The Annals of Statistics
2012, Vol. 40, No. 5, 2421–2451
DOI: 10.1214/12-AOS1027

ESTIMATION IN FUNCTIONAL REGRESSION FOR GENERAL EXPONENTIAL FAMILIES

BY WINSTON WEI DOU[2], DAVID POLLARD[1] AND HARRISON H. ZHOU[2]

Yale University

This paper studies a class of exponential family models whose canonical parameters are specified as linear functionals of an unknown infinite-dimensional slope function. The optimal minimax rates of convergence for slope function estimation are established. The estimators that achieve the optimal rates are constructed by constrained maximum likelihood estimation with parameters whose dimension grows with sample size. A change-of-measure argument, inspired by Le Cam's theory of asymptotic equivalence, is used to eliminate the bias caused by the nonlinearity of exponential family models.

1. Introduction. There has been extensive exploratory and theoretical study of *functional data analysis* (FDA) over the past two decades. Two monographs by Ramsay and Silverman (2002, 2005) provide comprehensive discussions on methods and applications.

Among many problems involving functional data, slope estimation in functional linear regression has received substantial attention in literature, for example, by Cardot, Ferraty and Sarda (2003), Li and Hsing (2007) and Hall and Horowitz (2007). In particular, Hall and Horowitz (2007) established minimax rates of convergence and proposed rate-optimal estimators based on spectral truncation (regression on functional principal components). They showed that the optimal rates depend on the smoothness of the slope function and the decay rate of the eigenvalues of the covariance kernel of the functional independent variable.

In this paper, we study optimal rates of convergence for slope estimation in functional generalized linear models, for which little theory is known. We introduce several new technical devices to overcome the problems caused by nonlinearity of the link function. To analyze our estimator, we establish a sharp approximation for maximum likelihood estimators for exponential families parametrized by linear functions of N-dimensional parameters, for an N that grows with sample size; see Lemma 1. We develop a change-of-measure argument—inspired by ideas from Le

Received August 2011; revised June 2012.
[1]Supported in part by NSF Grant MSPA-MCS-0528412.
[2]Supported in part by NSF Career Award DMS-06-45676 and NSF FRG Grant DMS-08-54975.
MSC2010 subject classifications. Primary 62J05, 60K35; secondary 62G20.
Key words and phrases. Approximation of compact operators, Assouad's lemma, exponential families, functional estimation, minimax rates of convergence.

Cam's theory of asymptotic equivalence of models—to eliminate the effect of bias terms caused by the nonlinearity of the link function; see Sections 4.2 and 4.3.

We consider problems where the observed data consist of independent, identically distributed pairs $(y_i, \mathbb{X}_i)$ where each $\mathbb{X}_i$ is a Gaussian process indexed by a compact subinterval of the real line, which with no loss of generality we take to be $[0, 1]$. Assume, for each i, that the random variable y_i conditional on the process $\mathbb{X}_i$, follows a distribution Q_{λ_i}, where $\{Q_\lambda : \lambda \in \mathbb{R}\}$ is a one-parameter exponential family. The density function of Q_λ is specified in equation (2). We take parameter λ_i to be a linear functional of $\mathbb{X}_i$ of the form

$$\lambda_i = a + \int_0^1 \mathbb{X}_i(t)\mathbb{B}(t)\,dt \tag{1}$$

for an unknown constant a and an unknown $\mathbb{B} \in L^2[0, 1]$.

Thus, the conditional joint distribution of $(y_1, \ldots, y_n)$ given $(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ is the product measure $\mathbb{Q}_{n,a,\mathbb{B},\mathbb{X}_1,\ldots,\mathbb{X}_n} = \bigotimes_{i\le n} Q_{\lambda_i}$. We abbreviate $\mathbb{Q}_{n,a,\mathbb{B},\mathbb{X}_1,\ldots,\mathbb{X}_n}$ to $\mathbb{Q}_{n,a,\mathbb{B}}$. Write $P_{\mu,K}$ for the distribution of each $\mathbb{X}_i$, where μ is the mean and K is the covariance of $\mathbb{X}_i$. The joint distribution of the sample processes is then $\mathbb{P}_{n,\mu,K} = P_{\mu,K}^n$. Therefore, our models $\mathbb{P}_{n,f} := \mathbb{P}_{n,\mu,K}\mathbb{Q}_{n,a,\mathbb{B}}$, where $f = (K, a, \mu, \mathbb{B})$, are the joint distributions of the sample $(y_1, \mathbb{X}_1), \ldots, (y_n, \mathbb{X}_n)$. The parameter set $\mathcal{F} \equiv \mathcal{F}(R, \alpha, \beta)$ depend on universal constants R, α and β. See Definition 1 (in Section 2) for the precise specification of the parameter set. The universal constant α controls the decay rate of eigenvalues of kernel K, and the universal constant β characterizes the "smoothness" of the slope function $\mathbb{B}$.

Denote the corresponding norm and inner product in the space $L^2[0, 1]$ by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$. We focus on the estimation of $\mathbb{B}$ using integrated squared error loss,

$$L(\widehat{\mathbb{B}}_n, \mathbb{B}) = \|\widehat{\mathbb{B}}_n - \mathbb{B}\|^2 = \int_0^1 \big(\widehat{\mathbb{B}}_n(t) - \mathbb{B}(t)\big)^2\,dt.$$

The two main results are as follows.

THEOREM 1 (Minimax upper bound). *Under the assumptions stated in Section 2, there exists an estimating sequence of $\widehat{\mathbb{B}}_n$'s for which: for each $\varepsilon > 0$ there exists a finite constant C_ε such that*

$$\sup_{f\in\mathcal{F}} \mathbb{P}_{n,f}\big\{\|\widehat{\mathbb{B}}_n - \mathbb{B}\|^2 > C_\varepsilon n^{(1-2\beta)/(\alpha+2\beta)}\big\} < \varepsilon \qquad \textit{for all large enough } n.$$

THEOREM 2 (Minimax lower bound). *Under the assumptions stated in Section 2,*

$$\liminf_{n\to\infty} n^{(2\beta-1)/(\alpha+2\beta)} \sup_{f\in\mathcal{F}} \mathbb{P}_{n,f}\|\widehat{\mathbb{B}}_n - \mathbb{B}\|^2 > 0 \qquad \textit{for every estimator } \{\widehat{\mathbb{B}}_n\}.$$

Two closely related works in the area of functional generalized linear models are Müller and Stadtmüller (2005) and Cardot and Sarda (2005), which provided theory for the convergence rate in functional generalized linear models. However, the rate optimality was not studied. In addition, Müller and Stadtmüller (2005) established an upper bound for rates of convergence assuming the negligibility of the bias due to the approximation of the infinite-dimensional model by a sequence of finite-dimensional models, the issue we overcome by using a change-of-measure argument. By contrast, more theoretical advances have been achieved in the functional linear regression setting, not only for estimation but also for prediction. For example, Cai and Hall (2006) and Crambes, Kneip and Sarda (2009) derived optimal rates of convergence for prediction in the fixed and random design cases. See also, Cardot, Mas and Sarda (2007) which derived a CLT for prediction in the fixed and random design cases and Cardot and Johannes (2010) which established a minimax optimal result for prediction at a random design using thresholding estimators. In a companion study to our paper, Dou [(2010), Chapter 5] considers optimal prediction in functional generalized linear regressions with an application to the economic problem of predicting recessions from the U.S. Treasury yield curve.

Our minimax upper bound result (Theorem 1) is proved in Section 4. The minimax lower bound result (Theorem 2) is established in Section 5. The proof of Theorem 1 depends on an approximation result (Lemma 1) for maximum likelihood estimators in exponential family models for parameters whose dimensions change with sample size. As an aid to the reader, we present our proof of Theorem 1 in two stages. In Section 4.2, we assume that both the mean μ and the covariance kernel K are known. This allows us to emphasize the key ideas in our proofs. We proceed in Section 4.3 to the case where μ and K are estimated. The proofs for the lemmas are collected together in Section 6. Some of them invoke the perturbation-theoretic results collected in the supplemental article [Dou, Pollard and Zhou (2012)].

2. Regularity conditions. Let $\{Q_\lambda : \lambda \in \mathbb{R}\}$ be a one-parameter exponential family,

$$dQ_\lambda / dQ_0 = f_\lambda(y) := \exp\bigl(\lambda y - \psi(\lambda)\bigr) \qquad \text{for all } \lambda \in \mathbb{R}. \tag{2}$$

Necessarily $\psi(0) = 0$. Remember that $e^{\psi(\lambda)} = Q_0 e^{\lambda y}$ and that the distribution Q_λ has mean $\dot\psi(\lambda)$ and variance $\ddot\psi(\lambda)$.

REMARK. We may assume that $\ddot\psi(\lambda) > 0$ for every real λ. Otherwise we would have $0 = \ddot\psi(\lambda_0) = \operatorname{var}_{\lambda_0}(y) = Q_0 f_{\lambda_0}(y)(y - \dot\psi(\lambda_0))^2$ for some λ_0, which would make $y = \dot\psi(\lambda_0)$ for almost all y under Q_0 and hence $Q_\lambda \equiv Q_{\lambda_0}$ for every λ.

REMARK. The main results in this paper can be extended to the functional quasi-likelihood regression models [see, e.g., Wedderburn (1974)] as follows:

$$y_i = \mu_i + \sigma_i \varepsilon_i,$$

where

$$\mu_i = g\left(a + \int_{\mathcal{I}} \mathbb{B}(t)\mathbb{X}_i(t)\,dt\right) \quad \text{and} \quad \sigma_i = v(\mu_i) \qquad \text{with known } g \text{ and } v.$$

However, a main goal of this paper is to provide a better understanding of the difficulties caused by nonlinearity in functional data analysis models and to propose a general approach to tackle them. The exponential families can provide a good representation of the quasi-likelihood regression models to this end. One of the gains of specifying exponential families is to simplify the proofs while still achieving our main goal and covering the most broadly used models, such as the functional logistic regression model, the functional probit regression model and the functional poisson regression model. The general nonparametric setting where the link functions g and v are unknown is studied by Müller and Stadtmüller (2005), assuming the negligibility of the bias due to the approximation of the infinite-dimensional model by a sequence of finite-dimensional models. Without ignoring the bias, the problem becomes much more difficult and would be an interesting topic for future research.

REMARK. A natural extension of our model is the classical generalized linear model with nuisance parameters ϕ as follows:

$$y_i|\mathbb{X}_i \sim f_{\lambda_i,\phi}(y) \qquad \text{with } \lambda_i = a + \int_{\mathcal{I}} \mathbb{B}(t)\mathbb{X}_i(t)\,dt$$

and

$$f_{\lambda,\phi}(y) := \exp\bigl[\alpha_1(\phi)\bigl(\lambda y - \psi(\lambda)\bigr) + \alpha_2(\phi, y)\bigr],$$

where $\alpha_1(\phi) > 0$ so that for each $\phi \in \mathbb{R}^d$ we have an exponential family. Under some regularity conditions on the known functions $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$, the exact maximum likelihood estimation analysis and the lower bound argument of this paper can still be employed to derive minimax results for the slightly more general setting.

We assume:

($\ddot{\Psi}$) For each $\varepsilon > 0$ there exists a finite constant C_ε for which $\ddot{\psi}(\lambda) \le C_\varepsilon \exp(\varepsilon\lambda^2)$ for all $\lambda \in \mathbb{R}$. Equivalently, $\ddot{\psi}(\lambda) \le \exp(o(\lambda^2))$ as $|\lambda| \to \infty$.

($\dddot{\Psi}$) There exists an increasing real function G on $\mathbb{R}^+$ such that

$$\bigl|\dddot{\psi}(\lambda + h)\bigr| \le \ddot{\psi}(\lambda) G\bigl(|h|\bigr) \qquad \text{for all } \lambda \text{ and } h.$$

Without loss of generality we assume $G(0) \ge 1$.

We also assume the observed data are i.i.d. pairs $(y_i, \mathbb{X}_i)$ for $i = 1, \ldots, n$, where:

(X) Each $\{\mathbb{X}_i(t) : 0 \le t \le 1\}$ is distributed like $\{\mathbb{X}(t) : 0 \le t \le 1\}$, a Gaussian process with mean $\mu(t)$ and covariance kernel $K(s, t)$.

(Y) $y_i|\mathbb{X}_i \sim Q_{\lambda_i}$ with $\lambda_i = a + \langle \mathbb{X}_i, \mathbb{B} \rangle$ for an unknown $\{\mathbb{B}(t) : 0 \le t \le 1\}$ in $L^2[0, 1]$ and $a \in \mathbb{R}$.

DEFINITION 1. *For real constants $\alpha > 1$ and $\beta > (\alpha + 3)/2$ and $R > 1$, define $\mathcal{F} = \mathcal{F}(R, \alpha, \beta)$ as the set of all $f = (K, a, \mu, \mathbb{B})$ that satisfy the following conditions*:

(K) *The covariance kernel is square integrable with respect to Lebesgue measure and has an eigenfunction expansion* (*as a compact operator on* $L^2[0, 1]$)

$$K(s, t) = \sum_{k \in \mathbb{N}} \theta_k \phi_k(s) \phi_k(t),$$

where the eigenvalues θ_k are decreasing with $Rk^{-\alpha} \ge \theta_k \ge \theta_{k+1} + (\alpha/R)k^{-\alpha-1}$.

(a) $|a| \le R$.

(μ) $\|\mu\| \le R$.

(B) $\mathbb{B}$ *has an expansion $\mathbb{B}(t) = \sum_{k \in \mathbb{N}} b_k \phi_k(t)$ with $|b_k| \le Rk^{-\beta}$, for the eigenfunctions defined by the kernel K.*

REMARK. The purpose of this paper is not to offer a universally optimal estimation procedure, but to provide a theory for the principal components regression in nonlinear models of functional data. As in Hall and Horowitz (2007) and Cai and Hall (2006), among others, assumptions (K) and (B) set up a natural theoretical framework to justify and analyze the principal components regression. In practice, principal components analysis has been one of the most widely and successfully used statistical methods. One example of successful application of principal components analysis is in analyzing the relationship between U.S. Treasury zero-coupon yield curves, which is a typical functional data, and the macroeconomic activities [see, e.g., Dou (2010), Estrella and Hardouvelis (1991), Wright (2006)]. In this analysis, the fixed basis such as wavelet basis or fourier basis fails to give a sparse representation of the yield curve data. Admittedly, under different regularity assumptions, by design the principal components regression approach may not be applicable, and accordingly, other estimation methods such as wavelet basis or spline basis may have better performance; see, for example, Crambes, Kneip and Sarda (2009), Efromovich and Koltchinskii (2001). In Efromovich and Koltchinskii (2001), the authors discussed an approach of using two different bases, one is for the slope function and the other is for the covariance kernel operator. This technique can be applied to some cases where the principal components regression fails. Nevertheless, the results in Efromovich and Koltchinskii [(2001), Theorem 3.1] requires a lower level of noise in the covariance kernel and a higher

degree of smoothness of the slope function in order to allow tractability in more severely ill-posed settings.

REMARK. The awkward lower bound for θ_k in assumption (K) implies, for all $k < j$,

$$\theta_k - \theta_j \geq R^{-1} \int_k^j \alpha x^{-\alpha-1}\, dx = R^{-1}\bigl(k^{-\alpha} - j^{-\alpha}\bigr). \tag{3}$$

If K and μ were known, we would only need the lower bound $\theta_k \geq R^{-1}k^{-\alpha}$ and not the lower bound for $\theta_k - \theta_{k+1}$. As explained by Hall and Horowitz [(2007), page 76], the stronger assumption is needed when one estimates the individual eigenfunctions of K. Note that the subset of $L^2[0,1]$ in which $\mathbb{B}$ lies, denoted as $\mathcal{B}_K$, depends on K. We regard the need for the stronger assumption on the eigenvalues and the irksome assumption (B) as artifacts of the method of proof, but we have not yet succeeded in removing either assumption.

REMARK. We discuss two extreme cases to help understand the regularity assumption $\beta > (\alpha+3)/2$. One case is that the eigenvalues $\{\theta_k\}$ decay exponentially fast and the slope coefficients $\{b_k\}$ decay with polynomial rates, where essentially we have α is much larger than β, for which it can be shown that the optimal rate of convergence is just logarithmic. The other case is that the eigenvalues $\{\theta_k\}$ decay polynomially fast, and the slope coefficients $\{b_k\}$ decay with exponential rates, where essentially we have β is much larger than α, for which it can be shown that the optimal convergence rate is nearly parametric up to a logarithmic term.

3. Methodology. In this section we introduce the methodology to construct a sequence of estimators, which achieve the optimal rates of convergence stated in Theorem 1. Our estimation features a two-step procedure. We first truncate at the first N principal components and replace the original model $\mathbb{P}_{n,f}$ by the truncated model $\widetilde{\mathbb{P}}_{n,f,N}$ defined in (7). The choice of N depends on an estimation-approximation trade-off: oversized N can compromise the performance of the MLE maximizing (11), whereas undersized N can make the model misspecification between $\mathbb{P}_{n,f}$ and its finite-dimensional approximation $\widetilde{\mathbb{P}}_{n,f,N}$ nonnegligible. Second, we further truncate the MLE at $m < N$ to form our estimator in (10). The choice of m depends on the standard variance-bias tradeoff as in nonparametric estimation problems. See Section 4.2 for more details.

Under the assumptions (X) and (K) from Section 2, the process $\mathbb{X}_i$ admits the eigen decomposition

$$\mathbb{X}_i(t) - \mu(t) = \mathbb{Z}_i(t) = \sum_{k\in\mathbb{N}} z_{i,k}\phi_k(t).$$

The random variables $z_{i,k} := \langle \mathbb{Z}_i, \phi_k \rangle$ are independent with $z_{i,k} \sim N(0, \theta_k)$.

Because μ and K are unknown, we estimate them in the usual way:

$$\widetilde{\mu}(t) = \overline{\mathbb{X}}(t) = n^{-1} \sum_{i \le n} \mathbb{X}_i(t) \tag{4}$$

and

$$\begin{aligned} \widetilde{K}(s,t) &= (n-1)^{-1} \sum_{i \le n} \big(\mathbb{X}_i(s) - \overline{\mathbb{X}}(s)\big)\big(\mathbb{X}_i(t) - \overline{\mathbb{X}}(t)\big) \\ &= (n-1)^{-1} \sum_{i \le n} \big(\mathbb{Z}_i(s) - \overline{\mathbb{Z}}(s)\big)\big(\mathbb{Z}_i(t) - \overline{\mathbb{Z}}(t)\big), \end{aligned} \tag{5}$$

which has spectral representation

$$\widetilde{K}(s,t) = \sum_{k \in \mathbb{N}} \widetilde{\theta}_k \widetilde{\phi}_k(s) \widetilde{\phi}_k(t) \tag{6}$$

with $\widetilde{\theta}_1 \ge \widetilde{\theta}_2 \ge \cdots \ge \widetilde{\theta}_{n-1} \ge 0$. In fact we must have $\widetilde{\theta}_k = 0$ for $k \ge n$ because all the eigenfunctions $\widetilde{\phi}_k$ corresponding to nonzero $\widetilde{\theta}_k$'s must lie in the $(n-1)$-dimensional space spanned by $\{\mathbb{Z}_i - \overline{\mathbb{Z}} : i = 1, 2, \ldots, n\}$.

Using the first N [to be specified in (13)] principal components, we can approximate the original infinite-dimensional model $\mathbb{P}_{n,f}$ by a sequence of truncated finite-dimensional models

$$\widetilde{\mathbb{P}}_{n,f,N} = P_{n,\mu,K} \widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N,\mathbb{X}_1,\ldots,\mathbb{X}_n}, \tag{7}$$

where $\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N,\mathbb{X}_1,\ldots,\mathbb{X}_n} := \bigotimes_{i \le n} Q_{\widetilde{\lambda}_{i,N}}$ with $y_i | \mathbb{X}_1, \ldots, \mathbb{X}_n \sim Q_{\widetilde{\lambda}_{i,N}}$ and

$$\widetilde{\lambda}_{i,N} = \widetilde{b}_0 + \sum_{1 \le k \le N} \widetilde{b}_k (\widetilde{z}_{i,k} - \widetilde{z}_{\cdot k}), \tag{8}$$

where $\widetilde{b}_0 = a + \langle \mathbb{B}, \overline{\mathbb{X}} \rangle$, and $\widetilde{b}_k = \langle \mathbb{B}, \widetilde{\phi}_k \rangle$ for $k \ge 1$, and $\widetilde{z}_{i,k} = \langle \mathbb{Z}_i, \widetilde{\phi}_k \rangle$ for all i, k, and $\widetilde{z}_{\cdot k} = n^{-1} \sum_{i \le n} \widetilde{z}_{i,k} = \langle \overline{\mathbb{Z}}, \widetilde{\phi}_k \rangle$. And hence $\widetilde{z}_{i,k} - \widetilde{z}_{\cdot k} = \langle \mathbb{Z}_i - \overline{\mathbb{Z}}, \widetilde{\phi}_k \rangle = \langle \mathbb{X}_i - \overline{\mathbb{X}}, \widetilde{\phi}_k \rangle$. We abbreviate $\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N,\mathbb{X}_1,\ldots,\mathbb{X}_n}$ to $\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N}$ in the rest of the paper. We introduce the following matrices and vectors for the purpose of notational convenience. Define:

- $z_i := (z_{i,1}, \ldots, z_{i,N})'$ and $\widetilde{z}_i := (\widetilde{z}_{i,1}, \ldots, \widetilde{z}_{i,N})'$;
- $z_\cdot := (z_{\cdot 1}, \ldots, z_{\cdot N})'$ and $\widetilde{z}_\cdot := (\widetilde{z}_{\cdot 1}, \ldots, \widetilde{z}_{\cdot N})'$;
- $D := \operatorname{diag}(1, \theta_1, \ldots, \theta_N)^{1/2}$, where θ_k's are defined in assumption (K);
- $\widetilde{D} := \operatorname{diag}(1, \widetilde{\theta}_1, \ldots, \widetilde{\theta}_N)^{1/2}$, where $\widetilde{\theta}_k$'s are defined in (6);
- $\xi_i := (1, z_i')'$ and $\widetilde{\xi}_i := (1, \widetilde{z}_i' - \widetilde{z}_\cdot')'$;
- $\eta_i := D^{-1} \xi_i$ and $\widetilde{\eta}_i := D^{-1} \widetilde{\xi}_i$;
- $\gamma := (b_0, b_1, \ldots, b_N)'$ and $\widetilde{\gamma} := (\widetilde{b}_0, \widetilde{b}_1, \ldots, \widetilde{b}_N)'$.

Thus, equation (8) can be rewritten as

$$\widetilde{\lambda}_{i,N} = \widetilde{\xi}_i' \widetilde{\gamma} = \widetilde{\eta}_i' D \widetilde{\gamma}. \tag{9}$$

We estimate $\mathbb{B}$ by

$$\widehat{\mathbb{B}}_n(t) = \sum_{k \le m} \widehat{b}_k \widetilde{\phi}_k(t), \tag{10}$$

where $(\widehat{b}_0, \ldots, \widehat{b}_N)$ is the conditional MLE for the truncated model $\widetilde{\mathbb{P}}_{n,f,N}$, and m is the optimal cutoff point according to the variance-bias tradeoff with $m < N$. More precisely, $(\widehat{b}_0, \ldots, \widehat{b}_N)$ is chosen to maximize the following conditional (on the $\mathbb{X}_i$'s) log likelihood over $g \equiv (g_0, g_1, \ldots, g_N)'$ in $\mathbb{R}^{N+1}$:

$$\mathcal{L}_n(g) = \sum_{i \le n} y_i \bigl(\widetilde{\xi}_i' g\bigr) - \psi\bigl(\widetilde{\xi}_i' g\bigr) \tag{11}$$

with cutoff points m and N chosen as

$$m \asymp n^{1/(\alpha+2\beta)} \tag{12}$$

and

$$N \asymp n^{\zeta} \qquad \text{with } (2+2\alpha)^{-1} > \zeta > (\alpha+2\beta-1)^{-1}. \tag{13}$$

Note that N is much larger than m. Such a ζ exists because the assumptions $\alpha > 1$ and $\beta > (\alpha+3)/2$ imply $\alpha + 2\beta - 1 > 2 + 2\alpha$. The universal constants α and β characterize the decay rate of the eigenvalues of kernel K and the smoothness of slope function $\mathbb{B}$ defined in Definition 1.

4. Proof of Theorem 1. The proof of Theorem 1 is divided into two stages. In the first stage, we prove the theorem assuming that the mean μ and the covariance kernel K are known. This case is relatively simple and of course artificial, but it captures the essence of our proof. For the Gaussian case, this is reduced to the setting considered in Goldenshluger and Tsybakov (2001). In the second stage where μ and K are unknown, we show that using the natural estimates $\widetilde{\mu}$ and $\widetilde{K}$ as in (4) and (5) will not affect the result achieved in the first stage.

In Section 4.1 we state the technical lemmas which serve as building blocks for establishing the main theorems. Their proofs are postponed to the Section 6. In Section 4.2 we prove Theorem 1 assuming μ and K are known, and then in Section 4.3 we apply Lemma 5 to complete the proof of Theorem 1 with unknown μ and K.

4.1. *Technical lemmas.* We write the lemmas in a notation that makes the applications in Sections 4.2 and 4.3 more straightforward. The notational cost is that the parameters are indexed by $\{0, 1, \ldots, N\}$ in Lemmas 1 and 2. Each of the lemmas stated in this subsection is a general result.

We first introduce an approximation result for maximum likelihood estimators in exponential family models for parameters whose dimensions change with sample size. This lemma combines ideas from Portnoy (1988) and from Hjort and

Pollard (1993). For each square matrix A, its spectral norm is defined by its largest absolute value of the eigenvalues, that is, $\|A\|_2 := \sup_{|v|\le 1} |Av|$ where $|v|$ denotes the l^2 norm of vector v. The proof can be found in Section 6.1.

LEMMA 1. *Let Q_λ be the one-parameter exponential family distribution defined as in (2) and satisfying regularity condition $(\ddot{\Psi})$. Suppose $\xi_1, \ldots, \xi_n$ are (nonrandom) vectors in $\mathbb{R}^{N+1}$. Suppose $\mathbb{Q} = \bigotimes_{i\le n} Q_{\lambda_i}$ with $\lambda_i = \xi_i'\gamma$ for a fixed $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_N)'$ in $\mathbb{R}^{N+1}$. Under $\mathbb{Q}$, the coordinate maps $y_1, \ldots, y_n$ are independent random variables with $y_i \sim Q_{\lambda_i}$.*

The log-likelihood for fitting the model is

$$L_n(g) = \sum_{i\le n} (\xi_i' g) y_i - \psi(\xi_i' g) \qquad \text{for } g \in \mathbb{R}^{N+1},$$

which is maximized (over $\mathbb{R}^{N+1}$) at the MLE $\widehat{g}$ $(=\widehat{g}_n)$. Define $\eta_i := D^{-1}\xi_i$ for some nonsingular matrix D, and define the matrix

$$J_n := \sum_{i\le n} \xi_i \xi_i' \ddot{\psi}(\lambda_i) = n D A_n D' \qquad \text{with } A_n := \frac{1}{n}\sum_{i\le n} \eta_i \eta_i' \ddot{\psi}(\lambda_i).$$

Assume B_n is another nonsingular matrix for which

$$\|A_n - B_n\|_2 \le \bigl(2\|B_n^{-1}\|_2\bigr)^{-1} \tag{14}$$

and assume

$$\max_{i\le n} |\eta_i| \le \frac{\varepsilon\sqrt{n}/(N+1)}{G(1)\sqrt{32\|B_n^{-1}\|_2}} \qquad \text{for some } 0 < \varepsilon < 1, \tag{15}$$

where $G(\cdot)$ is defined as in regularity condition $(\dddot{\Psi})$. Then, for each set of vectors $\kappa = \{\kappa_0, \ldots, \kappa_M\}$ in $\mathbb{R}^{N+1}$ there is a set $\mathcal{Y}_{\kappa,\varepsilon}$ with $\mathbb{Q}\mathcal{Y}_{\kappa,\varepsilon}^c < 2\varepsilon$ on which

$$\sum_{0\le j\le M} |\kappa_j'(\widehat{g} - \gamma)|^2 \le \frac{6\|B_n^{-1}\|_2}{n\varepsilon} \sum_{0\le j\le M} |D^{-1}\kappa_j|^2.$$

REMARK. This is a quite general result. In this paper, we are interested in one particular case where κ_j have all elements equal to zero except the jth element that equals one and $D = \operatorname{diag}(\theta_0, \ldots, \theta_M)$. In this case, the result can be rewritten as

$$\sum_{0\le j\le M} (\widehat{g}_j - \gamma_j)^2 \le \frac{6\|B_n^{-1}\|_2}{n\varepsilon} \sum_{0\le j\le M} \theta_j^{-2}.$$

The following approximation result for random matrices will be invoked in order to apply Lemma 1 to show Theorem 1. The proof can be found in Section 6.2.

LEMMA 2. *Suppose* $\{\eta_{i,k} : i, k \geq 1\}$ *are i.i.d. standard normal random variables. Let*

$$A_n := n^{-1} \sum_{i \leq n} \eta_i \eta_i' \ddot{\psi}(\gamma' D \eta_i), \tag{16}$$

where $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_N)'$, $\eta_i = (1, \eta_{i,1}, \ldots, \eta_{i,N})'$ *and* $D = \operatorname{diag}(D_0, \ldots, D_N)$. *Define* $B_n := \mathbb{P} A_n$, *and assume* ψ *satisfies regularity condition* $(\ddot{\Psi})$. *If we have* $\sum_{k \geq 1} D_k^2 \gamma_k^2 < \infty$ *and* $N = o(n^{1/2})$, *then it follows that*

$$\left\| B_n^{-1} \right\|_2 = O(1) \quad \text{and} \quad \mathbb{P} \| A_n - B_n \|_2^2 = o(1).$$

The following lemma establishes a bound on the Hellinger distance between members of an exponential family, which plays a key role in our change-of-measure argument. We write $\mathbf{h}(\cdot, \cdot)$ for the Hellinger distance. If both P and Q are dominated by some measure ν, with densities p and q, then $\mathbf{h}^2(P, Q) := \nu(\sqrt{p} - \sqrt{q})^2$. The proof can be found in Section 6.3.

LEMMA 3. *Suppose* $\{Q_\lambda : \lambda \in \mathbb{R}\}$ *is an exponential family defined as in* (2) *and satisfies regularity condition* $(\dddot{\Psi})$. *Then*

$$\mathbf{h}^2(Q_\lambda, Q_{\lambda+\delta}) \leq \delta^2 \ddot{\psi}(\lambda)\big(1 + |\delta|\big) G\big(|\delta|\big) \qquad \forall \lambda, \delta \in \mathbb{R}.$$

Here $G(\cdot)$ *is defined in the condition* $(\dddot{\Psi})$.

The following lemma provides a maximal inequality for weighted-chi-square variables, which easily leads to maximal inequalities for Gaussian processes and multivariate normal vectors. These inequalities will be repeatedly invoked. The proof can be found in Section 6.4.

LEMMA 4. *Suppose* $\{\eta_{i,k} : i, k \geq 1\}$ *are i.i.d. standard normal random variables. Let*

$$W_i = \sum_{k \in \mathbb{N}} \tau_{i,k} \eta_{i,k}^2 \qquad \text{for } i = 1, \ldots, n.$$

If the $\tau_{i,k}$*'s are nonnegative constants with* $T := \max_{i \leq n} \sum_{k \in \mathbb{N}} \tau_{i,k} < \infty$, *then it follows that*

$$\mathbb{P}\Big\{ \max_{i \leq n} W_i > 4T(\log n + x) \Big\} < 2e^{-x} \qquad \text{for each } x \geq 0.$$

The following lemma is to guarantee that the estimation of $\mathbb{B}$ using $\widetilde{\mu}$ and $\widetilde{K}$ basically has the same accuracy as using μ and K. We need some terminology before formally introducing the lemma, and these notations introduced below apply to the rest of the paper. When we want to indicate that a bound involving constants c, C, $C_1, \ldots$ holds uniformly over all models indexed by a set of parameters $\mathcal{F}$,

we write $c(\mathcal{F})$, $C(\mathcal{F})$, $C_1(\mathcal{F}), \ldots$. By the usual convention for eliminating subscripts, the values of the constants might change from one paragraph to the next: a constant $C_1(\mathcal{F})$ in one place need not be the same as a constant $C_1(\mathcal{F})$ in another place. For sequences of constants c_n that might depend on $f \in \mathcal{F}$, we write $c_n = O_{\mathcal{F}}(1)$ and $o_{\mathcal{F}}(1)$ and so on to show that the asymptotic bounds hold uniformly over $\mathcal{F}$. Denote H_p and $\widetilde{H}_p$ to be orthogonal projection operators associated with $\operatorname{span}\{\phi_1, \ldots, \phi_p\}$ and $\operatorname{span}\{\widetilde{\phi}_1, \ldots, \widetilde{\phi}_p\}$, respectively, where ϕ_k's are the eigenfunctions defined in assumption (K), and $\widetilde{\phi}_k$'s are their sample approximations defined in (6). We also need to define the following key quantities:

- $\widetilde{S} := \operatorname{diag}(\sigma_0, \ldots, \sigma_N)$ with $\sigma_0 = 1$ and $\sigma_k = \operatorname{sign}(\langle \phi_k, \widetilde{\phi}_k \rangle)$ for $k \geq 1$.
- $\Delta := \widetilde{K} - K$, where $\widetilde{K}$ is defined in (5).
- $\widetilde{A}_n := n^{-1} \sum_{i \leq n} \widetilde{\eta}_i \widetilde{\eta}_i' \ddot{\psi}(\widetilde{\lambda}_{i,N})$, where $\widetilde{\eta}_i$ and $\widetilde{\lambda}_{i,N}$ are defined in Section 3.
- $\widetilde{B}_n := \widetilde{S} B_n \widetilde{S}$, where B_n is defined in (19).

The proof of Lemma 5 can be found in Section 6.5.

LEMMA 5. *Assume the regularity conditions in Section 2 hold. Let m and N are chosen according to (12) and (13), respectively. For each $\varepsilon > 0$ there exists a set $\widetilde{\mathcal{X}}_{\varepsilon,n}$, depending on μ and K, with*

$$\sup_{\mathcal{F}} \mathbb{P}_{n,\mu,K} \widetilde{\mathcal{X}}_{\varepsilon,n}^c < \varepsilon \qquad \text{for all large enough } n$$

and on which, for some constant C_ε that does not depend on μ or K:

(i) $\|\Delta\| \leq C_\varepsilon n^{-1/2}$;
(ii) $\max_{i \leq n} \|\mathbb{Z}_i\| \leq C_\varepsilon \sqrt{\log n}$ *and* $\|\overline{\mathbb{Z}}\| \leq C_\varepsilon n^{-1/2}$;
(iii) $\|(\widetilde{H}_m - H_m)\mathbb{B}\|^2 = o_{\mathcal{F}}(n^{(1-2\beta)/(\alpha+2\beta)})$;
(iv) $\|(\widetilde{H}_N - H_N)\mathbb{B}\|^2 = O_{\mathcal{F}}(n^{-1-\nu})$ *for some universal constant* $\nu > 0$;
(v) $\max_{i \leq n} |\widetilde{\eta}_i|^2 = o_{\mathcal{F}}(\sqrt{n}/N)$;
(vi) $\|\widetilde{S}\widetilde{A}_n\widetilde{S} - A_n\|_2 = o_{\mathcal{F}}(1)$.

4.2. *Proof of Theorem 1 with known Gaussian distribution.* Initially we suppose that μ and K are known. We emphasize that this simpler case serves as an intermediate step to the more interesting unknown distribution case, and it captures the essential idea of the proof of Theorem 1.

Remember under $\mathbb{Q}_{n,a,\mathbb{B}}$, the y_i's are independent, conditional on $\mathbb{X}_1, \ldots, \mathbb{X}_n$, with $y_i \sim Q_{\lambda_i}$ and

$$\lambda_i = a + \langle \mathbb{X}_i, \mathbb{B} \rangle = b_0 + \sum_{k \in \mathbb{N}} z_{i,k} b_k \qquad \text{where } b_0 = a + \langle \mu, \mathbb{B} \rangle.$$

Our task is to estimate the b_k's with sufficient accuracy so that we are able to estimate $\mathbb{B}(t) = \sum_{k \in \mathbb{N}} b_k \phi_k(t)$ within an error of order $n^{(1-2\beta)/(\alpha+2\beta)}$. In fact it

will suffice to estimate the component $H_m\mathbb{B}$ of $\mathbb{B}$ in the subspace spanned by $\{\phi_1, \ldots, \phi_m\}$ with $m \asymp n^{1/(\alpha+2\beta)}$ because

$$\|H_m^{\perp}\mathbb{B}\|^2 = \sum_{k>m} b_k^2 = O_{\mathcal{F}}(m^{1-2\beta}) = O_{\mathcal{F}}(n^{(1-2\beta)/(\alpha+2\beta)}). \tag{17}$$

One might try to estimate the coefficients $(b_0, \ldots, b_m)$ by choosing $\widehat{g} = (\widehat{g}_0, \ldots, \widehat{g}_m)'$ to maximize a conditional log likelihood over all $g = (g_0, g_1, \ldots, g_m)'$ in $\mathbb{R}^{m+1}$:

$$\mathcal{L}_{n,m}(g) := \sum_{i\le n} y_i \Big(g_0 + \sum_{1\le k\le m} z_{i,k} g_k \Big) - \psi\Big(g_0 + \sum_{1\le k\le m} z_{i,k} g_k \Big).$$

To this end one might try to appeal to Lemma 1 stated at the beginning of the previous subsection, with κ_j equal to the unit vector with a 1 in its jth position for $j \le m$ and $\kappa_j = 0$ otherwise. That would give a bound for $\sum_{k\le m}(\widehat{g}_k - b_k)^2$. Unfortunately, we cannot directly invoke the lemma with $N = m$ to estimate $\gamma' = (b_0, b_1, \ldots, b_N)$ when we replace $\mathbb{Q}$, D, ξ_i and η_i (notations) in Lemma 1 by $\mathbb{Q}_{n,a,\mathbb{B}}$ (defined in Section 1), D, ξ_i and η_i (defined in Section 3), respectively, because $\lambda_i \ne \xi_i'\gamma$, a bias problem.

REMARK. We could modify Lemma 1 to allow $\lambda_i = \xi_i'\gamma + \text{bias}_i$, for a suitably small bias term, but at the cost of extra regularity conditions and a more delicate argument. The same difficulty arises whenever one investigates the asymptotics of maximum likelihood estimators with the true distribution outside the model family, that is, MLE under model misspecification.

Instead, we use a two-stage estimation procedure,

$$\widehat{\mathbb{B}}_n = \sum_{k\le m} \widehat{b}_k \phi_k, \tag{18}$$

where $(\widehat{b}_0, \ldots, \widehat{b}_N)$ is the conditional MLE for the truncated model and $m \le N$. More precisely, $(\widehat{b}_0, \ldots, \widehat{b}_N)$ is chosen to maximize the following conditional (on the $\mathbb{X}_i$'s) log likelihood over $g \equiv (g_0, g_1, \ldots, g_N)$ in $\mathbb{R}^{N+1}$:

$$\mathcal{L}_{n,N}(g) := \sum_{i\le n} y_i(\xi_i' g) - \psi(\xi_i' g)$$

with cutoff points m and N chosen as in (12) and (13), respectively. Note that this estimator differs from that in (10) in the sense that it uses ϕ_k and $z_{i,k}$ instead of the approximation correspondences $\widetilde{\phi}_k$ and $\widetilde{z}_{i,k} - \widetilde{z}_{\cdot k}$. This two-stage estimation procedure eliminates the bias term by a change-of-measure argument conditional on the $\mathbb{X}_i$'s. We present the proof in the following three steps.

Step 1. From the analysis above, one can see that the key in our proof is the change-of-measure argument and the application of Lemma 1. In this step, we construct a high probability set such that for each realization of the $\mathbb{X}_i$'s on the set the assumptions of Lemma 1 are satisfied.

Define γ, ξ_i, D and η_i as in Section 3. Note that in this case $\eta_{i,j} = z_{i,j}/\sqrt{\theta_i}$ for all i, j, and hence the $\eta_{i,j}$'s are i.i.d. standard normal variables. We define matrix A_n as in (16),

$$A_n = n^{-1} \sum_{i \le n} \eta_i \eta_i' \ddot{\psi}(\gamma' D \eta_i) \quad \text{and} \quad B_n := \mathbb{P}_{n,\mu,K} A_n. \tag{19}$$

Now, let us define $\mathcal{X}_n = \mathcal{X}_{\mathbb{Z},n} \cap \mathcal{X}_{\eta,n} \cap \mathcal{X}_{A,n}$, where

$$\mathcal{X}_{\mathbb{Z},n} := \Big\{ \max_{i \le n} \|\mathbb{Z}_i\|^2 \le C_0 \log n \Big\}, \tag{20}$$

$$\mathcal{X}_{\eta,n} := \Big\{ \max_{i \le n} |\eta_i|^2 \le C_0 N \log n \Big\}, \tag{21}$$

$$\mathcal{X}_{A,n} := \big\{ \|A_n - B_n\|_2 \le \big(2\|B_n^{-1}\|_2\big)^{-1} \big\}. \tag{22}$$

If we choose a large enough universal constant $C_0 = C_0(\mathcal{F})$, Lemma 4 ensures that $\mathbb{P}_{n,\mu,K} \mathcal{X}_{\mathbb{Z},n}^c \le 2/n$ and $\mathbb{P}_{n,\mu,K} \mathcal{X}_{\eta,n}^c \le 2/n$ by choosing $\tau_{i,k} = \theta_i$ and $\tau_{i,k} = \{i \le N\}$, respectively, for all i, k; and Lemma 2 shows that

$$\|B_n^{-1}\|_2 = O_{\mathcal{F}}(1) \quad \text{and} \quad \mathbb{P}_{n,\mu,K} \|A_n - B_n\|_2^2 = o_{\mathcal{F}}(1),$$

thus $\mathbb{P}_{n,\mu,K} \mathcal{X}_{A,n}^c = o_{\mathcal{F}}(1)$. And hence,

$$\mathbb{P}_{n,\mu,K} \mathcal{X}_n^c \le \mathbb{P}_{n,\mu,K} \mathcal{X}_{\mathbb{Z},n}^c + \mathbb{P}_{n,\mu,K} \mathcal{X}_{\eta,n}^c + \mathbb{P}_{n,\mu,K} \mathcal{X}_{A,n}^c = o_{\mathcal{F}}(1). \tag{23}$$

Step 2. In the previous step, we show the assumptions of Lemma 1 are satisfied on the set $\mathcal{X}_n$. In this step, we show that the change-of-measure argument is ready to work. Let us consider the truncated model

$$\mathbb{Q}_{n,a,\mathbb{B},N} := \bigotimes_{i \le n} Q_{\lambda_{i,N}} \qquad \text{with } \lambda_{i,N} := \xi_i' \gamma.$$

"Change of measure" means to view the data $y_1, \ldots, y_n$ as if they are generated from the conditional joint distribution $\mathbb{Q}_{n,a,\mathbb{B},N}$, though the true distribution is $\mathbb{Q}_{n,a,\mathbb{B}}$. In this step, we show that the divergence caused by replacing $\mathbb{Q}_{n,a,\mathbb{B}}$ by $\mathbb{Q}_{n,a,\mathbb{B},N}$ is small enough that it will not compromise the asymptotic results. A common control of this divergence is the total variation distance between $\mathbb{Q}_{n,a,\mathbb{B},N}$ and $\mathbb{Q}_{n,a,\mathbb{B}}$. We show that there exists a sequence of nonnegative constants c_n of order $o_{\mathcal{F}}(\log n)$ such that

$$\|\mathbb{Q}_{n,a,\mathbb{B}} - \mathbb{Q}_{n,a,\mathbb{B},N}\|_{\mathrm{TV}}^2 \le e^{2c_n} \sum_{i \le n} |\lambda_i - \lambda_{i,N}|^2 \qquad \text{on } \mathcal{X}_n. \tag{24}$$

To establish inequality (24) we use the bound

$$\|\mathbb{Q}_{n,a,\mathbb{B}} - \mathbb{Q}_{n,a,\mathbb{B},N}\|_{\mathrm{TV}}^2 \le \mathbf{h}^2(\mathbb{Q}_{n,a,\mathbb{B}}, \mathbb{Q}_{n,a,\mathbb{B},N}) \le \sum_{i\le n} \mathbf{h}^2(Q_{\lambda_i}, Q_{\lambda_{i,N}}).$$

By Lemma 3

$$\mathbf{h}^2(Q_{\lambda_i}, Q_{\lambda_{i,N}}) \le \delta_i^2 \ddot{\psi}(\lambda_i)(1+|\delta_i|)G(|\delta_i|) \qquad \text{with } \delta_i := \lambda_i - \lambda_{i,N},$$

where

$$\begin{aligned} |\delta_i| &= |\lambda_i - \lambda_{i,N}| = \big|\langle \mathbb{Z}_i, \mathbb{B}\rangle - \langle H_N \mathbb{Z}_i, \mathbb{B}\rangle\big| = \big|\langle \mathbb{Z}_i, H_N^{\perp}\mathbb{B}\rangle\big| \\ &\le \|\mathbb{Z}_i\| \, \|H_N^{\perp}\mathbb{B}\| \le O_{\mathcal{F}}\big(\sqrt{N^{1-2\beta}\log n}\big) = o_{\mathcal{F}}(1). \end{aligned} \tag{25}$$

Because $\delta_i = o_{\mathcal{F}}(1)$ for each i, we know all the $(1+|\delta_i|)G(|\delta_i|)$ factors can be bounded by a single $O_{\mathcal{F}}(1)$ term.

Further, for $(a, \mathbb{B}, \mu, K) \in \mathcal{F}(R, \alpha, \beta)$ and with the $\|\mathbb{Z}_i\|$'s on the set $\mathcal{X}_n$,

$$|\lambda_i| \le |a| + \big(\|\mu\| + \|\mathbb{Z}_i\|\big)\|\mathbb{B}\| \le C_2\sqrt{\log n} \tag{26}$$

for some constant $C_2 = C_2(\mathcal{F})$. Assumption $(\ddot{\Psi})$ then ensures that all the $\ddot{\psi}(\lambda_i)$ are bounded by a single $\exp(o_{\mathcal{F}}(\log n))$ term.

Therefore, inequality (24) is proved to hold. This bound for total variation distance legitimates the change-of-measure argument in the next step.

Step 3. We apply the change-of-measure argument and Lemma 1 to complete the proof. On the set $\mathcal{X}_n$, we can apply Lemma 1 directly with $\mathbb{Q} = \mathbb{Q}_{n,a,\mathbb{B},N}$, because the conditions of Lemma 1 hold: inequality (14) holds by construction of $\mathcal{X}_n$ and inequality (15) holds for large enough n because

$$\max_{i\le n} |\eta_i|^2 \le O_{\mathcal{F}}(N\log n) = o_{\mathcal{F}}(\sqrt{n}/N).$$

In the equation above, the first inequality is due to the construction of $\mathcal{X}_n$, and the second equality is due to $N = o_{\mathcal{F}}(n^{1/(2+2\alpha)})$.

For each realization of the $\mathbb{X}_i$'s lying in $\mathcal{X}_n$, we invoke Lemma 1, with η_i, A_n, B_n, D and $\mathbb{Q}$ (notations) in Lemma 1 replaced by η_i, A_n, B_n, D and $\mathbb{Q}_{n,a,\mathbb{B},N}$ defined in this subsection, respectively, and it gives a high probability set $\mathcal{Y}_{m,\varepsilon}$ with $\mathbb{Q}_{n,a,\mathbb{B},N}\mathcal{Y}_{m,\varepsilon}^c < 2\varepsilon$ on which

$$\sum_{1\le k\le m} |\widehat{b}_k - b_k|^2 = O_{\mathcal{F}}\Big(n^{-1}\sum_{1\le k\le m}\theta_k^{-1}\Big) = O_{\mathcal{F}}\big(m^{1+\alpha}/n\big) = O_{\mathcal{F}}\big(n^{(1-2\beta)/(\alpha+2\beta)}\big),$$

which implies

$$\|\widehat{\mathbb{B}}_n - \mathbb{B}\|^2 = \sum_{1\le k\le m} |\widehat{b}_k - b_k|^2 + \sum_{k>m} b_k^2 = O_{\mathcal{F}}\big(n^{(1-2\beta)/(\alpha+2\beta)}\big).$$

From inequality (24) it follows, for a large enough constant C_ε, that

$$\begin{aligned}
&\mathbb{P}_{n,\mu,K}\mathbb{Q}_{n,a,\mathbb{B}}\{\|\widehat{\mathbb{B}}_n - \mathbb{B}\|^2 > C_\varepsilon n^{(1-2\beta)/(\alpha+2\beta)}\} \\
&\quad \le \mathbb{P}_{n,\mu,K}\mathcal{X}_n^c + \mathbb{P}_{n,\mu,K}\mathcal{X}_n\big(\|\mathbb{Q}_{n,a,\mathbb{B}} - \mathbb{Q}_{n,a,\mathbb{B},N}\|_{\mathrm{TV}} + \mathbb{Q}_{n,a,\mathbb{B},N}\mathcal{Y}_{m,\varepsilon}^c\big) \\
&\quad \le o_{\mathcal{F}}(1) + 2\varepsilon + e^{c_n}\Bigg(\sum_{i\le n}\mathbb{P}_{n,\mu,K}|\lambda_i - \lambda_{i,N}|^2\Bigg)^{1/2}.
\end{aligned}$$

By construction,

$$\lambda_i - \lambda_{i,N} = \sum_{k>N} z_{i,k}b_k$$

with the $z_{i,k}$'s independent and $z_{i,k} \sim N(0,\theta_k)$. Thus

$$\sum_{i\le n}\mathbb{P}_{n,\mu,K}|\lambda_i - \lambda_{i,N}|^2 \le n\sum_{k>N}\theta_k b_k^2 = O_{\mathcal{F}}\big(nN^{1-\alpha-2\beta}\big) = o_{\mathcal{F}}\big(e^{-2c_n}\big),$$

because $\zeta > (\alpha+2\beta-1)^{-1}$. That is, we have an estimator that achieves the $O_{\mathcal{F}}(n^{(1-2\beta)/(\alpha+2\beta)})$ minimax rate.

4.3. *Proof of Theorem 1 with unknown Gaussian distribution.* Let $\widehat{\mathbb{B}}_n$ be the two-stage estimator defined in (10) with cutoff points m and N defined in (12) and (13), respectively. In this section, we show that $\widehat{\mathbb{B}}_n$ achieves the asymptotic rates of convergence stated in Theorem 1.

As in Section 4.2, most of the analysis will be conditional on the $\mathbb{X}_i$'s lying in a set with high probability on which the various estimators and other random quantities are well behaved. In fact, we choose the high probability set as $\widetilde{\mathcal{X}}_{\varepsilon,n}$ that is defined in Lemma 5. The set $\widetilde{\mathcal{X}}_{\varepsilon,n}$ is an analogy to $\mathcal{X}_n$ in Section 4.2.

As in Section 4.2, the component of $\mathbb{B}$ orthogonal to $\operatorname{span}\{\widetilde{\phi}_1,\dots,\widetilde{\phi}_m\}$ causes no trouble because

$$\|\widehat{\mathbb{B}} - \mathbb{B}\|^2 = \sum_{1\le k\le m}(\widehat{b}_k - \widetilde{b}_k)^2 + \big\|\widetilde{H}_m^{\perp}\mathbb{B}\big\|^2$$

and, by Lemma 5 part (iii),

$$\big\|\widetilde{H}_m^{\perp}\mathbb{B}\big\|^2 \le 2\big\|H_m^{\perp}\mathbb{B}\big\|^2 + 2\big\|(\widetilde{H}_m - H_m)\mathbb{B}\big\|^2 = O_{\mathcal{F}}\big(n^{(1-2\beta)/(\alpha+2\beta)}\big) \qquad \text{on } \widetilde{\mathcal{X}}_{\varepsilon,n}.$$

To handle $\sum_{1\le k\le m}(\widehat{b}_k - \widetilde{b}_k)^2$, we invoke Lemma 1 for $\mathbb{X}_i$'s lying in $\widetilde{\mathcal{X}}_{\varepsilon,n}$, with η_i, A_n, B_n, D and $\mathbb{Q}$ (notations) in Lemma 1 replaced by $\widetilde{\eta}_i$, $\widetilde{A}_n$, $\widetilde{B}_n$, D and $\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N}$, respectively, where

$$\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N} := \bigotimes_{i\le n} Q_{\widetilde{\lambda}_{i,N}}.$$

And, it gives a high probability set $\widetilde{\mathcal{Y}}_{m,\varepsilon}$ with $\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N}\widetilde{\mathcal{Y}}_{m,\varepsilon}^c < 2\varepsilon$ on which

$$\sum_{1\le k\le m}(\widehat{b}_k - \widetilde{b}_k)^2 = O_{\mathcal{F}}\big(n^{(1-2\beta)/(\alpha+2\beta)}\big).$$

The conditions of Lemma 1 are satisfied on $\widetilde{\mathfrak{X}}_{\varepsilon,n}$ when n is large, because of Lemma 5 part (v) and

$$\|\widetilde{A}_n - \widetilde{B}_n\|_2 \le \|\widetilde{A}_n - \widetilde{S}A_n\widetilde{S}\|_2 + \|\widetilde{S}A_n\widetilde{S} - \widetilde{S}B_n\widetilde{S}\|_2 = o_{\mathcal{F}}(1),$$

where the first part $\|\widetilde{A}_n - \widetilde{S}A_n\widetilde{S}\|_2 = o_{\mathcal{F}}(1)$ is due to Lemma 5 part (vi), and the second part $\|\widetilde{S}A_n\widetilde{S} - \widetilde{S}B_n\widetilde{S}\|_2 = o_{\mathcal{F}}(1)$ is due to Lemma 2.

Now, to complete the proof it suffices to show that $\|\mathbb{Q}_{n,a,\mathbb{B},N} - \widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N}\|_{\mathrm{TV}}$ tends to zero. First note that

$$\begin{aligned}\widetilde{\lambda}_{i,N} - \lambda_{i,N} &= a + \langle \mathbb{B}, \overline{\mathbb{X}}\rangle + \langle \widetilde{H}_N\mathbb{B}, \mathbb{Z}_i - \overline{\mathbb{Z}}\rangle - a - \langle \mathbb{B}, \mu\rangle - \langle H_N\mathbb{B}, \mathbb{Z}_i\rangle \\ &= \langle \widetilde{H}_N^{\perp}\mathbb{B}, \overline{\mathbb{Z}}\rangle - \langle H_N^{\perp}\mathbb{B}, \overline{\mathbb{Z}}\rangle + \langle H_N^{\perp}\mathbb{B}, \overline{\mathbb{Z}}\rangle + \langle \widetilde{H}_N\mathbb{B} - H_N\mathbb{B}, \mathbb{Z}_i\rangle,\end{aligned}$$

which implies that, on $\widetilde{\mathfrak{X}}_{\varepsilon,n}$,

$$\begin{aligned}|\widetilde{\lambda}_{i,N} - \lambda_{i,N}|^2 &\le 2|\langle H_N^{\perp}\mathbb{B}, \overline{\mathbb{Z}}\rangle|^2 + 2\|\widetilde{H}_N\mathbb{B} - H_N\mathbb{B}\|^2\big(\|\mathbb{Z}_i\| + \|\overline{\mathbb{Z}}\|\big)^2 \\ &\le O_{\mathcal{F}}\big(N^{1-2\beta}\big)C_\varepsilon^2 n^{-1} + O_{\mathcal{F}}\big(n^{-1-\nu}\big)C_\varepsilon^2\big(n^{-1/2} + \sqrt{\log n}\big)^2 \qquad (27) \\ &= O_{\mathcal{F}}\big(n^{-1-\nu'}\big) \qquad \text{for some } 0 < \nu' < \nu.\end{aligned}$$

Now we can argue as in step 2 of the proof for the case of known K: on $\widetilde{\mathfrak{X}}_{\varepsilon,n}$,

$$\begin{aligned}\|\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N} - \mathbb{Q}_{n,a,\mathbb{B},N}\|_{\mathrm{TV}}^2 &\le \sum_{i\le n}\mathbf{h}^2\big(Q_{\widetilde{\lambda}_{i,N}}, Q_{\lambda_{i,N}}\big) \\ &\le \exp\big(o_{\mathcal{F}}(\log n)\big)\sum_{i\le n}|\widetilde{\lambda}_{i,N} - \lambda_{i,N}|^2 \\ &= o_{\mathcal{F}}(1).\end{aligned}$$

Finish the argument as in Section 4.2, by splitting into contributions from $\widetilde{\mathfrak{X}}_{\varepsilon,n}^c$ and $\widetilde{\mathfrak{X}}_{\varepsilon,n} \cap \widetilde{\mathfrak{Y}}_{m,\varepsilon}^c$ and $\widetilde{\mathfrak{X}}_{\varepsilon,n} \cap \widetilde{\mathfrak{Y}}_{m,\varepsilon}$.

5. Proof of Theorem 2. We apply a slight variation on Assouad's lemma—combining ideas from Yu (1997) and from van der Vaart [(1998), Section 24.3]—to establish the minimax lower bound result in Theorem 2.

We consider behavior only for $\mu = 0$, $a = 0$ and a fixed K with spectral decomposition $\sum_{j\in\mathbb{N}}\theta_j\phi_j \otimes \phi_j$ satisfying assumption (K). For simplicity we abbreviate $\mathbb{P}_{n,0,K}$ to $\mathbb{P}$. Let $J = \{m+1, m+2, \ldots, 2m\}$ and $\Gamma = \{0,1\}^J := \{\gamma = (\gamma_{m+1}, \ldots, \gamma_{2m}) | \gamma_j = 0 \text{ or } \gamma_j = 1\}$. Let $\beta_j = Rj^{-\beta}$. For each γ in Γ define $\mathbb{B}_\gamma = \varepsilon\sum_{j\in J}\gamma_j\beta_j\phi_j$, for a small $\varepsilon > 0$ to be specified, and write $\mathbb{Q}_\gamma$ for the product measure $\bigotimes_{i\le n} Q_{\lambda_i(\gamma)}$ with $\lambda_i(\gamma) = \langle \mathbb{B}_\gamma, \mathbb{Z}_i\rangle = \varepsilon\sum_{j\in J}\gamma_j\beta_j z_{i,j}$.

For each j let $\Gamma_j = \{\gamma \in \Gamma : \gamma_j = 1\}$ and let ψ_j be the bijection on Γ that flips the jth coordinate but leaves all other coordinates unchanged. Let π be the uniform distribution on Γ, that is, $\pi_\gamma = 2^{-m}$ for each γ.

For each estimator $\widehat{\mathbb{B}} = \sum_{j\in\mathbb{N}} \widehat{b}_j \phi_j$ we have $\|\mathbb{B}_\gamma - \widehat{\mathbb{B}}\|^2 \geq \sum_{j\in J} (\varepsilon\gamma_j\beta_j - \widehat{b}_j)^2$, and so

$$\begin{aligned}\sup_{\mathcal{F}} \mathbb{P}_{n,f}\|\mathbb{B} - \widehat{\mathbb{B}}\|^2 &\geq \sum_{\gamma\in\Gamma} \pi_\gamma \sum_{j\in J} \mathbb{P}\mathbb{Q}_\gamma(\varepsilon\gamma_j\beta_j - \widehat{b}_j)^2 \\ &= 2^{-m} \sum_{j\in J} \sum_{\gamma\in\Gamma_j} \mathbb{P}\big(\mathbb{Q}_\gamma(\varepsilon\beta_j - \widehat{b}_j)^2 + \mathbb{Q}_{\psi_j(\gamma)}(0 - \widehat{b}_j)^2\big) \qquad (28) \\ &\geq 2^{-m} \sum_{j\in J} \sum_{\gamma\in\Gamma_j} \frac{1}{4}(\varepsilon\beta_j)^2 \mathbb{P}\|\mathbb{Q}_\gamma \wedge \mathbb{Q}_{\psi_j(\gamma)}\|,\end{aligned}$$

where the first lower bound is due to the fact that the supremum over $\mathcal{F}$ is not less than the average over a subset of $\mathcal{F}$, and the last lower bound comes from the fact that

$$(\varepsilon\beta_j - \widehat{b}_j)^2 + (0 - \widehat{b}_j)^2 \geq \tfrac{1}{4}(\varepsilon\beta_j)^2 \qquad \text{for all } \widehat{b}_j.$$

We assert that, if ε is chosen appropriately,

$$\min_{j,\gamma} \mathbb{P}\|\mathbb{Q}_\gamma \wedge \mathbb{Q}_{\psi_j(\gamma)}\| \qquad \text{stays bounded away from zero as } n \to \infty, \qquad (29)$$

which will ensure that the lower bound in (28) is eventually larger than a constant multiple of $\sum_{j\in J} \beta_j^2 \geq cn^{(1-2\beta)/(\alpha+2\beta)}$ for some constant $c > 0$. The inequality in Theorem 2 will then follow.

To prove (29), consider a γ in Γ and the corresponding $\gamma' = \psi_j(\gamma)$. By virtue of the inequality

$$\|\mathbb{Q}_\gamma \wedge \mathbb{Q}_{\gamma'}\| = 1 - \|\mathbb{Q}_\gamma - \mathbb{Q}_{\gamma'}\|_{\mathrm{TV}} \geq 1 - \Big(2 \wedge \sum_{i\leq n} \mathbf{h}^2(Q_{\lambda_i(\gamma)}, Q_{\lambda_i(\gamma')})\Big)^{1/2},$$

it is enough to show that

$$\limsup_{n\to\infty} \max_{j,\gamma} \mathbb{P}\Big(2 \wedge \sum_{i\leq n} \mathbf{h}^2(Q_{\lambda_i(\gamma)}, Q_{\lambda_i(\gamma')})\Big) < 1. \qquad (30)$$

Define $\mathcal{X}_n = \{\max_{i\leq n} \|\mathbb{Z}_i\|^2 \leq C_0 \log n\}$. Based on Lemma 4, we know that $\mathbb{P}\mathcal{X}_n^c = o(1)$ with the constant C_0 large enough. On $\mathcal{X}_n$ we have

$$\big|\lambda_i(\gamma)\big|^2 \leq \sum_{j\in J} \beta_j^2 \|Z_i\|^2 = O\big(n^{(1-2\beta)/(\alpha+2\beta)} \log n\big) = o(1),$$

and, by inequality in Lemma 3, there exits a universal constant $C > 0$ such that

$$\mathbf{h}^2(Q_{\lambda_i(\gamma)}, Q_{\lambda_i(\gamma')}) \leq C\big|\lambda_i(\gamma) - \lambda_i\big(\gamma'\big)\big|^2 \leq C\varepsilon^2\beta_j^2 z_{i,j}^2.$$

We deduce that

$$\begin{aligned}\mathbb{P}\Big(2 \wedge \sum_{i\leq n} h^2(Q_{\lambda_i(\gamma)}, Q_{\lambda_i(\gamma')})\Big) &\leq 2\mathbb{P}\mathcal{X}_n^c + C\sum_{i\leq n} \varepsilon^2\beta_j^2 \mathbb{P}\mathcal{X}_n z_{i,j}^2 \\ &\leq o(1) + C\varepsilon^2 n\beta_j^2\theta_j.\end{aligned}$$

The choice of J makes $\beta_j^2\theta_j \le R^2 m^{-\alpha-2\beta} \sim R^2/n$. Assertion (30) follows for any small enough ε.

6. Proofs of technical lemmas.

6.1. *Proof of Lemma 1.* We need to first show the following lemma. Note that $J_n = \sum_{i\le n} \xi_i \xi_i' \ddot{\psi}(\lambda_i)$. To avoid an excess of parentheses we write N_+ for $N+1$. We define $w_i := J_n^{-1/2}\xi_i$ and $W_n = \sum_{i\le n} w_i(y_i - \dot{\psi}(\lambda_i))$. Notice that $\mathbb{Q}W_n = 0$ and $\operatorname{var}_{\mathbb{Q}}(W_n) = \sum_{i\le n} w_i w_i' \ddot{\psi}(\lambda_i) = I_{N_+}$ and

$$\mathbb{Q}|W_n|^2 = \operatorname{trace}\bigl(\operatorname{var}_{\mathbb{Q}}(W_n)\bigr) = N_+.$$

LEMMA 6. *Suppose* $0 < \varepsilon_1 \le 1/2$ *and* $0 < \varepsilon_2 < 1$ *and*

$$\max_{i\le n}|w_i| \le \frac{\varepsilon_1\varepsilon_2}{2G(1)N_+} \qquad \textit{with } G \textit{ as in assumption } (\dddot{\Psi}).$$

Then, the MLE $\widehat{g}$ *has the decomposition* $\widehat{g} = \gamma + J_n^{-1/2}(W_n + r_n)$ *with* $|r_n| \le \varepsilon_1$ *on the set* $\{|W_n| \le \sqrt{N_+/\varepsilon_2}\}$, *which has* $\mathbb{Q}$*-probability greater than* $1-\varepsilon_2$.

PROOF. The equality $\mathbb{Q}|W_n|^2 = N_+$ and Chebyshev's inequality give

$$\mathbb{Q}\bigl\{|W_n| > \sqrt{N_+/\varepsilon_2}\bigr\} \le \varepsilon_2.$$

Reparametrize by defining $t = J_n^{1/2}(g-\gamma)$. The concave function

$$\mathcal{L}_n(t) := L_n\bigl(\gamma + J_n^{-1/2}t\bigr) - L_n(\gamma) = \sum_{i\le n} y_i w_i' t + \psi(\lambda_i) - \psi\bigl(\lambda_i + w_i' t\bigr)$$

is maximized at $\widehat{t}_n = J_n^{1/2}(\widehat{g}-\gamma)$. It has derivative

$$\dot{\mathcal{L}}_n(t) = \sum_{i\le n} w_i\bigl(y_i - \dot{\psi}(\lambda_i + w_i' t)\bigr).$$

For a fixed unit vector $u \in \mathbb{R}^{N_+}$ and a fixed $t \in \mathbb{R}^{N_+}$, consider the real-valued function of the real variable s,

$$H(s) := u'\dot{\mathcal{L}}_n(st) = \sum_{i\le n} u' w_i\bigl(y_i - \dot{\psi}(\lambda_i + s w_i' t)\bigr),$$

which has derivatives

$$\dot{H}(s) = -\sum_{i\le n}(u'w_i)(w_i' t)\ddot{\psi}(\lambda_i + s w_i' t),$$

$$\ddot{H}(s) = -\sum_{i\le n}(u'w_i)(w_i' t)^2\dddot{\psi}(\lambda_i + s w_i' t).$$

Notice that $H(0) = u'W_n$ and $\dot{H}(0) = -u'\sum_{i\le n} w_i w_i' \ddot{\psi}(\lambda_i)t = -u't$.

Write $M_n := \max_{i \le n} |w_i|$. By virtue of assumption $(\ddot{\Psi})$,

$$\begin{aligned}
|\ddot{H}(s)| &\le \sum_{i \le n} |u' w_i| (w_i' t)^2 \ddot{\psi}(\lambda_i) G(|s w_i' t|) \\
&\le M_n G(M_n |st|) t' \sum_{i \le n} w_i w_i' \ddot{\psi}(\lambda_i) t \\
&= M_n G(M_n |st|) |t|^2.
\end{aligned}$$

By Taylor expansion, for some $0 < s^* < 1$,

$$|H(1) - H(0) - \dot{H}(0)| \le \tfrac{1}{2} |\ddot{H}(s^*)| \le \tfrac{1}{2} M_n G(M_n |t|) |t|^2.$$

That is,

$$|u' (\dot{\mathcal{L}}_n(t) - W_n + t)| \le \tfrac{1}{2} M_n G(M_n |t|) |t|^2. \tag{31}$$

Approximation (31) will control the behavior of $\widetilde{\mathcal{L}}(s) := \mathcal{L}_n(W_n + su)$, a concave function of the real argument s, for each unit vector u. By concavity, the derivative $\dot{\widetilde{\mathcal{L}}}(s)$ is a decreasing function of s. Let us decompose $\dot{\widetilde{\mathcal{L}}}(s)$ in the following way:

$$\dot{\widetilde{\mathcal{L}}}(s) = u' \dot{\mathcal{L}}_n(W_n + su) = -s + R(s),$$

where

$$|R(s)| \le \tfrac{1}{2} M_n G(M_n |W_n + su|) |W_n + su|^2.$$

On the set $\{|W_n| \le \sqrt{N_+/\varepsilon_2}\}$ we have

$$|W_n \pm \varepsilon_1 u| \le \sqrt{N_+/\varepsilon_2} + \varepsilon_1.$$

Thus

$$M_n |W_n \pm \varepsilon_1 u| \le \frac{\varepsilon_1 \varepsilon_2}{2 G(1) N_+} (\sqrt{N_+/\varepsilon_2} + \varepsilon_1) < 1,$$

implying

$$\begin{aligned}
|R(\pm\varepsilon_1)| &\le \frac{1}{2} M_n G(1) |W_n \pm \varepsilon_1 u|^2 \le \frac{\varepsilon_1 \varepsilon_2}{G(1) N_+} (N_+/\varepsilon_2 + \varepsilon_1^2) \\
&\le \varepsilon_1 (1 + \varepsilon_1^2 \varepsilon_2 / N_+) < \frac{5}{8} \varepsilon_1.
\end{aligned}$$

Deduce that

$$\dot{\widetilde{\mathcal{L}}}(\varepsilon_1) = -\varepsilon_1 + R(\varepsilon_1) \le -\tfrac{3}{8} \varepsilon_1 \quad \text{and} \quad \dot{\widetilde{\mathcal{L}}}(-\varepsilon_1) = \varepsilon_1 + R(-\varepsilon_1) \ge \tfrac{3}{8} \varepsilon_1.$$

The concave function $s \mapsto \mathcal{L}_n(W_n + su)$ must achieve its maximum for some s in the interval $[-\varepsilon_1, \varepsilon_1]$, for each unit vector u. It follows that $|\widehat{t}_n - W_n| \le \varepsilon_1$. $\square$

First we establish a bound on the spectral distance between A_n^{-1} and B_n^{-1}. Define $H = B_n^{-1}A_n - I$. Then $\|H\|_2 \le \|B_n^{-1}\|_2\|A_n - B_n\|_2 \le 1/2$, which justifies the expansion

$$\|A_n^{-1} - B_n^{-1}\|_2 = \|((I+H)^{-1} - I)B_n^{-1}\|_2 \le \sum_{j\ge 1}\|H\|_2^k\|B_n^{-1}\|_2 \le \|B_n^{-1}\|_2.$$

As a consequence, $\|A_n^{-1}\|_2 \le 2\|B_n^{-1}\|_2$.

Choose $\varepsilon_1 = 1/2$ and $\varepsilon_2 = \varepsilon$ in Lemma 6. The bound on $\max_{i\le n}|\eta_i|$ gives the bound on $\max_{i\le n}|w_i|$ needed by the lemma

$$n|w_i|^2 = \eta_i' D(J_n/n)^{-1}D\eta_i = \eta_i' A_n^{-1}\eta_i \le \|A_n^{-1}\|_2|\eta_i|^2.$$

As shown in Lemma 6, the MLE $\widehat{g}$ can be decomposed as

$$\widehat{g} = \gamma + J_n^{-1/2}(W_n + r_n).$$

Define $K_j := J_n^{-1/2}\kappa_j$, so that $|\kappa_j'(\widehat{g}-\gamma)|^2 \le 2(K_j'W_n)^2 + 2(K_j'r_n)^2$. By Cauchy–Schwarz,

$$\sum_j (K_j'r_n)^2 \le \sum_j |K_j|^2|r_n|^2 = U_\kappa|r_n|^2,$$

where

$$U_\kappa := \sum_j \kappa_j' J_n^{-1}\kappa_j = \sum_j n^{-1}(D^{-1}\kappa_j)'A_n^{-1}D^{-1}\kappa_j \le 2n^{-1}\|B_n^{-1}\|_2\sum_j|D^{-1}\kappa_j|^2.$$

For the contribution $V_\kappa := \sum_j |K_j'W_n|^2$, the Cauchy–Schwarz bound is too crude. Instead, notice that $\mathbb{Q}V_\kappa = U_\kappa$, which ensures that the complement of the set

$$\mathcal{Y}_{\kappa,\varepsilon} := \{|W_n| \le \sqrt{N_+/\varepsilon}\} \cap \{V_\kappa \le U_\kappa/\varepsilon\}$$

has $\mathbb{Q}$ probability less that 2ε. On the set $\mathcal{Y}_{\kappa,\varepsilon}$,

$$\sum_{0\le j\le N}|\kappa_j'(\widehat{g}-\gamma)|^2 \le 2V_\kappa + 2U_\kappa|r_n|^2 \le 3U_\kappa/\varepsilon.$$

The asserted bound follows.

6.2. *Proof of Lemma 2.* Throughout this subsection, abbreviate $\mathbb{P}_{n,\mu,K}$ to $\mathbb{P}$. The matrix A_n is an average of n independent random matrices each of which is distributed like $\mathcal{N}\mathcal{N}'\ddot{\psi}(\gamma'D\mathcal{N})$, where $\mathcal{N} = (\mathcal{N}_0, \mathcal{N}_1, \ldots, \mathcal{N}_N)'$ with $\mathcal{N}_0 \equiv 1$, and the other $\mathcal{N}_j$'s are independent $N(0,1)$'s. Moreover, by rotational invariance of the spherical normal, we may assume with no loss of generality that $\gamma'D\mathcal{N} = \bar{a} + \kappa\mathcal{N}_1$, where

$$\kappa^2 = \sum_{k=1}^N D_k^2 b_k^2 = O_{\mathcal{F}}(1).$$

Thus

$$B_n = \mathbb{P}\mathcal{N}\mathcal{N}'\ddot{\psi}(\bar{a} + \kappa\mathcal{N}_1) = \operatorname{diag}(F, r_0 I_{N-1}),$$

where

$$r_j := \mathbb{P}\mathcal{N}_1^j \ddot{\psi}(\bar{a} + \kappa\mathcal{N}_1) \quad \text{and} \quad F = \begin{bmatrix} r_0 & r_1 \\ r_1 & r_2 \end{bmatrix}.$$

The block diagonal form of B_n simplifies calculation of spectral norms,

$$\begin{aligned} \|B_n^{-1}\|_2 &= \|\operatorname{diag}(F^{-1}, r_0^{-1} I_{N-1})\|_2 \\ &\le \max(\|F^{-1}\|_2, \|r_0^{-1} I_{N-1}\|_2) \le \max\left(\frac{r_0 + r_2}{r_0 r_2 - r_1^2}, r_0^{-1}\right). \end{aligned}$$

Assumption $(\ddot{\Psi})$ ensures that both r_0 and r_2 are $O_{\mathcal{F}}(1)$.

Continuity and strict positivity of $\ddot{\psi}$, together with $\max(|\bar{a}|, \kappa) = O_{\mathcal{F}}(1)$, ensure that $c_0 := \inf_{\bar{a},\kappa} \inf_{|x|\le 1} \ddot{\psi}(\bar{a} + \kappa x) > 0$. Thus

$$\sqrt{2\pi} r_0 \ge c_0 \int_{-1}^{+1} e^{-x^2/2}\, dx > 0.$$

Similarly,

$$\begin{aligned} \sqrt{2\pi}(r_0 r_2 - r_1^2) &= \sqrt{2\pi} r_0 \mathbb{P}\ddot{\psi}(\bar{a} + \kappa\mathcal{N}_1)(\mathcal{N}_1 - r_1/r_0)^2 \\ &\ge c_0 r_0 \int_{-1}^{+1} (x - r_1/r_0)^2 e^{-x^2/2}\, dx \\ &\ge c_0 r_0 \int_{-1}^{+1} x^2 e^{-x^2/2}\, dx. \end{aligned}$$

It follows that $\|B_n^{-1}\|_2 = O_{\mathcal{F}}(1)$.

The random matrix $A_n - B_n$ is an average of n independent random matrices, each distributed like $\mathcal{N}\mathcal{N}'\ddot{\psi}(\bar{a} + \kappa\mathcal{N}_1)$ minus its expected value. Thus

$$\mathbb{P}\|A_n - B_n\|_2^2 \le \mathbb{P}\|A_n - B_n\|_{\mathrm{F}}^2 = n^{-1} \sum_{0 \le j,k \le N} \operatorname{var}\bigl(\mathcal{N}_j \mathcal{N}_k \ddot{\psi}(\bar{a} + \kappa\mathcal{N}_1)\bigr),$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Assumption $(\ddot{\Psi})$ ensures that each summand is $O_{\mathcal{F}}(1)$, which leaves us with a $O_{\mathcal{F}}(N^2/n) = o_{\mathcal{F}}(1)$ upper bound.

6.3. *Proof of Lemma 3.* Let us temporarily write λ' for $\lambda + \delta$ and write $\overline{\lambda}$ for $(\lambda + \lambda')/2 = \lambda + \delta/2$,

$$\begin{aligned} 1 - \frac{1}{2}\mathbf{h}^2(Q_\lambda, Q_{\lambda'}) &= \int \sqrt{f_\lambda(y) f_{\lambda'}(y)} = \int \exp\left(\overline{\lambda} y - \frac{1}{2}\psi(\lambda) - \frac{1}{2}\psi(\lambda')\right) \\ &= \exp\left(\psi(\overline{\lambda}) - \frac{1}{2}\psi(\lambda) - \frac{1}{2}\psi(\lambda')\right) \\ &\ge 1 + \psi(\overline{\lambda}) - \frac{1}{2}\psi(\lambda) - \frac{1}{2}\psi(\lambda'). \end{aligned}$$

That is,

$$\mathbf{h}^2(Q_\lambda, Q_{\lambda'}) \le \psi(\lambda) + \psi(\lambda+\delta) - 2\psi(\lambda+\delta/2).$$

By Taylor expansion in δ around 0, the right-hand side is less than

$$\tfrac{1}{4}\delta^2\ddot{\psi}(\lambda) + \tfrac{1}{6}\delta^3\big(\dddot{\psi}(\lambda+\delta^*) - \tfrac{1}{4}\dddot{\psi}(\lambda+\delta^*/2)\big),$$

where $0 < |\delta^*| < |\delta|$. Invoke inequality $(\dddot{\Psi})$ twice to bound the coefficient of $\delta^3/6$ in absolute value by

$$\ddot{\psi}(\lambda)\big(G(|\delta|) + \tfrac{1}{4}G(|\delta|/2)\big) \le \tfrac{5}{4}\ddot{\psi}(\lambda)G(|\delta|).$$

The stated bound simplifies some unimportant constants.

6.4. *Proof of Lemma 4.* Without loss of generality, let us suppose $T = 1$. For $s = 1/4$, note that

$$\mathbb{P}[\exp(sW_i)] = \prod_{k\in\mathbb{N}}(1-2s\tau_{i,k})^{-1/2} \le \exp\Big(\sum_{k\in\mathbb{N}} s\tau_{i,k}\Big) \le e^{1/4}$$

by virtue of the inequality $-\log(1-t) \le 2t$ for $|t| \le 1/2$. With the same s, it then follows that

$$\begin{aligned}\mathbb{P}\Big\{\max_{i\le n} W_i > 4(\log n + x)\Big\} &\le \exp(-4s(\log n + x))\mathbb{P}\Big[\exp\Big(\max_{i\le n} sW_i\Big)\Big] \\ &\le e^{-x}\frac{1}{n}\sum_{i\le n}\mathbb{P}[\exp(sW_i)].\end{aligned}$$

The 2 is just a clean upper bound for $e^{1/4}$.

6.5. *Proof of Lemma 5.* We first show some preliminary lemmas in Section 6.5.1. Those preliminary results are used in the main proofs throughout Sections 6.5.2 to 6.6. For notational simplicity, we write $\sum_j^*$ for $\sum_{j\ne k}$.

6.5.1. *Preliminary lemmas.* Remember that θ_j's are the eigenvalues of K as defined in Definition 1. Many of the inequalities in the proof of Lemma 5 involve sums of functions of the θ_j's. The following result will save us a lot of repetition.

LEMMA 7. (i) *For each $r \ge 1$ there is a constant $C_r = C_r(\mathcal{F})$ for which*

$$\kappa_k(r,\gamma) := \sum_{j\in\mathbb{N}}\{j\ne k\}\frac{j^{-\gamma}}{|\theta_j-\theta_k|^r} \le \begin{cases} C_r\big(1+k^{r(1+\alpha)-\gamma}\big), & \text{if } r>1,\\ C_1\big(1+k^{1+\alpha-\gamma}\log k\big), & \text{if } r=1.\end{cases}$$

(ii) *For each p,*

$$\sum_{k\le p}\sum_{j>p}\frac{k^{-\alpha-2\beta}j^{-\alpha}}{|\theta_k-\theta_j|^2} = O_{\mathcal{F}}\big(p^{1-\alpha}\big).$$

PROOF. For (i), argue in the same way as Hall and Horowitz [(2007), page 85], using the lower bounds

$$|\theta_j - \theta_k| \geq \begin{cases} c_\alpha j^{-\alpha}, & \text{if } j < k/2, \\ c_\alpha |j-k| k^{-\alpha-1}, & \text{if } k/2 \leq j \leq 2k, \\ c_\alpha k^{-\alpha}, & \text{if } j > 2k, \end{cases}$$

where c_α is a positive constant.

For (ii), split the range of summation into two subsets: $\{(k, j) : j > \max(p, 2k)\}$ and $\{(k, j) : p/2 < k \leq p < j \leq 2k\}$. The first subset contributes at most

$$\sum_{k \leq p} k^{-\alpha-2\beta} \sum_{j > \max(p,2k)} j^{-\alpha} (c_\alpha k^{-\alpha})^{-2} = O_{\mathcal{F}}(p^{1-\alpha}),$$

because $\alpha - 2\beta < -3$. The second subset contributes at most

$$\sum_{p/2 < k \leq p} k^{-\alpha-2\beta} c_\alpha^{-2} k^{2\alpha+2} \sum_{j > p} j^{-\alpha} (j-k)^{-2} = O_{\mathcal{F}}(p^{2+\alpha-2\beta} p^{1-\alpha}),$$

which is of order $o_{\mathcal{F}}(p^{-\alpha})$. □

Remember that $z_{i,j} = \langle \mathbb{Z}_i, \phi_j \rangle$ and the standardized variables $\eta_{i,j} = z_{i,j}/\sqrt{\theta_j}$ are independent $N(0, 1)$'s. Define $\eta_{\cdot j} = n^{-1} \sum_{i \leq n} \eta_{i,j}$ and

$$\mathcal{C}_{j,k} := (n-1)^{-1} \sum_{i \leq n} (\eta_{i,j} - \eta_{\cdot j})(\eta_{i,k} - \eta_{\cdot k}),$$

the (j, k)-element of a sample covariance matrix of i.i.d. $N(0, I_N)$ random vectors. We further define

$$\Lambda_k := \sum_{j \in \mathbb{N}} \Lambda_{k,j} \phi_j \qquad \text{with } \Lambda_{k,j} := \begin{cases} \sqrt{\theta_j \theta_k}\, \mathcal{C}_{j,k}/(\theta_k - \theta_j), & \text{if } j \neq k, \\ 0, & \text{if } j = k. \end{cases}$$

In fact, most of the inequalities that we need for proving Lemma 5 come from simple moment bounds (Lemma 8) for the sample covariances $\mathcal{C}_{j,k}$ and the derived bounds (Lemma 9) for the Λ_k's. The distribution of $\mathcal{C}_{j,k}$ does not depend on the parameters of our model. By the rotation of axes we can rewrite $(n-1)\mathcal{C}_{j,k}$ as $U_j' U_k$, where $U_1, U_2, \ldots$ are independent $N(0, I_{n-1})$ random vectors. This representation gives some useful equalities and bounds.

LEMMA 8. *Uniformly over distinct* j, k, ℓ:

(i) $\mathbb{P}\mathcal{C}_{j,j} = 1$ *and* $\mathbb{P}(\mathcal{C}_{j,j} - 1)^2 = 2(n-1)^{-1}$;
(ii) $\mathbb{P}\mathcal{C}_{j,k} = \mathbb{P}\mathcal{C}_{j,k}\mathcal{C}_{j,\ell} = 0$;
(iii) $\mathbb{P}\mathcal{C}_{j,k}^2 = O(n^{-1})$.

PROOF. Assertion (i) is classical because $|U_j|^2 \sim \chi^2_{n-1}$. For assertion (ii) use $\mathbb{P}(U_1' U_2 | U_2) = 0$ and

$$\mathbb{P}(U_1' U_2 U_2' U_3 | U_2) = \operatorname{trace}\bigl(U_2 U_2' \mathbb{P}(U_3 U_1')\bigr) = 0.$$

For (iii) use $\mathbb{P}(U_1 U_1') = I_{n-1}$ and

$$\mathbb{P}(U_1' U_2 U_2' U_1 | U_2) = \operatorname{trace}\bigl(U_2 U_2' \mathbb{P}(U_1 U_1')\bigr) = \operatorname{trace}(U_2 U_2') = |U_2|^2. \qquad \square$$

LEMMA 9. *Uniformly over distinct j, k, ℓ:*

(i) $\mathbb{P}\Lambda_{k,j} = \mathbb{P}\Lambda_{k,j}\Lambda_{k,\ell} = 0$;
(ii) $\mathbb{P}\Lambda_{k,j}^2 = O_{\mathcal{F}}(n^{-1} k^{-\alpha} j^{-\alpha} (\theta_k - \theta_j)^{-2})$;
(iii) $\mathbb{P}\|\Lambda_k\|^2 = O_{\mathcal{F}}(n^{-1} k^2)$.

PROOF. Assertions (i) and (ii) follow from assertions (ii) and (iii) of Lemma 8. For (iii), note that

$$\mathbb{P}\|\Lambda_k\|^2 = \sum_j{}^{*} \mathbb{P}\Lambda_{j,k}^2 = O_{\mathcal{F}}\bigl(n^{-1} k^{-\alpha}\bigr) \kappa_k(2, \alpha)$$

and $\kappa_k(2, \alpha) = O_{\mathcal{F}}(k^{2+\alpha})$ from Lemma 7. $\square$

The following two lemmas related to perturbation theory for self-adjoint compact operators [cf., e.g., Birman and Solomjak (1987), Bosq (2000), Kato (1995)] are crucial in the development of Lemma 5. They are special cases of Lemmas 2 and 4 in the supplemental article [Dou, Pollard and Zhou (2012)] under the general perturbation-theoretic framework. For Lemma 10, similar results were established by other authors; see, for example, Hall and Hosseini-Nasab (2006), equation 2.8, and Cai and Hall (2006), Section 5.6. Lemma 11 extends the perturbation result for eigenprojections, obtained by Tyler [(1981), Lemma 4.1], from the matrix case to the general operator case.

Define

$$\varepsilon_k := \min\bigl\{|\theta_j - \theta_k| : j \neq k\bigr\}$$

and

$$f_k := \sigma_k \widetilde{\phi}_k - \phi_k \qquad \text{for all } k.$$

LEMMA 10. *If we have $\varepsilon_k > 5\|\Delta\|$, then it follows that*

$$\|f_k\| \leq 3\|\Lambda_k\|.$$

Define $H_J = \operatorname{span}\{\phi_j : j \in J\}$ and $\widetilde{H}_J = \operatorname{span}\{\widetilde{\phi}_j : j \in J\}$ for $J \subseteq \mathbb{N}$.

LEMMA 11. *If we have* $\min_{k\in J}\varepsilon_k > 5\|\Delta\|$, *then it follows that*

$$(\widetilde{H}_J - H_J)\mathbb{B} = \sum_{j\in J}\sum_{k\in J^c}\phi_j b_k(\Lambda_{j,k}+\Lambda_{k,j}) + e,$$

where $\|e\|^2$ *is bounded by a universal constant times* $R_1 + \|\Delta\|^2 R_2$ *with*

$$R_1 = \Bigl(\sum_{k\in J}\|\Lambda_k\|^2\Bigr)\sum_{k\in J}\Bigl(\sum_j{}^{*}\Lambda_{k,j}b_j\Bigr)^2,$$

$$R_2 = \sum_{k\in J}\|\Lambda_k\|^2\Bigl(\sum_j{}^{*}\frac{|b_j|}{|\theta_k-\theta_j|}\Bigr)^2 + \Bigl(\sum_{k\in J}\|\Lambda_k\||b_k|\sum_j{}^{*}\frac{1}{|\theta_k-\theta_j|}\Bigr)^2 + \sum_{k\in J}\|\Lambda_k\|^2|b_k|^2k^{2+2\alpha}.$$

6.5.2. *A high probability set* $\widetilde{\mathfrak{X}}_{\varepsilon,n}$. To prove Lemma 5 we define $\widetilde{\mathfrak{X}}_{\varepsilon,n}$ as an intersection of sets chosen to make the six assertions of the lemma hold,

$$\widetilde{\mathfrak{X}}_{\varepsilon,n} := \widetilde{\mathfrak{X}}_{\Delta,n}\cap\widetilde{\mathfrak{X}}_{\mathbb{Z},n}\cap\widetilde{\mathfrak{X}}_{\eta,n}\cap\widetilde{\mathfrak{X}}_{A,n}\cap\widetilde{\mathfrak{X}}_{\Lambda,n},$$

where the complement of each of the five sets appearing on the right-hand side has probability less than $\varepsilon/5$. More specifically, for a large enough constant C_ε, we define

$$\widetilde{\mathfrak{X}}_{\Delta,n} = \{\|\Delta\| \le C_\varepsilon n^{-1/2}\},$$

$$\widetilde{\mathfrak{X}}_{\mathbb{Z},n} = \Bigl\{\max_{i\le n}\|\mathbb{Z}_i\|^2 \le C_\varepsilon\log n \text{ and } \|\overline{\mathbb{Z}}\| \le C_\varepsilon n^{-1/2}\Bigr\},$$

$$\widetilde{\mathfrak{X}}_{\eta,n} = \Bigl\{\max_{i\le n}|\eta_i|^2 \le C_\varepsilon N\log n \text{ and } \Bigl\|\sum_{i\le n}\eta_i\eta_i'\Bigr\|_2 \le C_\varepsilon n\Bigr\},$$

$$\widetilde{\mathfrak{X}}_{A,n} = \Bigl\{\Bigl\|\sum_{i\le n}\widetilde{\eta}_i\widetilde{\eta}_i'\Bigr\|_2 \le C_\varepsilon n\Bigr\}.$$

The set of $\widetilde{\mathfrak{X}}_{\Lambda,n}$ is defined in a slightly more complicated way. It is defined by requiring various functions of the Λ_k's to be smaller than C_ε times their expected values. Calculate expected values for all the terms in R_1 and R_2 that appear in the bound of Lemma 11.

$$\mathbb{P}_{n,\mu,K}\sum_{k\le p}\Bigl(\sum_{j>p}\Lambda_{k,j}b_j\Bigr)^2 + \mathbb{P}_{n,\mu,K}\sum_{j>p}\Bigl(\sum_{k\le p}\Lambda_{k,j}b_k\Bigr)^2$$

$$= \sum_{k\le p}\sum_{j>p}\mathbb{P}_{n,\mu,K}\Lambda_{k,j}^2(b_j^2+b_k^2) \qquad \text{by Lemma 9 part (i)} \tag{32}$$

$$= O_{\mathcal{F}}(n^{-1}) \sum_{k\le p} \sum_{j>p} k^{-\alpha-2\beta} j^{-\alpha} (\theta_k - \theta_j)^{-2}$$

$$= O_{\mathcal{F}}(n^{-1} p^{1-\alpha}) \qquad \text{by Lemma 7}$$

and

$$\mathbb{P}_{n,\mu,K} \sum_{k\le p} b_k^2 \|\Lambda_k\|^2 k^{2+2\alpha} = O_{\mathcal{F}}(n^{-1}) \sum_{k\le p} k^{4+2\alpha-2\beta}$$

$$= O_{\mathcal{F}}(n^{-1})(1 + p^{5+2\alpha-2\beta} + \log p)$$

and

$$\mathbb{P}_{n,\mu,K} \sum_{k\le p} |b_k| \|\Lambda_k\|^2 = O_{\mathcal{F}}(n^{-1}) \sum_{k\in J} k^{2-\beta} = O_{\mathcal{F}}(n^{-1})(1 + p^{3-\beta} + \log p)$$

and

$$\mathbb{P}_{n,\mu,K} \sum_{k\le p} \|\Lambda_k\|^2 = O_{\mathcal{F}}(n^{-1} p^3)$$

and

$$\mathbb{P}_{n,\mu,K} \sum_{k\le p} \left(\sum_j^* \Lambda_{k,j} b_j \right)^2 = O_{\mathcal{F}}(n^{-1}) \sum_{k\le p} \sum_j^* k^{-\alpha} j^{-a-2\beta} (\theta_k - \theta_j)^{-2}$$

$$= O_{\mathcal{F}}(n^{-1}) \qquad \text{by Lemma 7} \tag{33}$$

and

$$\mathbb{P}_{n,\mu,K} \sum_{k\le p} \|\Lambda_k\|^2 \left(\sum_j^* \frac{|b_j|}{|\theta_k - \theta_j|} \right)^2 = O_{\mathcal{F}}(n^{-1})(p^3 + p^{5+2\alpha-2\beta} \log^2 p) \tag{34}$$

and by Lemma 7

$$\sum_{k\le p} b_k^2 \left(\sum_j^* \frac{1}{|\theta_k - \theta_j|} \right)^2 = O_{\mathcal{F}}(1 + p^{3+2\alpha-2\beta} \log^2 p). \tag{35}$$

For some constant $C_\varepsilon = C_\varepsilon(\mathcal{F})$, on a set $\mathcal{X}_{\Lambda,n}$ with $\mathbb{P}_{n,\mu,K} \mathcal{X}_{\Lambda,n}^c < \varepsilon/5$, each of the random quantities in the previous set of inequalities (for both $p = m$ and $p = N$) is bounded by C_ε times its $\mathbb{P}_{n,\mu,K}$ expected value. By virtue of Lemma 9 part (iii), we may also assume that $\|\Lambda_k\|^2 \le C_\varepsilon k^2/n$ on $\mathcal{X}_{\Lambda,n}$.

We now show that $\sup_{f\in\mathcal{F}} \mathbb{P}_{n,\mu,K} \widetilde{\mathcal{X}}_{\varepsilon,n}^c < \varepsilon$. From the construction of $\widetilde{\mathcal{X}}_{\Lambda,n}$ above, it follows directly that $\mathbb{P}_{n,\mu,K} \widetilde{\mathcal{X}}_{\Lambda,n}^c < \varepsilon/5$.

We analyze $\widetilde{K}$ by rewriting it using the eigenfunctions for K. Then

$$\mathbb{Z}_i(t) - \overline{\mathbb{Z}}(t) = \sum_{j\in\mathbb{N}} (z_{i,j} - z_{\cdot j}) \phi_j(t) = \sum_{j\in\mathbb{N}} \sqrt{\theta_j} (\eta_{i,j} - \eta_{\cdot j}) \phi_j(t)$$

and

$$\widetilde{K}(s,t)=\sum_{j,k\in\mathbb{N}}\widetilde{K}_{j,k}\phi_j(s)\phi_k(t)\qquad\text{with }\widetilde{K}_{j,k}=\sqrt{\theta_j\theta_k}\mathcal{C}_{j,k}. \tag{36}$$

Observe that

$$\begin{aligned}\mathbb{P}\|\Delta\|^2&=\sum_{j,k}\mathbb{P}_{n,\mu,K}\bigl(\widetilde{K}_{j,k}-\theta_j\{j=k\}\bigr)^2=\sum_{j,k}\theta_j\theta_k\mathbb{P}\bigl(\mathcal{C}_{j,k}-\{j=k\}\bigr)^2\\&\le\sum_j\theta_j O_{\mathcal{F}}\bigl(n^{-1}\bigr)+\sum_{j,k}\theta_j\theta_k O_{\mathcal{F}}\bigl(n^{-2}\bigr)=O_{\mathcal{F}}\bigl(n^{-1}\bigr).\end{aligned}$$

Thus, we have $\mathbb{P}_{n,\mu,K}\widetilde{\mathfrak{X}}^c_{\Delta,n}<\varepsilon/5$.

The set $\widetilde{\mathfrak{X}}_{A,n}$ is almost redundant in the sense that $\widetilde{\mathfrak{X}}_{\Delta,n}\subseteq\widetilde{\mathfrak{X}}_{A,n}$ when n and C_ε are large enough. From Definition 1 we know that

$$\min_{1\le j<j'\le N}|\theta_j-\theta_{j'}|\ge(\alpha/R)N^{-1-\alpha}\quad\text{and}\quad\min_{1\le j\le N}\theta_j\ge R^{-1}N^{-\alpha}.$$

The choice $N\asymp n^\zeta$ with $\zeta<(2+2\alpha)^{-1}$ ensures that $n^{1/2}N^{-1-\alpha}\to\infty$. On $\widetilde{\mathfrak{X}}_{\Delta,n}$ the spacing assumption used in Lemmas 10 and 11 holds for all n large enough; all the bounds from those lemmas are available to us on $\widetilde{\mathfrak{X}}_{\varepsilon,n}$. In particular,

$$\max_{j\le N}|\widetilde{\theta}_j/\theta_j-1|\le O_{\mathcal{F}}\bigl(N^\alpha\|\Delta\|\bigr)=o_{\mathcal{F}}(1), \tag{37}$$

where $\widetilde{\theta}_j$'s are defined in (6). Remember that

$$\mathbb{Z}_i(t)-\overline{\mathbb{Z}}(t)=\sum_{k\in\mathbb{N}}(\widetilde{z}_{i,k}-\widetilde{z}_{\cdot k})\widetilde{\phi}_k(t)$$

so that

$$\widetilde{\theta}_k\{j=k\}=\iint\widetilde{K}(s,t)\widetilde{\phi}_j(s)\widetilde{\phi}_k(t)\,ds\,dt=(n-1)^{-1}\sum_{i\le n}(\widetilde{z}_{i,j}-\widetilde{z}_{\cdot j})(\widetilde{z}_{i,k}-\widetilde{z}_{\cdot k}),$$

which implies $(n-1)^{-1}\sum_{i\le n}\widetilde{z}_i\widetilde{z}_i'=\widetilde{D}^2$ and

$$(n-1)^{-1}\sum_{i\le n}\widetilde{\eta}_i\widetilde{\eta}_i'=D^{-1}\widetilde{D}^2D^{-1}:=\operatorname{diag}(1,\widetilde{\theta}_1/\theta_1,\ldots,\widetilde{\theta}_N/\theta_N). \tag{38}$$

Inequality (37) and equality (38) together show that $\widetilde{\mathfrak{X}}_{\Delta,n}\subseteq\widetilde{\mathfrak{X}}_{A,n}$ eventually if we make sure $C_\varepsilon>1$. Thus, $\mathbb{P}_{n,\mu,K}\widetilde{\mathfrak{X}}^c_{A,n}\le\mathbb{P}_{n,\mu,K}\widetilde{\mathfrak{X}}^c_{\Delta,n}<\varepsilon/5$.

As the controls for the set defined in (20) and (21), Lemma 4 controls $\max_{i\le n}\|\mathbb{Z}_i\|^2$ and $\max_{i\le n}|\eta_i|^2$. In addition, we know that

$$\begin{aligned}\mathbb{P}\Bigl\|n^{-1}\sum_{i\le n}\eta_i\eta_i'-I_{N+1}\Bigr\|_2^2&\le\mathbb{P}\Bigl\|n^{-1}\sum_{i\le n}\eta_i\eta_i'-I_{N+1}\Bigr\|_{\mathrm{F}}^2\\&=n^{-1}\sum_{0\le j,k\le N}\operatorname{var}(\eta_{i,k}\eta_{i,j})=O_{\mathcal{F}}\bigl(N^2/n\bigr).\end{aligned}$$

Thus, $\mathbb{P}\|n^{-1}\sum_{i\le n}\eta_i\eta_i'\|_2 = 1 + o_{\mathcal{F}}(1)$. Therefore, we have $\mathbb{P}_{n,\mu,K}\widetilde{\mathcal{X}}^c_{\eta,n} < \varepsilon/5$. To control the $\overline{\mathbb{Z}}$ contribution, note that $n\|\overline{\mathbb{Z}}\|^2$ has the same distribution as $\|\mathbb{Z}_1\|^2$, which has expected value $\sum_{j\in\mathbb{N}}\theta_j < \infty$. Thus, we have $\mathbb{P}_{n,\mu,K}\widetilde{\mathcal{X}}^c_{\mathbb{Z},n} < \varepsilon/5$.

Therefore, there exists $C_\varepsilon > 0$ such that

$$\mathbb{P}_{n,\mu,K}\widetilde{\mathcal{X}}^c_{\varepsilon,n} \le \mathbb{P}_{n,\mu,K}\big(\widetilde{\mathcal{X}}^c_{\Delta,n} + \widetilde{\mathcal{X}}^c_{\mathbb{Z},n} + \widetilde{\mathcal{X}}^c_{\eta,n} + \widetilde{\mathcal{X}}^c_{A,n} + \widetilde{\mathcal{X}}^c_{\Lambda,n}\big) < \varepsilon.$$

6.6. *Proof of the assertions on* $\widetilde{\mathcal{X}}_{\varepsilon,n}$. The assertions (i) and (ii) hold on the set $\widetilde{\mathcal{X}}_{\varepsilon,n}$ as a direct consequence of the construction. From Lemma 11, it follows that on the set $\mathcal{X}_{\Delta,n}\cap\mathcal{X}_{\Lambda,n}$, if $p \le N$,

$$\big\|(\widetilde{H}_p - H_p)\mathbb{B}\big\|^2 = O_{\mathcal{F}}\big(n^{-1}p^{1-\alpha}\big).$$

This inequality leads to the asserted conclusions in (iii) and (iv) when $p = m$ or $p = N$.

Now we show assertion (v) holds on the set $\widetilde{\mathcal{X}}_{\varepsilon,n}$. By construction, $\widetilde{\eta}_{i1} = 1$ for every i, and for $j \ge 2$,

$$\sqrt{\theta_j}\widetilde{\eta}_{i,j} = (\widetilde{z}_{i,j} - \widetilde{z}_{\cdot j}) = \langle \mathbb{Z}_i - \overline{\mathbb{Z}}, \widetilde{\phi}_j\rangle.$$

Thus, for $j \ge 2$,

$$\sigma_j\widetilde{\eta}_{i,j} = \theta_j^{-1/2}\langle \mathbb{Z}_i - \overline{\mathbb{Z}}, \phi_j + f_j\rangle = \eta_{i,j} + \widetilde{\delta}_{i,j}$$

with $\widetilde{\delta}_{i,j}$ satisfying the following bound, due to Lemma 10:

$$|\widetilde{\delta}_{i,j}|^2 \le \theta_j^{-1}\big(\|\mathbb{Z}_i\| + \|\overline{\mathbb{Z}}\|\big)^2\|f_j\|^2 \le O_{\mathcal{F}}\left(\frac{j^{2+\alpha}\log n}{n}\right) \qquad \text{on } \widetilde{\mathcal{X}}_{\varepsilon,n}.$$

In vector form,

$$\widetilde{S}\widetilde{\eta}_i = \eta_i + \widetilde{\delta}_i \qquad \text{with } |\widetilde{\delta}_i|^2 = O_{\mathcal{F}}\left(\frac{N^{3+\alpha}\log n}{n}\right) \le o_{\mathcal{F}}\big(n/N^2\big) \qquad \text{on } \widetilde{\mathcal{X}}_{\varepsilon,n}. \tag{39}$$

It follows that

$$\max_{i\le n}|\widetilde{\eta}_i| = \max_{i\le n}|\widetilde{S}\widetilde{\eta}_i| \le \max_{i\le n}|\eta_i| + o_{\mathcal{F}}(\sqrt{n}/N) = O_{\mathcal{F}}(\sqrt{n}/N) \qquad \text{on } \widetilde{\mathcal{X}}_{\varepsilon,n}.$$

In the end, we show that on $\widetilde{\mathcal{X}}_{\varepsilon,n}$ assertion (vi) holds. From inequality (27) we know that

$$\widetilde{\varepsilon}_N := \max_{i\le n}|\widetilde{\lambda}_{i,N} - \lambda_{i,N}| = O_{\mathcal{F}}\big(n^{-(1+\nu')/2}\big) \qquad \text{on } \widetilde{\mathcal{X}}_{\varepsilon,n},$$

and from bounds (25) and (26) in Section 4.2, we have $\max_{i\le n}|\lambda_{i,N}| = O_{\mathcal{F}}(\sqrt{\log n})$. Assumption ($\ddot{\Psi}$) in Section 2 and the mean-value theorem then give

$$\max_{i\le n}\big|\ddot{\psi}(\widetilde{\lambda}_{i,N}) - \ddot{\psi}(\lambda_{i,N})\big| \le \widetilde{\varepsilon}_N\ddot{\psi}(\lambda_{i,N})G(\widetilde{\varepsilon}_N) = o_{\mathcal{F}}(1).$$

If we replace $\ddot{\psi}(\widetilde{\lambda}_{i,N})$ in the definition of $\widetilde{A}_n$ by $\ddot{\psi}(\lambda_{i,N})$, we make a change

$$\Pi = (n-1)^{-1}\sum_{i\le n}\widetilde{\eta}_i\widetilde{\eta}_i'\big(\ddot{\psi}(\widetilde{\lambda}_{i,N}) - \ddot{\psi}(\lambda_{i,N})\big)$$

with $\|\Pi\|_2 \le o_{\mathcal{F}}(1)\|(n-1)^{-1}\sum_{i\le n}\widetilde{\eta}_i\widetilde{\eta}_i'\|_2$, which, by equality (38), is of order $o_{\mathcal{F}}(1)$ on $\widetilde{\mathfrak{X}}_{\varepsilon,n}$.

From assumption $(\ddot{\Psi})$ we have $d_n := \log\max_{i\le n}\ddot{\psi}(\lambda_{i,N}) = o_{\mathcal{F}}(\log n)$. By triangular inequality and decomposition (39), we have

$$\begin{aligned}\|\widetilde{S}\widetilde{A}_n\widetilde{S} - A_n\|_2 &\le \|\Pi\|_2 + \Big\|(n-1)^{-1}\sum_{i\le n}\ddot{\psi}(\lambda_{i,N})\big(\widetilde{S}\widetilde{\eta}_i\widetilde{\eta}_i'\widetilde{S} - \eta_i\eta_i'\big)\Big\|_2\\
&\le o_{\mathcal{F}}(1) + O_{\mathcal{F}}\big(n^{-1}e^{d_n}\big)\Big\|\sum_{i\le n}\widetilde{\delta}_i\widetilde{\delta}_i'\Big\|_2 \qquad (40)\\
&\quad + O_{\mathcal{F}}\big(n^{-1}\big)\Big\|\sum_{i\le n}\ddot{\psi}(\lambda_{i,N})\big(\widetilde{\delta}_i\eta_i' + \eta_i\widetilde{\delta}_i'\big)\Big\|_2 \qquad \text{on } \widetilde{\mathfrak{X}}_{\varepsilon,n}.\end{aligned}$$

Uniformly over all unit vectors u in $\mathbb{R}^{N+1}$, we have

$$u'\Big(\sum_{i\le n}\widetilde{\delta}_i\widetilde{\delta}_i'\Big)u \le \sum_{i\le n}|\widetilde{\delta}_i|^2 \le n\max_{i\le n}|\widetilde{\delta}_i|^2 = O_{\mathcal{F}}\big(N^{3+\alpha}\log n\big) \qquad \text{on } \widetilde{\mathfrak{X}}_{\varepsilon,n}$$

and by the Cauchy–Schwarz inequality,

$$\begin{aligned}\Big|u'\Big(\sum_{i\le n}\ddot{\psi}(\lambda_{i,N})\big(\widetilde{\delta}_i\eta_i' + \eta_i\widetilde{\delta}_i'\big)\Big)u\Big| &\le O_{\mathcal{F}}\big(n^{1/2}e^{d_n}\big)\max_{i\le n}|\widetilde{\delta}_i|\Big\|\sum_{i\le n}\eta_i\eta_i'\Big\|_2^{1/2}\\
&= O_{\mathcal{F}}\big(e^{d_n}\sqrt{n\log n}N^{(3+\alpha)/2}\big) \qquad \text{on } \widetilde{\mathfrak{X}}_{\varepsilon,n}.\end{aligned}$$

Therefore, the following two bounds hold:

$$\Big\|\sum_{i\le n}\widetilde{\delta}_i\widetilde{\delta}_i'\Big\|_2 = O_{\mathcal{F}}\big(N^{3+\alpha}\log n\big) \qquad \text{on } \widetilde{\mathfrak{X}}_{\varepsilon,n},$$

$$\Big\|\sum_{i\le n}\ddot{\psi}(\lambda_{i,N})\big(\eta_i\widetilde{\delta}_i' + \widetilde{\delta}\eta_i'\big)\Big\|_2 = O_{\mathcal{F}}\big(e^{d_n}\sqrt{n\log n}N^{(3+\alpha)/2}\big) \qquad \text{on } \widetilde{\mathfrak{X}}_{\varepsilon,n}.$$

By plugging into (40), we can obtain that $\|\widetilde{S}\widetilde{A}_n\widetilde{S} - A_n\|_2 = o_{\mathcal{F}}(1)$ on $\widetilde{\mathfrak{X}}_{\varepsilon,n}$.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation in functional regression for general exponential families." (DOI: 10.1214/12-AOS1027SUPP; .pdf). We introduce some useful results in spectral theory and perturbation theory in general Hilbert spaces. They serve as powerful tools that allow us to tackle some of the statistical approximation problems in an elegant way. Some of the results are well-established, while others we believe are new.

REFERENCES

BIRMAN, M. S. and SOLOMJAK, M. Z. (1987). *Spectral Theory of Selfadjoint Operators in Hilbert Space*. Reidel, Dordrecht. Translated from the 1980 Russian original by S. Khrushchëv and V. Peller. MR1192782

BOSQ, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138

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

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

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

CARDOT, H. and SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92** 24–41. MR2102242

CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

DOU, W. W. (2010). Functional regressions for general exponential families: A theoretical and applied study. Ph.D. dissertation, Yale Univ.

DOU, W. W., POLLARD, D. and ZHOU, H. H. (2012). Supplement to "Estimation in functional regression for general exponential families." DOI:10.1214/12-AOS1027SUPP.

EFROMOVICH, S. and KOLTCHINSKII, V. (2001). On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** 2876–2894. MR1872847

ESTRELLA, A. and HARDOUVELIS, G. A. (1991). The term structure as a predictor of real economic activity. *J. Finance* **46** 555–576.

GOLDENSHLUGER, A. and TSYBAKOV, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.* **29** 1601–1619. MR1891740

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

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

HJORT, N. L. and POLLARD, D. (1993). Asymptotics for minimisers of convex processes. Available at arXiv:1107.3806v1 [math.ST].

KATO, T. (1995). *Perturbation Theory for Linear Operators*. Springer, Berlin. Reprint of the 1980 edition. MR1335452

LI, Y. and HSING, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Anal.* **98** 1782–1804. MR2392433

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876

RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. Springer, New York. MR1910407

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

TYLER, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9** 725–736. MR0619278

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447. MR0375592

WRIGHT, J. H. (2006). The yield curve and predicting recessions. Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board at Washington, DC. Available at http://www.federalreserve.gov/pubs/feds/2006.

YU, B. (1997). Assouad, Fano, and Le Cam. In *A Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 423–435. Springer, New York.

STATISTICS DEPARTMENT
YALE UNIVERSITY
24 HILLHOUSE AVE
NEW HAVEN, CONNECTICUT
USA
E-MAIL: Winston.Wei.Dou@aya.yale.edu
David.Pollard@yale.edu
Huibin.Zhou@yale.edu
URL: http://www.stat.yale.edu/