1	1.1	1.09	0.857	1.054	0.888	0.935	0.851
	2	0.949	0.854	0.821	0.694	0.725	0.697
β_2	1.1	1.055	0.822	1.015	0.835	0.887	0.808
	2	0.896	0.813	0.763	0.647	0.681	0.647
β_3	1.1	0.051	0.05	0.045	0.043	0.058	0.051
	2	0.049	0.048	0.046	0.043	0.051	0.05

Table 1 gives the MSEs of the three estimators (averaged over 1000 Monte-Carlo iterations, namely, $(1000)^{-1}\sum_{i=1}^{1000}||\widehat{\beta}_i - \beta||^2$ with $\widehat{\beta}_i$ being the estimator in the ith itertation computed using the entire sample) for the simulated models considered earlier when $n = 100$. In Table 1, the "true MSE" refers to the minimum MSE obtained under each model when the minimization is done over a range of tuning parameter values (i.e., the estimator and thus the MSE is computed for all choices of tuning parameters in that range). The "GCV" MSE is computed with the GCV choice of the tuning parameters. The results for $n = 50$ and $n = 300$ are reported in the Supplementary Material (Chakraborty and Panaretos, 2016). The Monte-Carlo standard deviations of the MSEs are mostly of the order of 10^{-3} with some exceptions, but even these do not exceed 0.019. All the significance statements made later take these standard deviations into account.

It is observed from Table 1 that under the well-spaced scenario, the MSEs (true as well as cross-validated) of the hybrid estimator are significantly smaller than those of the Tikhonov estimator for $\beta = \beta_1$ and β_2. Somewhat surprisingly, the true MSEs of the hybrid estimator and the spectral estimator are not dissimilar. Although the cross-validation MSE of the spectral estimator for β_1 as well as β_2 is significantly smaller than that of the hybrid estimator for $\alpha = 1.1$, these MSEs are quite close when $\alpha = 2$. In the closely-spaced case, the cross-validation MSEs of the hybrid estimator are significantly smaller than those of the spectral and the Tikhonov estimators for all choices of α under β_1 and β_2. For these β's, the true MSEs of the spectral estimator and hybrid estimator are comparable for $\alpha = 1.1$, but the former become significantly larger when $\alpha = 2$. For $\beta = \beta_3$, it is found that the MSEs (true as well as cross-validated) of the three estimators are not significantly different from one another when $\alpha = 2$. In case $\alpha = 1.1$, the cross-validated MSE of the hybrid estimator is marginally larger than those of other two estimators. Further, the true MSE of the hybrid estimator is marginally larger than that of the spectral estimator in the well-spaced scenario.

Table 2. MSEs of the hybrid regularisation estimator when $n = 100$

		well-spaced		closely-spaced	
β	α	$L = 5$	$L = 10$	$L = 5$	$L = 10$
β_1	1.1	0.267	0.349	0.948	0.936
	2	0.321	0.347	0.706	0.701
β_2	1.1	0.248	0.388	0.854	0.863
	2	0.296	0.315	0.679	0.658
β_3	1.1	0.156	0.179	0.051	0.052
	2	0.089	0.147	0.05	0.052

As mentioned in the earlier simulation study, these findings do not contradict the domination result in Theorem 3. It seems that in both the well-spaced and the closely spaced situations, the cross-validation method for ρ is slightly unstable when the eigenvalues decay slowly. This may be attributed to the fact that the cross-validation estimate is based on prediction error, whose difficulty reduces as the eigenvalues decay faster (see the discussion in p. 3428 in Yuan and Cai (2010)).

We next compare the MSEs of the hybrid estimator under the above models when the estimator is computed using the algorithm given in subsection 4.3 for two choices of L, namely, $L = 5$ and $L = 10$. These choices are made by keeping in mind the eigenvalue sequence used in the simulations. The MSEs are reported in Table 2 for both the well-spaced and the closely-spaced regimes when $n = 100$. It is observed that for the closely-spaced regime, the MSEs for both choices of L do not differ significantly from the MSEs obtained using the automated double cross-validation method earlier. The situation is very different in the well-spaced regime. Here, the MSEs for specified values of L are worse than the MSEs using the double cross-validation method when $\beta = \beta_3$, and the increase is even more when $L = 10$ compared to $L = 5$. This is in line with our findings in the plots in the third row of Figure 1, where higher values of r resulted in increased MSEs of the hybrid estimator. When $\beta = \beta_3$, a very conservative choice of r is needed than those obtained using $L = 5$ or $L = 10$. Indeed, as seen from Figure 1, $r = 1$ is optimal in this case. For the other choices of β, the MSEs for both $L = 5$ and $L = 10$ are slightly larger than the MSEs obtained earlier when $\alpha = 2$, although between these two choices of L, the MSEs are quite similar. This is because a faster decay of the eigenvalues implies that the covariance is more ill-conditioned so that there is a delicate estimation bias-variance trade-off between choosing a higher condition number (increased variance but reduced bias) and a lower condition number (increased bias but reduced variance). However, when $\alpha = 1.1$ and $L = 5$, the MSEs are very close to the true MSEs. The results for $n = 50$ and $n = 300$ are very similar and we report them in the Supplementary Material (Chakraborty and Panaretos, 2016)..

The investigation done so far in this section indicates that when β is spanned by all eigenfunctions or at least has significant contribution from the leading eigenfunctions, then the condition number choice of r is in general better. On the other hand, the double cross-validation method for choosing both r and ρ may be preferred if one has no such prior knowledge about β and wants to safeguard against estimation bias. Also, the choice of L should be done by looking at the scree-plot of the empirical eigenvalues, which will give information about whether the spectrum is well-spaced or closely-spaced. This is useful in deciding how large L should be in order to balance the bias and the variance in estimation.

Sometimes in practice, the functional covariate may be observed with error, i.e., instead of observing $X_i(t)$, we observe $W_i(t) = X_i(t) + \xi_{i,t}$, where $\{\xi_{i,t}; t \in [0,1]\}$ is a collection of i.i.d. standard Gaussian errors independent of the X_i's. The above equality in only formal (see Section 2 of the Supplementary Material (Chakraborty and Panaretos, 2016) for more details). We have also compared the performance of the estimators under this setting, and the results are reported in the Supplementary Material (Chakraborty

and Panaretos, 2016).

7. Proofs of Formal Statements

In order to prove Theorem 2, we first prove a Lemma that will allow us to connect the Fourier coefficient decay of β, the eigenvalue decay of $\mathscr{K}$, and the ridge parameter ρ.

Lemma 1. *Suppose that $\lambda_1 > \lambda_2 > \ldots > 0$ is a sequence of reals and $\rho > 0$. Assume that the λ_j's satisfy Assumption (A2) in Section 4.2 for some $\alpha > 1$ and for all sufficiently large $j \geq 1$. Let $\{b_j\}_{j\geq 1}$ be another sequence of reals such that $|b_j| \leq j^{-\eta}$ for some $\eta > 0$ and for all sufficiently large j. Then, for any $b \geq a \geq 0$ and any $c \geq 0$ with $2c\eta + a\alpha > 1$, we have*

$$\sum_{j=1}^{\infty} b_j^{2c}\lambda_j^a/(\lambda_j+\rho)^b \leq \text{const.}\rho^{\frac{2c\eta}{\alpha}-b+a-\frac{1}{\alpha}}$$

if $2c\eta < \alpha(b-a)+1$. Further, if $2c\eta > \alpha(b-a)+1$, then $\sup_{\rho>0}\sum_{j=1}^{\infty} b_j^{2c}\lambda_j^a/(\lambda_j+\rho)^b = \sum_{j=1}^{\infty} b_j^{2c}\lambda_j^a/\lambda_j^b < \infty$.

Proof. Consider the case when $2c\eta < \alpha(b-a)+1$, and fix $J = \rho^{-1/\alpha}$. Note that

$$\begin{aligned}\sum_{j>J}\frac{b_j^{2c}\lambda_j^a}{(\lambda_j+\rho)^b} &\leq \text{const.}\rho^{-b}\sum_{j>J} b_j^{2c}\lambda_j^a \leq \text{const.}\rho^{-b}\sum_{j>J} j^{-2c\eta-a\alpha}\\ &\leq \text{const.}\rho^{-b}\int_J^{\infty} x^{-2c\eta-a\alpha}dx \leq \text{const.}\rho^{\frac{2c\eta}{\alpha}-b+a-\frac{1}{\alpha}}.\end{aligned}$$

Also,

$$\begin{aligned}\sum_{j\leq J}\frac{b_j^{2c}\lambda_j^a}{(\lambda_j+\rho)^b} &\leq \text{const.}\sum_{j\leq J} b_j^{2c}\lambda_j^{-b+a} \leq \text{const.}\sum_{j\leq J} j^{-2c\eta+\alpha(b-a)}\\ &\leq \text{const.}\int_0^{J} x^{-2c\eta+\alpha(b-a)}dx \leq \text{const.}\rho^{\frac{2c\eta}{\alpha}-b+a-\frac{1}{\alpha}}.\end{aligned}$$

This completes the proof of the first part of the lemma.

Next consider the case when $2c\eta > \alpha(b-a)+1$. Note that $\sum_{j=1}^{\infty} b_j^{2c}\lambda_j^a/(\lambda_j+\rho)^b \leq \sum_{j=1}^{\infty} b_j^{2c}\lambda_j^{-b+a}$ for all $\rho > 0$. Further,

$$\sum_{j=1}^{\infty} b_j^{2c}\lambda_j^{-b+a} \leq \text{const.}\sum_{j=1}^{\infty} j^{-2c\eta+\alpha(b-a)} < \infty.$$

This proves the second part of the lemma. □

Proof of Theorem 2. As in the oracle case, define $Y_i = \sum_{j=1}^{r}\langle X_i, \phi_j\rangle\phi_j$ and $Z_i = X_i - Y_i$ for all $i = 1, 2, \ldots, n$ (these random variables are not observed in practice). By choice of the Y_i's and the Z_i's, their population covariance operators are $\mathscr{K}_1 = \sum_{j=1}^{r}\lambda_j\phi_j\otimes\phi_j$ and $\mathscr{K}_2 = \sum_{j=r+1}^{r}\lambda_j\phi_j\otimes\phi_j$, respectively. So, the corresponding eigenspaces are orthogonal. Also, define $\mathscr{K}_{\rho,2} = \mathscr{K}_2 + \rho\mathscr{P}_2$.

Now, observe that

$$\hat{\mathscr{K}}_{\rho,2}^{-} - \mathscr{K}_{\rho,2}^{-} = \hat{\mathscr{K}}_{\rho}^{-1} - \mathscr{K}_{\rho}^{-1} + \sum_{j=1}^{r}(\lambda_j+\rho)^{-1}\phi_j\otimes\phi_j - \sum_{j=1}^{r}(\hat{\lambda}_j+\rho)^{-1}\hat{\phi}_j\otimes\hat{\phi}_j.$$

Define $\hat{\mathscr{F}}_a = \sum_{j=1}^{r}(\lambda_j + a)^{-1}\phi_j \otimes \phi_j - \sum_{j=1}^{r}(\hat{\lambda}_j + a)^{-1}\hat{\phi}_j \otimes \hat{\phi}_j$ for any $a \geq 0$. In this notation, $\hat{\mathscr{K}}_1^- - \mathscr{K}_1^- = -\hat{\mathscr{F}}_0$. Also, $\hat{C}_1 - \widetilde{C}_1 = (\hat{\mathscr{P}}_1 - \mathscr{P}_1)\hat{C}$ and $\hat{C}_2 - \widetilde{C}_2 = (\hat{\mathscr{P}}_2 - \mathscr{P}_2)\hat{C} = (\mathscr{P}_1 - \hat{\mathscr{P}}_1)\hat{C}$.

Note that

$$\hat{\beta}_{HR} = \widetilde{\beta}_{HR} + \sum_{l=1}^{8} U_l,$$

where

$$\begin{aligned}
U_1 &= (\hat{\mathscr{K}}_1^- - \mathscr{K}_1^-)(\hat{C}_1 - \widetilde{C}_1) = -\hat{\mathscr{F}}_0(\hat{\mathscr{P}}_1 - \mathscr{P}_1)\hat{C} \\
U_2 &= (\hat{\mathscr{K}}_{\rho,2}^- - \mathscr{K}_{\rho,2}^-)(\hat{C}_2 - \widetilde{C}_2) = (\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1})(\mathscr{P}_1 - \hat{\mathscr{P}}_1)\hat{C} + \hat{\mathscr{F}}_\rho(\mathscr{P}_1 - \hat{\mathscr{P}}_1)\hat{C} \\
U_3 &= \mathscr{K}_1^-(\hat{C}_1 - \widetilde{C}_1) = \mathscr{K}_1^-(\hat{\mathscr{P}}_1 - \mathscr{P}_1)\hat{C} \\
U_4 &= \mathscr{K}_{\rho,2}^-(\hat{C}_2 - \widetilde{C}_2) = \mathscr{K}_{\rho,2}^-(\mathscr{P}_1 - \hat{\mathscr{P}}_1)\hat{C} \\
U_5 &= (\hat{\mathscr{K}}_1^- - \mathscr{K}_1^-)(\widetilde{C}_1 - C_1) = -\hat{\mathscr{F}}_0(\widetilde{C}_1 - C_1) \\
U_6 &= (\hat{\mathscr{K}}_{\rho,2}^- - \mathscr{K}_{\rho,2}^-)(\widetilde{C}_2 - C_2) = (\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1})(\widetilde{C}_2 - C_2) + \hat{\mathscr{F}}_\rho(\widetilde{C}_2 - C_2) \\
U_7 &= (\hat{\mathscr{K}}_1^- - \mathscr{K}_1^-)C_1 = -\hat{\mathscr{F}}_0 C_1 \\
U_8 &= (\hat{\mathscr{K}}_{\rho,2}^- - \mathscr{K}_{\rho,2}^-)C_2 = (\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1})C_2 + \hat{\mathscr{F}}_\rho C_2.
\end{aligned}$$

Putting the pieces together, we get

$$\begin{aligned}
& \mathbb{E}\{(\hat{\beta}_{HR} - \beta) \otimes (\hat{\beta}_{HR} - \beta)\} \\
= \; & \mathbb{E}\{(\widetilde{\beta}_{HR} - \beta) \otimes (\widetilde{\beta}_{HR} - \beta)\} + \mathbb{E}\left\{\left(\sum_{l=1}^{8} U_l\right) \otimes \left(\sum_{l=1}^{8} U_l\right)\right\} \\
& + \mathbb{E}\left\{(\widetilde{\beta}_{HR} - \beta) \otimes \left(\sum_{l=1}^{8} U_l\right)\right\} + \mathbb{E}\left\{\left(\sum_{l=1}^{8} U_l\right) \otimes (\widetilde{\beta}_{HR} - \beta)\right\}.
\end{aligned}$$

So,

$$\mathrm{MSE}(\hat{\beta}_{HR}) = \mathrm{MSE}(\widetilde{\beta}_{HR}) + \mathbb{E}\left(\left\|\sum_{l=1}^{8} U_l\right\|^2\right) + 2\mathbb{E}\left(\langle \sum_{l=1}^{8} U_l, \widetilde{\beta}_{HR} - \beta\rangle\right).$$

Using the Cauchy-Schwarz inequality, one has

$$\begin{aligned}
& |\mathrm{MSE}(\hat{\beta}_{HR}) - \mathrm{MSE}(\widetilde{\beta}_{HR})| \\
& \leq O(1)\left[\sum_{l=1}^{8}\mathbb{E}(\|U_l\|^2) + \mathbb{E}^{1/2}\{\|\widetilde{\beta}_{HR} - \beta\|^2\}\left\{\sum_{l=1}^{8}\mathbb{E}(\|U_l\|^2)\right\}^{1/2}\right].
\end{aligned} \tag{7.1}$$

Note that $\mathbb{E}\{\|\widetilde{\beta}_{HR} - \beta\|^2\} = \mathrm{MSE}(\widetilde{\beta}_{HR})$ so that it can be obtained from the general expression in equation (1.3) in the Supplementary Material (Chakraborty and Panaretos, 2016). The second term in the right hand side of equation (1.3) in the Supplementary Material (Chakraborty and Panaretos, 2016) equals (by the definition of Z_1 and the assumptions in the theorem)

$$n^{-1}\sum_{j=r+1}^{\infty}(\lambda_j + \rho)^{-2}\left[\langle\beta, \phi_j\rangle^2\lambda_j^2 + \lambda_j\{\langle \mathscr{K}\beta, \beta\rangle + \sigma^2\}\right] = O(n^{-1})\left[1 + \sum_{j=r+1}^{\infty}\frac{\lambda_j}{\lambda_j + \rho^2}\right] = \frac{O(1)}{n\rho^{1+\frac{1}{\alpha}}} \tag{7.2}$$

Here, the last equality follows from Lemma 1 by taking $a = 1, b = 2$ and $c = 0$ in the statement of that lemma. It also follows from Lemma 1 by taking $a = 0, b = 2$ and $c = 1$ that

$$\sum_{j=r+1}^{\infty} (\lambda_j + \rho)^{-2} \langle \beta, \phi_j \rangle^2 = O(\rho^L), \tag{7.3}$$

where $L = (2\eta - 1)/\alpha - 2$ or $L = 0$ according as $2\eta < 2\alpha + 1$ or $2\eta > 2\alpha + 1$. Put $m = L + 2$. So, the third term in the right hand side of equation (1.3) in the Supplementary Material (Chakraborty and Panaretos, 2016) is $O(\rho^m)$, where $m = (2\eta - 1)/\alpha$ or $m = 2$ according as $\alpha > \eta - 1/2$ or $\alpha < \eta - 1/2$. Combining this bound with (7.2) and the fact that first term in the right hand side of equation (1.3) in the Supplementary Material (Chakraborty and Panaretos, 2016) is $O(n^{-1})$, we obtain

$$\mathbb{E}\{\|\widetilde{\beta}_{HR} - \beta\|^2\} = O(1) \left\{ \frac{1}{n\rho^{1+\frac{1}{\alpha}}} + \rho^m \right\}, \tag{7.4}$$

where $m = (2\eta - 1)/\alpha$ or $m = 2$ depending on whether $\alpha > \eta - 1/2$ or $\alpha < \eta - 1/2$.

We will now consider bounds for $\mathbb{E}(\|U_l\|^2)$ for $l = 1, 2, \ldots, 8$. First note that for any $a \geq 0$, we have

$$\begin{aligned} \hat{\mathscr{F}}_a &= -\sum_{j=1}^{r} \{(\hat{\lambda}_j + a)^{-1} - (\lambda_j + a)^{-1}\}(\hat{\phi}_j \otimes \hat{\phi}_j) - \\ &\quad \sum_{j=1}^{r} (\lambda_j + a)^{-1} \{\hat{\phi}_j \otimes (\hat{\phi}_j - \phi_j) + (\hat{\phi}_j - \phi_j) \otimes \phi_j\} \\ \Rightarrow \quad |||\hat{\mathscr{F}}_a||| &\leq \sum_{j=1}^{r} |(\hat{\lambda}_j + a)^{-1} - (\lambda_j + a)^{-1}| + 2 \sum_{j=1}^{r} (\lambda_j + a)^{-1} \|\hat{\phi}_j - \phi_j\|. \end{aligned}$$

Some straightforward but tedious moment calculations yield $\mathbb{E}\{|||\hat{\mathscr{K}} - \mathscr{K}|||^8\} = O(n^{-4})$ so that $\mathbb{E}\{|||\hat{\mathscr{K}} - \mathscr{K}|||^4\} = O(n^{-2})$. Thus, using Lemma 2.2 and 2.3 in Horváth and Kokoszka (2012), we have that for any $a \geq 0$

$$\mathbb{E}\{|||\hat{\mathscr{F}}_a|||^4\} = O(n^{-2}) \tag{7.5}$$

as $n \to \infty$. We will use this fact often in the proof. We will also use the fact that

$$\begin{aligned} & \mathbb{E}\{|||\hat{\mathscr{P}}_1 - \mathscr{P}_1|||^8\} \\ \leq \quad & \mathbb{E}\{|||\sum_{j=1}^{r} \{\hat{\phi}_j \otimes (\hat{\phi}_j - \phi_j) + (\hat{\phi}_j - \phi_j) \otimes \phi_j\}|||^8\} \\ \leq \quad & O(1) \sum_{j=1}^{r} \mathbb{E}\{\|\hat{\phi}_j - \phi_j\|^8\} \;\leq\; O(1)\mathbb{E}\{|||\hat{\mathscr{K}} - \mathscr{K}|||^8\} = O(n^{-4}) \end{aligned} \tag{7.6}$$

as $n \to \infty$. The third inequality follows from Lemma 2.3 in Horváth and Kokoszka (2012).

Note that $\mathbb{E}(\|U_1\|^2) \leq O(1)E^{1/2}\{|||\hat{\mathscr{F}}_0|||_\infty^4\}E^{1/4}\{|||\hat{\mathscr{P}}_1 - \mathscr{P}_1|||_\infty^8\}E^{1/4}\{\|\hat{C}\|^8\}$. It directly follows that $\mathbb{E}\{\|\widetilde{C}\|^8\} = O(1)$ as $n \to \infty$. Thus using (7.5) and (7.6) along with the fact that the operator norm is bounded above by the Hilbert-Schmidt norm, we have

$$\mathbb{E}(\|U_1\|^2) = O(n^{-2}) \tag{7.7}$$

as $n \to \infty$.

Next note that $\mathbb{E}(\|U_3\|^2) \leq \left\|\mathscr{K}_1^-\right\|_\infty^2 E^{1/2}\{\left\|\hat{\mathscr{P}}_1 - \mathscr{P}_1\right\|^4\} E^{1/2}\{\|\hat{C}\|^4\}$. Using the fact that $\left\|\mathscr{K}_1^-\right\|_\infty = \lambda_r^{-1}$, we get that

$$\mathbb{E}(\|U_3\|^2) = O(n^{-1}) \tag{7.8}$$

as $n \to \infty$.

Next note that $\mathbb{E}(\|U_5\|^2) \leq E^{1/2}\{\left\|\hat{\mathscr{F}}_0\right\|^4\} E^{1/2}\{\|\widetilde{C}_1 - C_1\|^4\}$. It is easy to show that $\mathbb{E}\{\|\widetilde{C}_1 - C_1\|^4\} = O(n^{-2})$ as $n \to \infty$. So, it follows from (7.5) that

$$\mathbb{E}(\|U_5\|^2) = O(n^{-2}). \tag{7.9}$$

Similar calculations also show that

$$\mathbb{E}(\|U_7\|^2) = O(n^{-1}) \tag{7.10}$$

as $n \to \infty$. Next, observe that

$$\begin{aligned} \mathbb{E}(\|U_6\|^2) \leq\ & 2E^{1/2}\{\left\|\hat{\mathscr{F}}_\rho\right\|^4\} E^{1/2}\{\|\widetilde{C}_2 - C_2\|^4\} \\ & + 2E\{\|(\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1})(\widetilde{C}_2 - C_2)\|^2\}. \end{aligned} \tag{7.11}$$

From the fact that $\mathbb{E}\{\|\widetilde{C}_2 - C_2\|^4\} = O(n^{-2})$ as $n \to \infty$ and using (7.5), it follows that the first term on the right hand side of (7.11) is $O(n^{-2})$ as $n \to \infty$. Further,

$$\begin{aligned} & E\{\|(\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1})(\widetilde{C}_2 - C_2\|^2\} \\ \leq\ & E\{\left\|\mathscr{K}_\rho^{-1}\right\|_\infty \|(\hat{\mathscr{K}} - \mathscr{K})\mathscr{K}_\rho^{-1}(\widetilde{C}_2 - C_2\|^2\} \\ \leq\ & \rho^{-2} E^{1/2}\{\left\|\hat{\mathscr{K}} - \mathscr{K}\right\|^4\} E^{1/2}\{\|\mathscr{K}_\rho^{-1}(\widetilde{C}_2 - C_2\|^4\} \\ \leq\ & O(n^{-1}\rho^{-2}) \left[E^{1/2}\{\|\mathscr{K}_{\rho,1}^-(\widetilde{C}_2 - C_2\|^4\} + E^{1/2}\{\|\mathscr{K}_{\rho,2}^-(\widetilde{C}_2 - C_2\|^4\}\right], \end{aligned} \tag{7.12}$$

where $\mathscr{K}_{\rho,1}^- = \sum_{j=1}^r (\lambda_j + \rho)^{-1}(\phi_j \otimes \phi_j)$. The last inequality also uses the bound for $\mathbb{E}\{\left\|\hat{\mathscr{K}} - \mathscr{K}\right\|_\infty^4\}$ obtained for deriving (7.19).

Now, using the fact that for any $j = r+1, r+2, \ldots$, we have $E\{\langle \widetilde{C}_2 - C_2, \phi_j\rangle^4\} = O(n^{-2}) E^2\{\langle y_1 Z_1 - \mathscr{K}_2\beta, \phi_j\rangle^2\} = O(n^{-2})\{\langle \beta, \phi_j\rangle^2 \lambda_j^2 + \lambda_j(\sigma^2 + \langle \mathscr{K}\beta, \beta\rangle)\}^2$, we get that

$$\begin{aligned} & E\{\|\mathscr{K}_{\rho,2}^-(\widetilde{C}_2 - C_2\|^4\} \\ =\ & E\left[\left\{\sum_{j=r+1}^\infty (\lambda_j + \rho)^{-2}\langle \widetilde{C}_2 - C_2, \phi_j\rangle^2\right\}^2\right] \\ =\ & E\left\{\sum_{j_1, j_2 = r+1}^\infty (\lambda_{j_1} + \rho)^{-2}(\lambda_{j_2} + \rho)^{-2}\langle \widetilde{C}_2 - C_2, \phi_{j_1}\rangle^2 \langle \widetilde{C}_2 - C_2, \phi_{j_2}\rangle^2\right\} \\ \leq\ & \sum_{j_1, j_2 = r+1}^\infty (\lambda_{j_1} + \rho)^{-2}(\lambda_{j_2} + \rho)^{-2} E^{1/2}\{\langle \widetilde{C}_2 - C_2, \phi_{j_1}\rangle^4\} E^{1/2}\{\langle \widetilde{C}_2 - C_2, \phi_{j_2}\rangle^4\} \\ \leq\ & O(n^{-2}) \left[\sum_{j=r+1}^\infty (\lambda_j + \rho)^{-2}\{\langle \beta, \phi_j\rangle^2 \lambda_j^2 + \lambda_j(\sigma^2 + \langle \mathscr{K}\beta, \beta\rangle)\}\right]^2 \end{aligned}$$

$$\leq \quad O(n^{-2}) \left[1 + \left\{ \sum_{j=r+1}^{\infty} (\lambda_j + \rho)^{-2} \lambda_j \right\}^2 \right] = O(n^{-2} \rho^{-2-2/\alpha}),$$

by an application of Lemma 1. Now, using (7.11) and (7.12), we get that

$$\mathbb{E}(\|U_6\|^2) = o(n^{-1} \rho^{-1-1/\alpha}) \tag{7.13}$$

as $n \to \infty$.

Next note that $\mathbb{E}(\|U_8\|^2) \leq \mathbb{E}\{\|(\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1}) C_2\|^2\} + \mathbb{E}\{\|\hat{\mathscr{F}}_\rho C_2\|^2\}$. From earlier calculations and using (7.5), it follows that $\mathbb{E}\{\|\hat{\mathscr{F}}_\rho C_2\|^2\} = O(n^{-1})$ as $n \to \infty$. Next, note that $\mathbb{E}\{\|(\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1}) C_2\|^2\} = \mathbb{E}\{\|\hat{\mathscr{K}}_\rho^{-1}(\hat{\mathscr{K}} - \mathscr{K}) \mathscr{K}_\rho^{-1} C_2\|^2\} \leq E^{1/2}\{\|\hat{\mathscr{K}}_\rho^{-1} \mathscr{K}_\rho\|_\infty^4\} E^{1/2}\{\|\mathscr{K}_\rho^{-1}(\hat{\mathscr{K}} - \mathscr{K}) \mathscr{K}_\rho^{-1} C_2\|^4\}$. Observe that $\|\hat{\mathscr{K}}_\rho^{-1} \mathscr{K}_\rho\|_\infty = \|\hat{\mathscr{K}}_\rho^{-1}(\mathscr{K}_\rho - \hat{\mathscr{K}}_\rho) + \mathscr{I}\|_\infty \leq \rho^{-1} \|\mathscr{K} - \hat{\mathscr{K}}\|_\infty + 1 \leq \rho^{-1} \|\hat{\mathscr{K}} - \mathscr{K}\| + 1$. So, we have $\mathbb{E}\{\|\hat{\mathscr{K}}_\rho^{-1} \mathscr{K}_\rho\|_\infty^4\} \leq 1 + \rho^{-4} \mathbb{E}\{\|\hat{\mathscr{K}} - \mathscr{K}\|^4\} = 1 + O(n^{-2} \rho^{-4}) = O(1)$ since $n\rho^2 \to \infty$. We have

$$\begin{aligned}
\mathbb{E}\{\|\mathscr{K}_\rho^{-1}(\hat{\mathscr{K}} - \mathscr{K}) \mathscr{K}_\rho^{-1} C_2\|^4\} &= \mathbb{E} \left[\sum_{j=1}^{\infty} \langle \mathscr{K}_\rho^{-1}(\hat{\mathscr{K}} - \mathscr{K}) \mathscr{K}_\rho^{-1} C_2, \phi_j \rangle^2 \right]^2 \\
&= \mathbb{E} \left[\sum_{j=1}^{\infty} \langle (\hat{\mathscr{K}} - \mathscr{K}) \mathscr{K}_\rho^{-1} C_2, \mathscr{K}_\rho^{-1} \phi_j \rangle^2 \right]^2 \\
&= \mathbb{E} \left[\sum_{j=1}^{\infty} \langle (\hat{\mathscr{K}} - \mathscr{K}) \mathscr{K}_\rho^{-1} \mathscr{K}_2 \beta, (\lambda_j + \rho)^{-1} \phi_j \rangle^2 \right]^2
\end{aligned}$$

We denote the above expectation by T. Now,

$$\begin{aligned}
T = \sum_{j_1, j_2 = 1}^{\infty} \sum_{l_1, l_2, l_3, l_4 = r+1}^{\infty} \Bigg[& \frac{\prod_{u=1}^4 (\langle \beta, \phi_{l_u} \rangle \lambda_{l_u})}{(\lambda_{j_1} + \rho)^2 (\lambda_{j_2} + \rho)^2 \prod_{u=1}^4 (\lambda_{l_u} + \rho)} \times \\
& \mathbb{E} \left\{ \prod_{i=1}^{2} \langle (\hat{\mathscr{K}} - \mathscr{K}) \phi_{j_1}, \phi_{l_u} \rangle \prod_{i=3}^{4} \langle (\hat{\mathscr{K}} - \mathscr{K}) \phi_{j_2}, \phi_{l_u} \rangle \right\} \Bigg].
\end{aligned} \tag{7.14}$$

Direct calculation yields that if $j_1 = j_2$ in the expression of T above, then

$$\begin{aligned}
& \mathbb{E} \left\{ \prod_{i=1}^{2} \langle (\hat{\mathscr{K}} - \mathscr{K}) \phi_{j_1}, \phi_{l_u} \rangle \prod_{i=3}^{4} \langle (\hat{\mathscr{K}} - \mathscr{K}) \phi_{j_2}, \phi_{l_u} \rangle \right\} \\
\leq \quad & O(n^{-2}) \left\{ [\lambda_{j_1}^2 \mathbf{1}\{j_1 = l_1 = l_2\} + \lambda_{j_1} \lambda_{l_1} \mathbf{1}\{j_1 \neq l_1 = l_2\}] \times \right. \\
& \left. [\lambda_{j_1}^2 \mathbf{1}\{j_1 = l_3 = l_4\} + \lambda_{j_1} \lambda_{l_3} \mathbf{1}\{j_1 \neq l_3 = l_4\}] \right\}.
\end{aligned}$$

On the other hand if $j_1 \neq j_2$, then

$$\begin{aligned}
& \mathbb{E} \left\{ \prod_{i=1}^{2} \langle (\hat{\mathscr{K}} - \mathscr{K}) \phi_{j_1}, \phi_{l_u} \rangle \prod_{i=3}^{4} \langle (\hat{\mathscr{K}} - \mathscr{K}) \phi_{j_2}, \phi_{l_u} \rangle \right\} \\
\leq \quad & O(n^{-2}) \left\{ [\lambda_{j_1}^2 \mathbf{1}\{j_1 = l_1 = l_2\} + \lambda_{j_1} \lambda_{l_1} \mathbf{1}\{j_1 \neq l_1 = l_2\}] \times \right.
\end{aligned}$$

$$[\lambda_{j_2}^2 \mathbf{1}\{j_2 = l_3 = l_4\} + \lambda_{j_2}\lambda_{l_3}\mathbf{1}\{j_2 \neq l_3 = l_4\}]$$
$$+ \lambda_{j_1}^2\lambda_{j_2}^2\mathbf{1}\{j_1 = l_3 = l_4\}\mathbf{1}\{j_2 = l_1 = l_2\}\Big\}$$

So, we have

$$T = O(n^{-2})\Bigg[\left(\sum_{j_1=1}^{r}\frac{\lambda_{j_1}}{(\lambda_{j_1}+\rho)^2}\right)^2 \times \left(\sum_{l_1,l_2=r+1}^{\infty}\frac{\langle\beta,\phi_{l_1}\rangle^2\langle\beta,\phi_{l_2}\rangle^2\lambda_{l_1}^3\lambda_{l_3}^2}{(\lambda_{l_1}+\rho)^2(\lambda_{l_2}+\rho)^2}\right) +$$
$$2\sum_{j_1=1}^{r}\sum_{j_2=r+1}^{\infty}\sum_{l_1=r+1}^{\infty}\frac{\langle\beta,\phi_{l_1}\rangle^2\langle\beta,\phi_{j_2}\rangle^2\lambda_{l_1}^3\lambda_{j_2}^4\lambda_{j_1}}{(\lambda_{l_1}+\rho)^2(\lambda_{j_1}+\rho)^2(\lambda_{j_2}+\rho)^4} +$$
$$4\sum_{j_1=1}^{r}\sum_{j_2=r+1}^{\infty}\sum_{\substack{l_1,l_2=r+1\\ l_2\neq j_2}}^{\infty}\frac{\langle\beta,\phi_{l_1}\rangle^2\langle\beta,\phi_{l_2}\rangle^2\lambda_{l_1}^3\lambda_{l_2}^3\lambda_{j_1}\lambda_{j_2}}{(\lambda_{l_1}+\rho)^2(\lambda_{j_1}+\rho)^2(\lambda_{l_2}+\rho)^2(\lambda_{j_2}+\rho)^2} +$$
$$\sum_{j_1=r+1}^{\infty}\frac{\langle\beta,\phi_{j_1}\rangle^4\lambda_{j_1}^4}{(\lambda_{j_1}+\rho)^8} + 2\sum_{\substack{j_1,l_1=r+1\\ j_1\neq l_1}}^{\infty}\frac{\langle\beta,\phi_{j_1}\rangle^2\langle\beta,\phi_{l_1}\rangle^2\lambda_{j_1}^5\lambda_{l_1}^3}{(\lambda_{j_1}+\rho)^6(\lambda_{l_1}+\rho)^2} +$$
$$\sum_{\substack{j_1,l_1,l_2=r+1\\ j_1\neq l_1,j_1\neq l_2}}^{\infty}\frac{\langle\beta,\phi_{l_1}\rangle^2\langle\beta,\phi_{l_2}\rangle^2\lambda_{j_1}^2\lambda_{l_1}^3\lambda_{l_2}^3}{(\lambda_{j_1}+\rho)^4(\lambda_{l_1}+\rho)^2(\lambda_{l_2}+\rho)^2} +$$
$$2\sum_{\substack{j_1,j_2=r+1\\ j_1\neq j_2}}^{\infty}\frac{\langle\beta,\phi_{j_1}\rangle^2\langle\beta,\phi_{j_2}\rangle^2\lambda_{j_1}^4\lambda_{j_2}^4}{(\lambda_{j_1}+\rho)^4(\lambda_{j_2}+\rho)^4} +$$
$$2\sum_{\substack{j_1,j_2,l_1=r+1\\ j_1\neq j_2,j_2\neq l_1}}^{\infty}\frac{\langle\beta,\phi_{j_1}\rangle^2\langle\beta,\phi_{l_2}\rangle^2\lambda_{j_1}^4\lambda_{j_2}\lambda_{l_2}^3}{(\lambda_{j_1}+\rho)^4(\lambda_{j_2}+\rho)^2(\lambda_{l_2}+\rho)^2} +$$
$$2\sum_{\substack{j_1,j_2,l_1,l_2=r+1\\ j_1\neq j_2,j_1\neq l_1,j_2\neq l_2}}^{\infty}\frac{\langle\beta,\phi_{l_1}\rangle^2\langle\beta,\phi_{l_2}\rangle^2\lambda_{j_1}\lambda_{j_2}\lambda_{l_1}^3\lambda_{l_2}^3}{(\lambda_{j_1}+\rho)^2(\lambda_{j_2}+\rho)^2(\lambda_{l_1}+\rho)^2(\lambda_{l_2}+\rho)^2}\Bigg].$$

Using the simple bound that $\langle\beta,\phi_l\rangle^2 \leq \|\beta\|^2$ and applying Lemma 1 to the above expression with $c = 0$ and appropriately chosen a and b for each infinite sum, we get that $T = O(n^{-2}\rho^{-2-2/\alpha})$ as $n \to \infty$. This together with the fact that $\mathbb{E}\{\|\!|\hat{\mathscr{K}}_\rho^{-1}\mathscr{K}_\rho\|\!|_\infty^4\} = O(1)$ as $n \to \infty$ implies that $\mathbb{E}\{\|(\hat{\mathscr{K}}_\rho^{-1} - \mathscr{K}_\rho^{-1})C_2\|^2\} = O(n^{-1}\rho^{-1-1/\alpha})$ as $n \to \infty$. So, we have

$$\mathbb{E}(\|U_8\|^2) = O(n^{-1}\rho^{-1-1/\alpha}) \tag{7.15}$$

as $n \to \infty$.

We now turn to controlling $\mathbb{E}(\|U_4\|^2)$. First we decompose U_4 as

$$U_4 = \mathscr{K}_{\rho,2}^{-}(\mathscr{P}_1 - \hat{\mathscr{P}}_1)(\hat{C} - C) + \mathscr{K}_{\rho,2}^{-}(\mathscr{P}_1 - \hat{\mathscr{P}}_1)C,$$

and denote the first and the second terms by U_{41} and U_{42}, respectively. Calculations similar to those carried out earlier yield $\mathbb{E}(\|U_{41}\|^2) = O(n^{-2}\rho^{-2})$ as $n \to \infty$.

To bound $\mathbb{E}(\|U_{42}\|^2)$, set $M_n^2 = An^{-1}\rho^{-2}$ for some $A > 0$, and define the set

$$G_n = \left\{ \max_{j=1,2,\ldots,r} |\hat{\lambda}_j - \lambda_j| \leq M_n \right\}.$$

Since $\max_{j=1,2,\ldots,r} \mathbb{E}\{(\hat{\lambda}_j - \lambda_j)^2\} = O(n^{-1})$ as $n \to \infty$, Markov's inequality yields$P(G_n^c) < \rho^2$ as $n \to \infty$ for an appropriate choice of A. Thus, $\mathbb{E}\{\|U_{42}\|^2 \mathbf{1}(G_n^c)\} \leq \rho^{-2} E^{1/2}\{|||\hat{\mathscr{P}}_1 - \mathscr{P}_1|||^4\}\sqrt{P(G_n^c)} \leq \rho^{-1} E^{1/2}\{|||\hat{\mathscr{K}} - \mathscr{K}|||^4\} = O(n^{-1}\rho^{-1}) = o(n^{-1}\rho^{-1-1/\alpha})$ as $n \to \infty$. Consequently, it suffices to bound $\mathbb{E}\{\|U_{42}\|^2 \mathbf{1}(G_n)\}$. Using the resolvent formalism, we represent $\mathscr{P}_1$ as

$$\mathscr{P}_1 = \frac{1}{2\pi i} \int_\Gamma (\mathscr{K} - z\mathscr{I})^{-1} dz,$$

where $i^2 = -1$ and Γ is the boundary of a closed disk containing $\{\lambda_j : j = 1, ..., r\}$ and excluding $\{\lambda_j : j > r\}$ (see Hsing and Eubank (2015)). Similarly,

$$\hat{\mathscr{P}}_1 = \frac{1}{2\pi i} \int_{\hat{\Gamma}} (\hat{\mathscr{K}} - z\mathscr{I})^{-1} dz,$$

where $\hat{\Gamma}$ is the boundary of a closed disk containing $\{\hat{\lambda}_j : j = 1, ..., r\}$ and excluding $\{\hat{\lambda}_j : j > r\}$. Since $M_n \to 0$ as $n \to \infty$, so for all sufficiently large n, $M_n < (\lambda_r - \lambda_{r+1})/4$. Thus, for all sufficiently large n, $\hat{\Gamma}$ can be chosen to be Γ for all sample points in the set G_n. Thus, for all sufficiently large n, we have

$$\begin{aligned}
& \mathbb{E}\{\|U_{42}\|^2 \mathbf{1}(G_n)\} \\
= \; & \mathbb{E}\left\{ \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- [(\hat{\mathscr{K}} - z\mathscr{I})^{-1} - (\mathscr{K} - z\mathscr{I})^{-1}] C dz \right\|^2 \mathbf{1}(G_n) \right\} \\
\leq \; & \mathbb{E}\left\{ \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- (\hat{\mathscr{K}} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \mathbf{1}(G_n) \right\} \\
\leq \; & 2\mathbb{E}\left\{ \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- [(\hat{\mathscr{K}} - z\mathscr{I})^{-1} - (\mathscr{K} - z\mathscr{I})^{-1}](\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \right\} \\
& + 2\mathbb{E}\left\{ \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- (\mathscr{K} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \right\} \\
= \; & 2\mathbb{E}\left\{ \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- (\hat{\mathscr{K}} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \right\} \quad (7.16) \\
& + 2\mathbb{E}\left\{ \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- (\mathscr{K} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \right\}
\end{aligned}$$

Now note that

$$\begin{aligned}
& \left\| \frac{1}{2\pi i} \int_\Gamma \mathscr{K}_{\rho,2}^- (\hat{\mathscr{K}} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \\
\leq \; & |||\mathscr{K}_{\rho,2}^-|||_\infty^2 \left\| \frac{1}{2\pi i} \int_\Gamma (\hat{\mathscr{K}} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C dz \right\|^2 \\
\leq \; & \rho^{-2} \frac{L^2}{4\pi^2} \sup_\Gamma \left| (\hat{\mathscr{K}} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} (\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C \right|^2,
\end{aligned}$$

where L denotes the arc length of the contour Γ. This last inequality follows from properties of complex contour integrals (see Conway (1978)). Now let us note that

$$\begin{aligned} & \sup_{\Gamma} \left| (\hat{\mathscr{K}} - z\mathscr{I})^{-1}(\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1}(\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C \right|^2 \\ \leq & \; |||\hat{\mathscr{K}} - \mathscr{K}|||^4 \sup_{\Gamma} |||(\hat{\mathscr{K}} - z\mathscr{I})^{-1}|||_\infty \} |||(\mathscr{K} - z\mathscr{I})^{-1}|||_\infty^2 \\ = & \; |||\hat{\mathscr{K}} - \mathscr{K}|||^4 \sup_{\Gamma} |z|^{-3} \leq \text{const.} |||\hat{\mathscr{K}} - \mathscr{K}|||^4, \end{aligned}$$

where the last inequality follows because Γ only encompasses $\lambda_1, \lambda_2, \ldots, \lambda_r$ and all of them are bounded away from zero. Thus, from the above facts, it follows that the first expectation in the right hand side of (7.16) is bounded above by $O(1)\rho^{-2}\mathbb{E}\{|||\hat{\mathscr{K}} - \mathscr{K}|||^4\} = O(n^{-2}\rho^{-2})$ as $n \to \infty$.

We continue by noting that

$$\begin{aligned} & \mathbb{E} \left| \frac{1}{2\pi i} \int_\Gamma \langle (\mathscr{K} - z\mathscr{I})^{-1}(\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z\mathscr{I})^{-1} C, \phi_j \rangle dz \right|^2 \\ = & \; \frac{1}{(2\pi i)^2} \int_\Gamma \int_\Gamma \mathbb{E} \Big\{ \langle (\mathscr{K} - z_1\mathscr{I})^{-1}(\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z_1\mathscr{I})^{-1} \mathscr{K}\beta, \phi_j \rangle \times \\ & \qquad \langle (\mathscr{K} - z_2\mathscr{I})^{-1}(\hat{\mathscr{K}} - \mathscr{K})(\mathscr{K} - z_2\mathscr{I})^{-1} \mathscr{K}\beta, \phi_j \rangle dz_1 dz_2 \Big\} \\ = & \; \frac{1}{(2\pi i)^2} \int_\Gamma \int_\Gamma (\lambda_j - z_1)^{-1}(\lambda_j - z_2)^{-1} \sum_{l_1, l_2 = 1}^{\infty} \frac{\langle \beta, \phi_{l_1} \rangle \langle \beta, \phi_{l_2} \rangle \lambda_{l_1} \lambda_{l_2}}{(\lambda_{l_1} - z_1)(\lambda_{l_2} - z_2)} \times \\ & \qquad \mathbb{E}\{ \langle (\hat{\mathscr{K}} - \mathscr{K})\phi_{l_1}, \phi_j \rangle \langle (\hat{\mathscr{K}} - \mathscr{K})\phi_{l_2}, \phi_j \rangle \} dz_1 dz_2 \\ = & \; \frac{1}{(2\pi i)^2} \int_\Gamma \int_\Gamma (\lambda_j - z_1)^{-1}(\lambda_j - z_2)^{-1} \sum_{l_1, l_2 = 1}^{\infty} \frac{\langle \beta, \phi_{l_1} \rangle \langle \beta, \phi_{l_2} \rangle \lambda_{l_1} \lambda_{l_2}}{(\lambda_{l_1} - z_1)(\lambda_{l_2} - z_2)} \times \\ & \qquad n^{-1} \{ \lambda_j^2 \mathbf{1}\{j = l_1 = l_2\} + \lambda_j \lambda_l \mathbf{1}\{j \neq l_1 = l_2\} \} dz_1 dz_2 \\ = & \; \frac{n^{-1}}{(2\pi i)^2} \int_\Gamma \int_\Gamma (\lambda_j - z_1)^{-1}(\lambda_j - z_2)^{-1} \Bigg\{ \sum_{l=1}^{r} \frac{\langle \beta, \phi_l \rangle^2 \rangle \lambda_l^3 \lambda_j}{(\lambda_l - z_1)(\lambda_l - z_2)} + \\ & \qquad \sum_{l=r+1}^{\infty} \frac{\langle \beta, \phi_l \rangle^2 \rangle \lambda_l^2 [\lambda_j^2 \mathbf{1}\{j = l_1 = l_2\} + \lambda_j \lambda_l \mathbf{1}\{j \neq l_1 = l_2\}]}{(\lambda_l - z_1)(\lambda_l - z_2)} \Bigg\} dz_1 dz_2. \end{aligned}$$

Since Γ does not contain $\lambda_{r+1}, \lambda_{r+2}, \ldots$, it follows by the Cauchy integral theorem (see Conway (1978)) that when l and j varies over $r+1, r+2, \ldots$, we have

$$\int_\Gamma (\lambda_l - z)^{-1}(\lambda_j - z)^{-1} dz = 0.$$

Furthermore, for any $l = 1, 2, \ldots, r$, we have

$$\begin{aligned} & \frac{1}{2\pi i} \int_\Gamma \frac{dz}{(\lambda_j - z)(\lambda_l - z)} \\ = & \; \frac{1}{2\pi i} \int_\Gamma \left(\frac{1}{\lambda_j - z} - \frac{1}{\lambda_l - z} \right) dz \times \frac{1}{\lambda_l - \lambda_j} \\ = & \; -\frac{1}{\lambda_l - \lambda_j} \times \frac{1}{2\pi i} \int_\Gamma \frac{dz}{\lambda_l - z} = -\frac{1}{\lambda_l - \lambda_j}. \end{aligned}$$

The third inequality follows from Cauchy integral theorem along with the fact that Γ does not contain $\lambda_{r+1}, \lambda_{r+2}, \ldots$ and the fact that j varies over $r+1, r+2, \ldots$. The last equality follows from Cauchy formula (see Conway (1978)), stating that the integral is the winding number of Γ around λ_l, which equals one.

Combining all of the above facts, we finally deduce

$$\begin{aligned}\mathbb{E}\{\|U_{42}\|^2\mathbf{1}(G_n)\} &\leq \sum_{j=r+1}^{\infty}(\lambda_j+\rho)^{-2}\sum_{l=1}^{r}\frac{\langle\beta,\phi_l\rangle^2\rangle\lambda_l^3\lambda_j}{n(\lambda_l-\lambda_j)^2}\\ &\leq n^{-1}\sum_{j=r+1}^{\infty}\frac{\lambda_j}{(\lambda_j+\rho)^2}\left[\sum_{l=1}^{r}\frac{\langle\beta,\phi_l\rangle^2\rangle\lambda_l^3}{n(\lambda_l-\lambda_{r+1})^2}\right]\\ &= O(n^{-1}\rho^{-1-1/\alpha})\end{aligned}$$

as $n\to\infty$ by using (7.5). Thus, we have

$$\mathbb{E}(\|U_4\|^2) = O(n^{-1}\rho^{-1-1/\alpha}) \tag{7.17}$$

as $n\to\infty$.

Finally, we provide a bound for $\mathbb{E}(\|U_2\|^2)$. Note that $\mathbb{E}(\|U_2\|^2) \leq 2\mathbb{E}\{\|(\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)\hat{C}\|^2\} + 2\mathbb{E}\{\|\hat{\mathscr{F}}_\rho(\hat{\mathscr{P}}_1-\mathscr{P}_1)\hat{C}\|^2\}$.

Similar arguments as above show that $\mathbb{E}\{\|\hat{\mathscr{F}}_\rho(\hat{\mathscr{P}}_1-\mathscr{P}_1)\hat{C}\|^2\}$ is $O(n^{-2})$ as $n\to\infty$. Further, $\mathbb{E}\{\|(\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)\hat{C}\|^2\} \leq 2\mathbb{E}\{\|(\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)(\hat{C}-C)\|^2\} + \mathbb{E}\{\|(\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^2\}$. Now,

$$\begin{aligned}&\mathbb{E}\{\|(\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)(\hat{C}-C)\|^2\}\\ \leq\ & E^{1/2}\{|||\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1}|||_\infty^4\}E^{1/4}\{|||\hat{\mathscr{P}}_1-\mathscr{P}_1|||_\infty^8\}E^{1/4}\{\|\hat{C}-C\|^8\}\\ \leq\ & O(n^{-3}\rho^{-4}) = o(n^{-1}\rho^{-1-1/\alpha})\end{aligned}$$

as $n\to\infty$ by using (7.6), the fact that $E\{\|\hat{C}-C\|^8\} = O(n^{-4})$ as $n\to\infty$ and arguments similar to those used earlier. Next,

$$\begin{aligned}&\mathbb{E}\{\|(\hat{\mathscr{K}_\rho}^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^2\}\\ \leq\ & \mathbb{E}\{\|\hat{\mathscr{K}_\rho}^{-1}(\hat{\mathscr{K}}-\mathscr{K})\mathscr{K}_\rho^{-1}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^2\}\\ \leq\ & \rho^{-2}\mathbb{E}^{1/2}\{|||\mathscr{K}-\hat{\mathscr{K}}|||^4\}\mathbb{E}^{1/2}\{\|\mathscr{K}_\rho^{-1}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\}\\ \leq\ & O(n^{-1}\rho^{-2})\mathbb{E}^{1/2}\{\|\mathscr{K}_\rho^{-1}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\}.\end{aligned} \tag{7.18}$$

Now,

$$\begin{aligned}&\mathbb{E}^{1/2}\{\|\mathscr{K}_\rho^{-1}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\}\\ \leq\ & O(1)[\mathbb{E}^{1/2}\{\|\mathscr{K}_{\rho,1}^{-}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\} + \mathbb{E}^{1/2}\{\|\mathscr{K}_{\rho,2}^{-}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\}].\end{aligned}$$

The first term on the right hand side of the above inequality is $O(n^{-1})$ as $n\to\infty$ and we need to bound the term $\mathbb{E}\{\|\mathscr{K}_{\rho,2}^{-}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\}$. To do this, we will follows the same arguments as those used to bound $\mathbb{E}(\|U_{42}\|^2)$ earlier.

Proceeding as in the case of bounding $\mathbb{E}(\|U_{42}\|^2)$, it is easy to see that to obtain a bound for $\mathbb{E}\{\|\mathscr{K}_{\rho,2}^{-}(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^4\}$, it is enough to obtain a bound for $\mathbb{E}\{(2\pi i)^{-1}\int_\Gamma \mathscr{K}_{\rho,2}^{-}(\mathscr{K}-z\mathscr{I})^{-1}(\hat{\mathscr{K}}-\mathscr{K})(\mathscr{K}-z\mathscr{I})^{-1}Cdz\|^4\}$,

where Γ is the same contour as considered in the case of $\mathbb{E}(\|U_{42}\|^2)$. Now, expanding the latter term, we get that

$$\begin{aligned}
&\mathbb{E}\left\{\left\|(2\pi i)^{-1}\int_\Gamma \mathscr{K}^-_{\rho,2}(\mathscr{K}-z\mathscr{I})^{-1}(\hat{\mathscr{K}}-\mathscr{K})(\mathscr{K}-z\mathscr{I})^{-1}Cdz\right\|^4\right\}\\
=\;& \frac{1}{(2\pi i)^4}\sum_{j_1,j_2=r+1}^{\infty}\sum_{l_1,l_2,l_3,l_4=1}^{\infty}\{(\lambda_{j_1}+\rho)^{-2}(\lambda_{j_2}+\rho)^{-2}\times\\
&\int_\Gamma\int_\Gamma\int_\Gamma\int_\Gamma\prod_{u=1}^{2}\frac{\langle\beta,\phi_{l_u}\rangle\lambda_{l_u}}{(\lambda_{l_u}-z_u)(\lambda_{j_1}-z_u)}\prod_{u=3}^{4}\frac{\langle\beta,\phi_{l_u}\rangle\lambda_{l_u}}{(\lambda_{l_u}-z_u)(\lambda_{j_2}-z_u)}S\prod_{u=1}^{4}dz_{l_u}\Bigg\},
\end{aligned}$$

where

$$S=\mathbb{E}\left\{\prod_{u=1}^{2}\langle(\hat{\mathscr{K}}-\mathscr{K})\phi_{l_u},\phi_{j_1}\rangle\prod_{u=3}^{4}\langle(\hat{\mathscr{K}}-\mathscr{K})\phi_{l_u},\phi_{j_2}\rangle\right\}.$$

We obtained the expression of S after (7.14) while bounding $\mathbb{E}(\|U_8\|^2)$ earlier. Plugging-in those expressions and using the Cauchy integral theorem arguments used while bounding $\mathbb{E}(\|U_{42}\|^2)$ earlier, we get that

$$\begin{aligned}
&\mathbb{E}\left\{\left\|(2\pi i)^{-1}\int_\Gamma \mathscr{K}^-_{\rho,2}(\mathscr{K}-z\mathscr{I})^{-1}(\hat{\mathscr{K}}-\mathscr{K})(\mathscr{K}-z\mathscr{I})^{-1}Cdz\right\|^4\right\}\\
\le\;& O(n^{-2})\sum_{j_1,j_2=r+1}^{\infty}\sum_{l_1,l_2=1}^{r}\frac{\langle\beta,\phi_{l_1}\rangle^2\langle\beta,\phi_{l_2}\rangle^2\lambda_{l_1}^3\lambda_{l_2}^3\{\lambda_{j_1}^2I(j_1=j_2)+\lambda_{j_1}\lambda_{j_2}I(j_1\neq j_2)\}}{(\lambda_{j_1}+\rho)^2(\lambda_{j_2}+\rho)^2\prod_{u=1}^{2}(\lambda_{l_u}-\lambda_{j_1})\prod_{u=3}^{4}(\lambda_{l_u}-\lambda_{j_2})}\\
\le\;& O(n^{-2})\left[\sum_{j=r+1}^{\infty}\frac{\lambda_j}{(\lambda_j+\rho)^2}\right]^2 \;=\; O(n^{-2}\rho^{-2-2/\alpha})
\end{aligned}$$

as $n\to\infty$ by Lemma 1. Thus, it follows from (7.18) that $\mathbb{E}\{\|(\hat{\mathscr{K}}_\rho^{-1}-\mathscr{K}_\rho^{-1})(\mathscr{P}_1-\hat{\mathscr{P}}_1)C\|^2\}=o(n^{-1}\rho^{-1-1/\alpha})$ and hence

$$\mathbb{E}(\|U_2\|^2)=o(n^{-1}\rho^{-1-1/\alpha}) \tag{7.19}$$

as $n\to\infty$.

The bound for $|\mathrm{MSE}(\hat{\beta}_{HR})-\mathrm{MSE}(\widetilde{\beta}_{HR})|$ given in the statement of Theorem 2 now follows from (7.1) and using the bounds (7.4), (7.7), (7.8), (7.9), (7.10), (7.13), (7.15), (7.17) and (7.19).

The bound for $\mathrm{MSE}(\hat{\beta}_{HR})$ is obtained by combining the above bound for $|\mathrm{MSE}(\hat{\beta}_{HR})-\mathrm{MSE}(\widetilde{\beta}_{HR})|$ with the bound obtained for $\mathrm{MSE}(\widetilde{\beta}_{HR})$ in (7.4).

The proofs of the results for $\hat{\beta}_{TR}$ are directly analogous to those for $\hat{\beta}_{HR}$ and are therefore omitted. □

Proof of Theorem 3. It was obtained in the proof of part(b) of Theorem 1 that $\mathrm{MSE}(\widetilde{\beta}_{TR})-\mathrm{MSE}(\widetilde{\beta}_{HR})=O(1)\rho^2$ as $n\to\infty$, where the $O(1)$ term is bounded below by a positive number for all sufficiently large n. Let κ_1 be a positive number which is less than this $O(1)$ term for all sufficiently large n. Then,

$$\mathrm{MSE}(\widetilde{\beta}_{TR})-\mathrm{MSE}(\widetilde{\beta}_{HR})>\kappa_1\rho^2 \tag{7.20}$$

as $n\to\infty$. Note that this bound is irrespective of whether $\alpha>\eta-1/2$ or $\alpha<\eta-1/2$.

Now, if $\rho\sim cn^{-\varepsilon}$ for any $\varepsilon<\alpha/(5\alpha-2\eta+2)$ when $\alpha>\eta-1/2$ or for any $\varepsilon<\alpha/(3\alpha+1)$ when $\alpha<\eta-1/2$, it can be checked from (4.2) that $|\mathrm{MSE}(\hat{\beta}_{HR})-\mathrm{MSE}(\widetilde{\beta}_{HR})|=o(\rho^2)$ as $n\to\infty$. So, by

Theorem 2, it follows that $|\text{MSE}(\hat{\beta}_{TR}) - \text{MSE}(\widetilde{\beta}_{TR})| = o(\rho^2)$ as $n \to \infty$. Fix κ_0 to be any positive number less than κ_1. Thus, using the inequality

$$\begin{aligned}\text{MSE}(\hat{\beta}_{TR}) - \text{MSE}(\hat{\beta}_{HR}) > \{\text{MSE}(\widetilde{\beta}_{TR}) - \text{MSE}(\widetilde{\beta}_{HR})\} - |\text{MSE}(\hat{\beta}_{TR}) - \text{MSE}(\widetilde{\beta}_{TR})| \\ -|\text{MSE}(\hat{\beta}_{HR}) - \text{MSE}(\widetilde{\beta}_{HR})|\end{aligned}$$

along with (7.20) and the rates of convergences of $|\text{M}|\text{SE}(\hat{\beta}_{TR}) - \text{M}|\text{SE}(\widetilde{\beta}_{TR})|$ and $|\text{MSE}(\hat{\beta}_{HR}) - \text{MSE}(\widetilde{\beta}_{HR})|$ obtained above, it follows that

$$\text{MSE}(\hat{\beta}_{TR}) - \text{MSE}(\hat{\beta}_{HR}) > \kappa_0 n^{-2\varepsilon}$$

for all sufficiently large n and for the above choices of ε. □

Proof of Theorem 4. Note that $m^{-1}MSE(\hat{\beta}_{HR}^{(m)})$ is equal to the MSE of $\hat{\beta}_{HR}^{(m)}$ as an estimator of $\Phi_m(\beta)$ when we compute it based on

$$\Phi_m(X_1), \Phi_m(X_2), \ldots, \Phi_m(X_n),$$

where $\Phi_m(X) = X^{(m)}/\sqrt{m}$ and $\Phi_m(\beta) = \beta^{(m)}/\sqrt{m}$. Since Theorem 2 applies to any separable Hilbert space, we will follow the proof of this theorem for the above-mentioned random variables and parameter.

First observe that when deriving bounds for $E(\|U_l\|^2)$ in the proof of Theorem 2, we required bounds for $\sum_{j\geq 1} \lambda_j^a/(\lambda_j+\rho)^b$ for $b \geq a > 0$. So, in the discrete case, we need bounds for $\sum_{j=1}^m (\lambda_j^{(m)})^a/(\lambda_j^{(m)}+\rho)^b$ for $b \geq a > 0$. But by assumption (A2') and using the arguments in the proof of Lemma 1, it follows that

$$\sum_{j=1}^{m} \frac{(\lambda_j^{(m)})^a}{(\lambda_j^{(m)}+\rho)^b} \leq O(1)\rho^{-b+a-\frac{1}{\alpha}}$$

as $m \to \infty$, where the $O(1)$ term is uniform over m.

Next, we bound $\sum_{j=1}^m \langle \beta^{(m)}, \phi_j^{(m)}\rangle^2/(\lambda_j^{(m)}+\rho)^2$, which is the discrete version of (7.3) used in the proof of Theorem 2. First, consider the case when $\alpha > \eta' - 1/2$. Observe that since $m > \rho^{-2}$ and $\alpha > 1$, in this case, we have $\rho^{-1/\alpha} < m^{1/\eta'}$. So, defining $J = [\rho^{-1/\alpha}]$ as in the proof of Lemma 1 and by using assumption (A3'), we have

$$\sum_{j=1}^{J} \frac{\langle \beta^{(m)}, \phi_j^{(m)}\rangle^2}{(\lambda_j^{(m)}+\rho)^2} \quad \leq \quad O(1)\rho^{\frac{2\eta'}{\alpha}-2-\frac{1}{\alpha}},$$

where the $O(1)$ term is uniform over m. Further,

$$\begin{aligned}\sum_{j=J+1}^{[m^{1/\eta'}]} \frac{\langle \beta^{(m)}, \phi_j^{(m)}\rangle^2}{(\lambda_j^{(m)}+\rho)^2} \quad &\leq \quad O(1)\rho^{-2}\sum_{j>J} j^{-2\eta'} \quad \leq \quad O(1)\rho^{\frac{2\eta'}{\alpha}-2-\frac{1}{\alpha}}, \\ \sum_{j>[m^{1/\eta'}]}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)}\rangle^2}{(\lambda_j^{(m)}+\rho)^2} \quad &\leq \quad O(1)\rho^{-2}\sum_{j>[m^{1/\eta'}]} m^{-2} \quad \leq \quad O(1)\rho^{-2}/m,\end{aligned}$$

where all the $O(1)$ terms above are uniform in m. Combining all the above inequalities, and using the facts that $m > \rho^{-2}$ and $(2\eta'-1)/\alpha < 2$, we have

$$\sum_{j=1}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)}\rangle^2}{(\lambda_j^{(m)}+\rho)^2} \quad \leq \quad O(1)\rho^{\frac{2\eta'-1}{\alpha}-2},$$

where the $O(1)$ term is uniform over m. We then consider the case $\alpha < \eta' - 1/2$. In this case, we may either have $m > \rho^{-\eta'/\alpha}$ or $m \leq \rho^{-\eta'/\alpha}$. In the first scenario, as in the proof of Lemma 1, we have

$$\begin{aligned}
\sup_{\rho>0} \sum_{j=1}^{[m^{1/\eta'}]} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} &\leq O(1) \sum_{j=1}^{[m^{1/\eta'}]} j^{-2\eta'+2\alpha} \leq O(1), \\
\sum_{j>[m^{1/\eta'}]}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} &\leq O(1)\rho^{-2} \sum_{j>[m^{1/\eta'}]} m^{-2} \leq O(1)\rho^{-2}/m,
\end{aligned}$$

where all the $O(1)$ terms are uniform in m. Since $m > \rho^{-2}$, we have

$$\sum_{j=1}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} \leq O(1),$$

with the $O(1)$ term being uniform in m. In the other scenario, when $m \leq \rho^{-\eta'/\alpha}$, we have

$$\begin{aligned}
\sup_{\rho>0} \sum_{j=1}^{[m^{1/\eta'}]} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} &\leq O(1) \sum_{j=1}^{[m^{1/\eta'}]} j^{-2\eta'+2\alpha} \leq O(1), \\
\sum_{j>[m^{1/\eta'}]}^{[\rho^{-1/\alpha}]} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} &\leq O(1) \sum_{j\leq[\rho^{-1/\alpha}]} m^{-2} j^{2\alpha} \leq O(1)\rho^{-2-1/\alpha}/m^2, \\
\sum_{j>[\rho^{-1/\alpha}]}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} &\leq O(1)\rho^{-2} \sum_{j>[\rho^{-1/\alpha}]}^{m} m^{-2} \leq O(1)\rho^{-2}/m,
\end{aligned}$$

where all the $O(1)$ terms are uniform in m. Since $\alpha > 1$ and $m > \rho^{-2}$, we again have

$$\sum_{j=1}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} \leq O(1),$$

with the $O(1)$ term being uniform in m. Thus, analogous to the bound in (7.3) in the proof of Theorem 2, we have

$$\rho^2 \sum_{j=1}^{m} \frac{\langle \beta^{(m)}, \phi_j^{(m)} \rangle^2}{(\lambda_j^{(m)} + \rho)^2} \leq O(\rho^M),$$

where $M = (2\eta' - 1)/\alpha$ or $M = 2$ according as $\alpha > \eta' - 1/2$ or $\alpha < \eta' - 1/2$.

The proof of the present theorem is now complete by using arguments similar to those used in the proof of Theorem 3, using the above bounds and noting that the other $O(1)$ terms in the proof of Theorem 2, namely, those involved with the $E(\|U_l\|^2)$'s and that of $m^{-1}E\{\|\widetilde{\beta}_{HR}^{(m)} - \beta^{(m)}\|^2\}$ are uniform over m. □

Supplementary Material

A companion supplement contains the proof of Theorem 1, the verification of Assumption (A3') for the case of standard Brownian motion, the results of a simulation study for sample sizes $n = 50$ and $n = 300$ as well as the case when the functional covariate is observed with error, and comparative results when all three regularisation methods are applied to real data set.

References

Alquier, P., Gautier, E. and Stoltz, G., eds. (2011). *Inverse problems and high-dimensional estimation. Lecture Notes in Statistics—Proceedings* **203**. Springer, Heidelberg Lecture notes from the "Stats in the Château" Summer School held in Jouy-en-Josas, August 31–September 4, 2009. MR2867623

Amini, A. A. and Wainwright, M. J. (2012). Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *Ann. Statist.* **40** 2483–2510. MR3097610

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305 (97i:62062)

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *Ann. Statist.* **39** 2330–2355. MR2906870

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162 (2004e:62072)

Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *J. Multivariate Anal.* **101** 395–408. MR2564349

Cardot, H., Mas, A. and Sarda, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields* **138** 325–361. MR2299711 (2007m:60055)

Cardot, H. and Sarda, P. (2006). Linear Regression Models for Functional Data. In *The Art of Semiparametrics. Contributions to Statistics* 49-66. Physica-Verlag HD.

Chakraborty, A. and Panaretos, V. M. (2016). Supplement to "Hybrid regularisation of functional linear models".

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697 (2011c:62187)

Comte, F. and Johannes, J. (2012). Adaptive functional linear regression. *Ann. Statist.* **40** 2765–2797. MR3097959

Conway, J. B. (1978). *Functions of one complex variable*, second ed. *Graduate Texts in Mathematics* **11**. Springer-Verlag, New York-Berlin. MR503901 (80c:30003)

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344 (2010i:62089)

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression: The case of fixed design and functional response. *Canadian Journal of Statistics* **30** 285–300.

Ferraty, F. and Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* **330** 139 - 142.

Grenander, U. (1981). *Abstract inference*. Wiley New York.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269 (2008k:62134)

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 109–126. MR2212577

Hocking, R. R. (2003). *Methods and applications of linear models*, second ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR1963885 (2004b:62002)

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications. Springer Series in Statistics*. Springer, New York. MR2920735

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction*

to linear operators. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. MR3379106

Jolliffe, I. T. (2002). *Principal component analysis*, second ed. *Springer Series in Statistics*. Springer-Verlag, New York. MR2036084 (2004k:62010)

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Anal.* **98** 1782–1804. MR2392433 (2009e:62174)

Lukas, M. A. (1993). Asymptotic optimality of generalized cross-validation for choosing the regularization parameter. *Numer. Math.* **66** 41–66. MR1240702

Lukas, M. A. (2006). Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems* **22** 1883–1902. MR2261272

Marx, B. D. and Eilers, P. H. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics* **41** 1–13.

Meister, A. (2009). *Deconvolution problems in nonparametric statistics. Lecture Notes in Statistics* **193**. Springer-Verlag, Berlin. MR2768576

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* 539–572.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*, second ed. *Springer Series in Statistics*. Springer, New York. MR2168993

Shao, P. Y.-S. and Strawderman, W. E. (1994). Improving on the James-Stein positive-part estimator. *Ann. Statist.* **22** 1517–1538. MR1311987

Theobald, C. M. (1974). Generalizations of Mean Square Error Applied to Ridge Regression. *Journal of the Royal Statistical Society. Series B (Methodological)* **36** 103-106.

Tikhonov, A. and Arsenin, V. *Solutions of ill-posed problems*.

Utreras, F. I. (1987). On generalized cross-validation for multivariate smoothing spline functions. *SIAM J. Sci. Statist. Comput.* **8** 630–643. MR892310

Wahba, G. (1990). *Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR1045442

Yao, F., Muller, H.-G. and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857 (2012b:62237)