An empirical Bayes shrinkage method for functional data

Joshua Derenski, Yingying Fan, Gareth James, and Min Xu*

October 5, 2022

Abstract

Shrinkage methods have been commonly used in practice and studied, dating back at least as far as the James-Stein shrinkage estimator. The issue of shrinkage arises, albeit with additional complications, for functional data. In this paper we propose an empirical Bayes method to analyze functional data contaminated with noise by constructing posterior mean estimates of the true mean functions. We first reduce the dimension of the functional data by projecting the curves into the finite dimensional space spanned by some prechosen functional basis whose dimension can diverge with sample size. This allows us to explicitly construct an empirical Bayes estimate of the posterior mean by utilizing a multivariate version of Tweedie's formula, converting the original problem to one of estimation of the score function of the basis coefficients resulting from the projection of the curves. For more flexible modeling and efficient estimation, we impose the independent component analysis (ICA) assumption on the projected basis coefficients. This ICA structure enables us to estimate the score function efficiently without making any parametric distribution assumption, yielding our final "Functional Empirical Bayes" (FEmBa) estimate. We formally investigate the theoretical properties of FEmBa and show that it possesses desirable theoretical properties. Furthermore, we demonstrate through extensive simulations and real data analyses that our approach can achieve the desired bias reduction with improved accuracy relative to possible competitors.

Keywords— Tweedie's Formula, Shrinkage Estimation, Functional Data Analysis

1 Introduction

In a *Functional Data Analysis* (FDA) approach one treats an entire curve, or function, $X(t)$ as the unit of observation. Over the last two decades FDA has become a widely used tool in many fields

*The fourth author is supported by the U.S. National Science Foundation under grant DMS-2113671.

ranging from medicine, to the physical and social sciences and even business. Standard statistical tools such as principal components analysis (Silverman, 1996; Goldsmith et al., 2014), regression (James et al., 2009; Fan et al., 2014, 2015), clustering (Abraham et al., 2003; James and Sugar, 2003) and classification (Alonso et al., 2012; Li and Yu, 2008), to mention just a few, have all been extended to functional data settings. See Ramsay et al. (2005) for a more complete summary of these various applications.

Indeed functional data is now becoming so ubiquitous that one often encounters some of the "big data" problems that occur with more traditional scalar observations. In particular here we are interested in the setting where one observes a large number of functions and wishes to select a subset of the most extreme curves. As a concrete example, consider Figure 1 which plots the magnitudes of filtered normalized light emissions from a given star during a fixed time period, with the blue line representing a smoothed fit to the data. This curve is derived from what is called a light curve, which records the apparent brightness of an object in the sky over time. If one is observing a star with an orbiting exoplanet, the relative brightness of the star decreases slightly when the planet passes in front of the host star. Detecting this slight decrease (i.e., dip in Figure 1) helps infer the existence of an exoplanet and derive its properties (such as its size). To date, light curves from hundreds of thousands of stars have been observed, and the most extreme ones (in the sense of having large dips) have been received special attention in the search for exoplanets. One challenge in this application is that there is a substantial amount of preprocessing involved in the process, leading to noisy data. This noise carries over into the process of detecting the most extreme curves. Correcting for such noise can help increase the accuracy of exoplanet detection. In this paper we investigate an empirical Bayes approach for correcting the bias that is introduced in selecting the most extreme among a large number of observed functional curves.

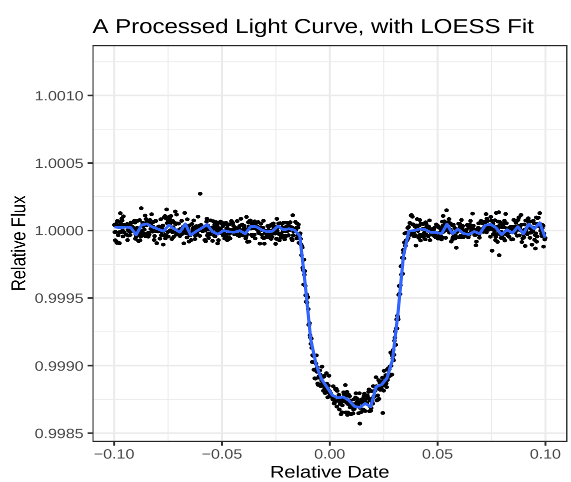

Figure 1: A preprocessed light curve. The transit event is evident.

To clarify this, let us briefly review the empirical Bayes approach for posterior mean estimation in the scalar case. Suppose we have independent observations X_i, $i = 1, \ldots, n$, where $X_i|\mu_i \sim N(\mu_i, \sigma^2)$ with μ_i unobservable and independently generated from some unknown prior distribution $G(\mu)$ and $\sigma^2 > 0$ the deterministic variance. The goal is to detect the most extreme μ_i's based on the observed X_i's. Selection bias refers to the well known property that simply identifying the minimum/maximum values of X_i typically under/over estimates the corresponding values of μ_i. This problem has been extensively studied and numerous approaches have been suggested to address the issue. Most of these methods impose some form of shrinkage with the seminal James-Stein estimator (James and Stein, 1961) being the most well known example. There are several popular classes of methods including; linear shrinkage estimators (Efron and Morris, 1975; Ikeda et al., 2016), non-linear approaches utilizing sparse priors (Abramovich et al., 2006; Bickel and Levina, 2008; Ledoit and Wolf, 2012), Bayesian estimators (Gelman and Shalizi, 2012) and empirical Bayes methods (Jiang and Zhang, 2009; Petrone et al., 2014). Tweedie's formula (Robbins, 1956) is a particularly elegant empirical Bayes approach, which works by directly estimating the marginal distribution of X_i and makes few assumptions about the prior on μ_i. It has been shown to be an effective non-parametric approach for addressing selection bias (Efron, 2011; Zhang, 1997).

To the best of our knowledge, none of these shrinkage approaches have been extended to handle functional data. This is a challenging problem, partly as a result of the infinite dimensional nature of functional data, but also because there are many ways that functional data may be considered "extreme." For example, a curve may be selected based on its maximum value, its average value, or some other functional of the curve. We propose a general approach called "Functional Empirical Bayes" (FEmBa), which extends Tweedie's formula to the functional setting, to reduce the selection bias by denoising each functional curve.

To reduce the infinite dimensionality of functional curves, we propose to project them onto the space spanned by some pre-chosen functional basis whose dimension K is finite but can slowly diverge with the sample size. Then our working data become the basis coefficient vectors in $\mathbb{R}^K$, denoted as $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, which result from the projection. The posterior means of the projected true mean functions, conditional on the $\boldsymbol{\theta}_i$'s, can then be explicitly derived using the multivariate version of Tweedie's formula. However, despite the explicit form, empirically estimating such posterior means is challenging because it requires estimating the multidimensional score function $\mathbf{v}_0(\boldsymbol{\theta}_i)$ of the nonparametric multivariate density of $\boldsymbol{\theta}_i$. We address this challenge by imposing the independent component analysis (ICA) assumption on the $\boldsymbol{\theta}_i$'s, which reduces the original problem to the estimation of the unmixing matrix W_0 and the corresponding univariate score functions $u_{0k}(Z_k)$ of the unmixed independent components Z_k, in an analogous fashion to multiple index models for regression. This is particularly valuable in the functional data setting as the dimensionality K of the working basis elements must increase with the sample size n to ensure that all curves can be approximated uniformly well. Due to the complicated nature of our problem, the practical estimations of W_0 and $\boldsymbol{u}_0 = (u_{01}, \ldots, u_{0K})^\top$ are difficult. We further propose a joint estimation approach for both W_0 and $\boldsymbol{u}_0$ by directly minimizing the risk function, where the geodesic gradient descent method in Plumbley (2005) (for ICA estimation) is adapted for our purpose to estimate W_0 and $\boldsymbol{u}_0$. This allows us to construct the plug-in estimate of the original score function $\mathbf{v}_0(\boldsymbol{\theta}_i)$, leading to our final FEmBa estimate of the projected posterior mean function.

To theoretically justify the performance of FEmBa, we establish both in-sample and out-of-sample risk bounds for the FEmBa estimate compared to the oracle posterior mean estimate, under the assumption that the score function $\mathbf{v}_0(\boldsymbol{\theta})$ is known. Our theoretical analyses also provide standalone contributions to the ICA literature. We prove, under a sub-Weibull tail assumption on the $\boldsymbol{\theta}_i$'s and an m-degree smoothness assumption on the true score function, that the risk of our score function estimator converges to 0 at a rate of $O(\frac{K^{3+\frac{1}{m}} \log^2 K}{n})^{2m/2m+3}\text{polylog}(n)$. This result shows that we may choose K as high as n^α for $\alpha < \frac{m}{3m+1}$ and still consistently estimate the optimal shrinkage function. To the best of our knowledge, this is the first rate of convergence result for nonparametric estimation in an ICA model framework, as previous analyses of ICA either only focus on parametric estimation of the unmixing matrix (Chen and Bickel, 2006) or only prove consistency (Samworth and Yuan, 2012). Our proof uses techniques from empirical process theory as well as a new analysis of the risk function showing that it is locally strongly convex around 0.

Our paper is laid out as follows. In Section 2 we present the FEmBa framework. Section 3 proposes a criterion for score function estimation under an ICA structure, resulting in an approach for joint estimation of the needed unmixing matrix and the desired score function. We also provide theoretical justification for the FEmBa estimator by establishing both in-sample and out-of-sample risk bounds. Section 4 discusses the practical implementation of FEmBa. Finite sample simulation comparisons of FEmBa with several other estimators are provided in Section 5 and an application of FEmBa to the planetary data is discussed in Section 6. All technical proofs, detailed description of algorithm and additional simulation results are included in a separate supplementary file.

2 FEmBa Model

In this paper we focus on fully observed curves. We examine the case of discretely but densely observed curves in simulation studies in Section 5. Denote independent and identically distributed (iid) curves observed at all time points as $X_1(t), \dots, X_n(t)$ where $t \in [0,1]$. We model these functions via

$$X_i(t) = m_{*i}(t) + g_i(t), \quad i = 1, \dots n, \tag{1}$$

where the $m_{*i}(t)$'s are the underlying mean curves, generated from an unknown prior distribution, and the $g_i(t)$'s represent noise curves, sampled from a mean zero Gaussian process, independently from the m_{*i}'s. To simplify notation we assume the functional curves are demeaned so that $\mathbb{E}[X_i(t)] = 0$ for all t and all $i = 1, \dots, n$. However, in practice our method, and the theoretical results, can be applied equally well without this assumption.

Our goal is to estimate the mean curves $m_{*i}(t)$ using the observed curves $X_i(t)$. Such a problem is not always feasible because the $m_{*i}(t)$'s are not identifiable without additional assumptions. However, in the case where the distribution of the noise is known, estimation of $m_{*i}(t)$ is possible. Here we discuss our proposed method assuming the covariance function of the $g_i(t)$'s, denoted $\Gamma(s,t) = \text{Cov}(g_i(s), g_i(t))$, is known. This assumption is inherited from Tweedie's formula (Robbins, 1956), without which the estimation of $m_{*i}(t)$ from $X_i(t)$ is impossible. In applications such as the exoplanet light curves discussed in Section 6, $\Gamma(s,t)$ can be well estimated because repeated measurements for a subset of curves are available.

We select our mean curve estimator by minimizing the loss function:

$$\mathbb{E}\|m_{*i} - m_i\|^2, \tag{2}$$

where $m_i = \mathcal{T}(X_i)$ is some estimator for $m_{*i}(t)$ constructed from the observed data $X_i(t)$ through a common operator $\mathcal{T}$, $\|g\|^2 = \int g(t)^2 dt$ for any square integrable function $g(t)$, and the expectation is taken with respect to both $m_{*i}(t)$ and $X_i(t)$. Here, both g and $g(t)$ represent the same function and we will use this convention throughout the paper. Under some mild conditions, it can be shown that the conditional expectation $\mathbb{E}(m_{*i}(t)|X_i(t))$ minimizes (2) among the class of estimators $\mathcal{T}(X_i)$. However, since the distribution of $m_{*i}(t)$ is generally unavailable to us, $\mathbb{E}(m_{*i}(t)|X_i(t))$ cannot be directly calculated. Moreover, since the curves $X_i(t)$ are infinite dimensional, we must impose some kind of dimension reduction in order to estimate $m_{*i}(t)$ efficiently.

2.1 Dimension Reduction through Projection

To reduce the dimensionality, we first project the curves $X_i(t)$'s onto a finite dimensional basis–a standard approach in Functional Data Analysis. Let $\mathbf{s}(t) = (s_1(t), \ldots, s_K(t))^\top \in \mathbb{R}^K$ represent any basis that we choose for representing our curves, where K can slowly diverge with sample size n. We then approximate our curves with the following:

$$X_i^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \boldsymbol{\theta}_i^{\mathbf{s}}, \tag{3}$$

where $\boldsymbol{\theta}_i^{\mathbf{s}} = (\theta_{i1}^{\mathbf{s}}, \ldots, \theta_{iK}^{\mathbf{s}})^\top = \Sigma_{\mathbf{s}}^{-1} \int X_i(t)\mathbf{s}(t)dt$ with $\Sigma_{\mathbf{s}} = \int \mathbf{s}(t)\mathbf{s}(t)^\top dt$ is the basis coefficient vector corresponding to $\mathbf{s}$. With such dimension reduction, we can treat the $\boldsymbol{\theta}_i^{\mathbf{s}}$'s as the working data. Correspondingly, instead of minimizing (2) over the entire functional space, we propose searching for the best estimate of $m_{*i}(t)$ within the class of all functions in the space spanned by $\mathbf{s}(t)$, that is, our final estimate will take the form $\mathbf{s}(t)^\top \boldsymbol{\mu}$ with $\boldsymbol{\mu} \in \mathbb{R}^K$. The problem in (2) is then reduced to minimizing the following risk

$$\mathbb{E} \int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu})^2 dt \tag{4}$$

with respect to $\boldsymbol{\mu} = t^*(\boldsymbol{\theta}_i^{\mathbf{s}})$, where $t^* : \mathbb{R}^K \to \mathbb{R}^K$ is some Borel measurable function.

Standard calculations show that, for any $\boldsymbol{\mu} \in \mathbb{R}^K$,

$$\mathbb{E} \int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu})^2 dt = \mathbb{E}\left((\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \boldsymbol{\mu})^\top \Sigma_{\mathbf{s}} (\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \boldsymbol{\mu}) \right) + \mathbb{E} \int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}})^2 dt, \tag{5}$$

where $\boldsymbol{\mu}_{*i}^{\mathbf{s}} := \Sigma_{\mathbf{s}}^{-1} \int m_{*i}(t)\mathbf{s}(t)dt$ is the coefficient vector corresponding to $m_{*i}(t)$'s projection onto the basis $\mathbf{s}(t)$, denoted $m_{*i}^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}}$. Here, since K is allowed to diverge, under some mild assumptions, the last term in (5) is expected to vanish as K diverges. Thus, for a given $\boldsymbol{\theta}_i^{\mathbf{s}}$, our original loss criterion (2) can be approximately optimized by minimizing (5) with respect to $\boldsymbol{\mu}$.

The following proposition characterizes the minimizer of the risk function. To simplify the notation, we suppress the dependence of various quantities on $\mathbf{s}$ whenever there is no confusion. For example, below we write $\boldsymbol{\theta}_i^{\mathbf{s}}$ as $\boldsymbol{\theta}_i$, and $\boldsymbol{\mu}_{*i}^{\mathbf{s}}$ as $\boldsymbol{\mu}_{*i}$.

Proposition 1. *The minimizer of* (4) *is* $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i} = \mathbb{E}[\boldsymbol{\mu}_{*i}|\boldsymbol{\theta}_i]$.

Denote by $\tilde{m}_{X_i}^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}$ the corresponding functional curve, which is indeed the population target that FEmBa aims to estimate. Here, $\tilde{m}_{X_i}^{\mathbf{s}}$ only minimizes our objective function (4), but, as we demonstrate in later sections, it also allows us to reduce the selection bias for extreme functions. We observe from Proposition 1 that if $m_{*i}(t)$ lies exactly in the space spanned by $\mathbf{s}(t)$ then the term $\mathbb{E}\|m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}\|^2$ in (5) disappears and the estimation error of $\tilde{m}_{X_i}^{\mathbf{s}}$ is completely characterized by the estimation error between the posterior mean estimate $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}$, and $\boldsymbol{\mu}_{*i}$. The term $\mathbb{E}\|m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}\|^2$ can be viewed as the fixed cost of dimensionality reduction. Table 1 further

	Observed Data	True Mean	Posterior Mean
Functional Space	$X_i(t)$	$m_{*i}(t)$	$\tilde{m}_{X_i}(t) = \mathbb{E}[m_{*i}(t)\|X_i(t)]$
Basis Space	$X_i^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \boldsymbol{\theta}_i^{\mathbf{s}}$	$m_{*i}^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}}$	$\tilde{m}_{X_i}^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}^{\mathbf{s}}$
Basis Coefficients	$\boldsymbol{\theta}_i^{\mathbf{s}}$	$\boldsymbol{\mu}_{*i}^{\mathbf{s}}$	$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}^{\mathbf{s}} = \mathbb{E}[\boldsymbol{\mu}_{*i}^{\mathbf{s}}\|\boldsymbol{\theta}_i^{\mathbf{s}}]$

Table 1: Explanation of notation used for the observed data, true mean curve and corresponding conditional expectation. The basis expansion is with respect to some prechosen $\mathbf{s}(t)$.

explains our notation and the relationships among the various quantities discussed in this section. For the rest of the paper, we discuss how to estimate $\tilde{m}_{X_i}^{\mathbf{s}}(t)$.

2.2 Tweedie's Formula on Functional Data

We see from Proposition 1 that the key to constructing $\tilde{m}_{X_i}^{\mathbf{s}}$ is calculating the posterior mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}$, which generally requires knowledge of the joint distribution of $\boldsymbol{\mu}_{*i}$ and $\boldsymbol{\theta}_i$. However, in some special cases, the problem can be simplified. We next demonstrate that Tweedie's formula can be extended to our functional setting to help solve the problem.

Proposition 2. *For data generated according to Model* (1)*, it holds that*

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i} = \mathbb{E}(\boldsymbol{\mu}_{*i}|\boldsymbol{\theta}_i) = \boldsymbol{\theta}_i + \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i), \tag{6}$$

$$\tilde{m}_{X_i}^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}, \tag{7}$$

where $\mathbf{v}_0(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta})$ *and* $f(\boldsymbol{\theta})$ *are respectively the score function and marginal density of* $\boldsymbol{\theta}$*, and* $\Sigma_{\boldsymbol{\gamma}} = \int\int \Gamma(s,t)\mathbf{s}(s)\mathbf{s}(t)^T dtds$*.*

Furthermore, Theorem 1 below provides an explicit formula for the reduction in risk from using $\tilde{m}_{X_i}^{\mathbf{s}}(t)$ over the natural, but somewhat naive, estimator $X_i^{\mathbf{s}}(t)$ in (3).

Theorem 1. *It holds that*

$$\mathbb{E}\|m_{*i} - \tilde{m}_{X_i}^{\mathbf{s}}\|^2 = \mathbb{E}\|m_{*i} - X_i^{\mathbf{s}}\|^2 - \mathbb{E}\|\Sigma_{\mathbf{s}}^{1/2}\Sigma_{\boldsymbol{\gamma}}\mathbf{v}_0(\boldsymbol{\theta}_i)\|^2.$$

Equations (6) and (7) suggest a two step approach, which FEmBa adopts for estimating the projected posterior mean curves $\tilde{m}_{X_i}^{\mathbf{s}}$. FEmBa first estimates the $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i}$'s according to (6), and then constructs curve estimates via (7). The second step is trivial once the first step is completed. To conclude this section, we note that the choice of basis is not unique. In fact, under some conditions, all the theoretical results in Section 3 remain true for any basis constructed independently from the training data.

2.3 Curve Estimation via Risk Minimization

Motivated by Proposition 2, we estimate $\tilde{m}_{X_i}^{\mathbf{s}}(t)$ via risk minimization. Let $\check{m}_{X_i}^{\mathbf{s}}(t) = \mathbf{s}(t)^\top(\boldsymbol{\theta}_i + \Sigma_{\boldsymbol{\gamma}}\check{\mathbf{v}}(\boldsymbol{\theta}_i))$ be a candidate estimate for $\tilde{m}_{X_i}^{\mathbf{s}}(t)$ with $\check{\mathbf{v}}(\boldsymbol{\theta}_i) : \mathbb{R}^K \to \mathbb{R}^K$ some multivariate multiresponse function. To derive the optimal estimator, we consider the following loss function:

$$\int (\tilde{m}^{\mathbf{s}}_{X_i}(t) - \check{m}^{\mathbf{s}}_{X_i}(t))^2 dt = (\check{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i))^\top \tilde{\Sigma} (\check{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i)), \tag{8}$$

where $\tilde{\Sigma} = \Sigma_{\boldsymbol{\gamma}} \int \mathbf{s}(t)\mathbf{s}(t)^\top dt \Sigma_{\boldsymbol{\gamma}}$. Notice that the loss function is now written entirely in terms of the score function we wish to estimate. Thus, deriving an estimator for $\tilde{m}^{\mathbf{s}}_{X_i}(t)$ is equivalent to deriving an estimator for $\mathbf{v}_0(\boldsymbol{\theta}_i)$. We thus wish to minimize the following risk function:

$$\begin{aligned} R(\check{\mathbf{v}}) &= \mathbb{E}\left((\check{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i))^\top \tilde{\Sigma} (\check{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i)) \right) \\ &\propto \mathbb{E}\left(\check{\mathbf{v}}(\boldsymbol{\theta}_i)^\top \tilde{\Sigma} \check{\mathbf{v}}(\boldsymbol{\theta}_i) \right) - 2\mathbb{E}\left(\mathbf{v}_0(\boldsymbol{\theta}_i)^\top \tilde{\Sigma} \check{\mathbf{v}}(\boldsymbol{\theta}_i) \right) \\ &= \mathbb{E}\left(\check{\mathbf{v}}(\boldsymbol{\theta}_i)^\top \tilde{\Sigma} \check{\mathbf{v}}(\boldsymbol{\theta}_i) \right) - 2\sum_{k,l} a_{kl} \mathbb{E}\left(\check{v}_k(\boldsymbol{\theta}_i) v_{0l}(\boldsymbol{\theta}_i) \right), \end{aligned} \tag{9}$$

where in the second step above we have dropped terms that do not depend on $\mathbf{v}_0(\boldsymbol{\theta}_i)$. Here, $a_{kl} = \tilde{\Sigma}_{kl}$, and $\check{v}_k(\boldsymbol{\theta}_i)$ and $v_{0l}(\boldsymbol{\theta}_i)$ are the k-th and l-th components of $\check{\mathbf{v}}(\boldsymbol{\theta}_i)$ and $\mathbf{v}_0(\boldsymbol{\theta}_i)$, respectively. Let us now analyze $\mathbb{E}\left(\check{v}_k(\boldsymbol{\theta}_i) v_{0l}(\boldsymbol{\theta}_i) \right)$, for a fixed k and l. We will confine ourselves to bounded $\check{v}_k$'s. Then under a mild condition on the density function of $\boldsymbol{\theta}_i$, where $\lim_{\|\boldsymbol{\theta}_i\|_\infty \to \infty} f(\boldsymbol{\theta}_i) = 0$, we have that

$$\mathbb{E}\left[\check{v}_k(\boldsymbol{\theta}_i) v_l(\boldsymbol{\theta}_i) \right] = \int \check{v}_k(\boldsymbol{\theta}_i) \partial_l f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = -\int f(\boldsymbol{\theta}_i) \partial_l \check{v}_k(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = -\mathbb{E}[\partial_l \check{v}_k(\boldsymbol{\theta}_i)].$$

Plugging the above expression into (9), we have that

$$R(\check{\mathbf{v}}) \propto \mathbb{E}\left(\check{\mathbf{v}}(\boldsymbol{\theta}_i)^\top \tilde{\Sigma} \check{\mathbf{v}}(\boldsymbol{\theta}_i) \right) + 2\mathbb{E}\left(\mathbf{1}^\top (\tilde{\Sigma} \circ J\check{\mathbf{v}}(\boldsymbol{\theta}_i)) \mathbf{1} \right), \tag{10}$$

where $J\check{\mathbf{v}}(\boldsymbol{\theta}_i)$ is the Jacobian matrix of $\check{\mathbf{v}}$ at value $\boldsymbol{\theta}_i$ and where $\tilde{\Sigma} \circ J\check{\mathbf{v}}(\boldsymbol{\theta}_i)$ denotes the element-wise (Hadamard) product of these two matrices.

Minimizing (10) directly with respect to $\check{\mathbf{v}}$ can be challenging, especially when K is large. Next, we present an ICA framework for $\boldsymbol{\theta}_i$ which can help simply both the theoretical derivations and the development of the estimation algorithm.

2.4 A Functional ICA Assumption

To motivate utilizing the ICA model framework, we present the following lemma.

Lemma 1. *Suppose there exists a functional basis $\{s^*_j(t)\}_{j=1}^\infty$ such that the functional representation of $X_i(t) = \sum_{j=1}^\infty s^*_j(t)\theta^{\mathbf{s}^*}_{ij}$ satisfies that $\theta^{\mathbf{s}^*}_{ij} \perp\!\!\!\perp \theta^{\mathbf{s}^*}_{il}$ for all $j \neq l$. Assume that $\lim_{K\to\infty} \sum_{j>K} |\theta^{\mathbf{s}^*}_{ij}| = 0$ almost surely, and that $\sup_{j\geq 1} \int s^*_j(t)^2 dt < \infty$. For the chosen K-dimensional working basis $\mathbf{s}(t)$, assume that $\max_{1\leq i\leq K} \mathbf{e}_i^\top \Sigma_{\mathbf{s}}^{-1} \mathbf{e}_i$ is uniformly bounded for all K large enough, where recall $\Sigma_{\mathbf{s}} = \int \mathbf{s}(t)\mathbf{s}(t)^\top dt$. Then the basis coefficient vector $\boldsymbol{\theta}_i$ corresponding to $\mathbf{s}(t)$ satisfies that*

$$\boldsymbol{\theta}_i = \sum_{k=1}^K \theta^{\mathbf{s}^*}_{ik} \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) s^*_k(t) dt + o_{o.s.}(1), \tag{11}$$

for all K sufficiently large.

Since $\theta_{ik}^{\mathbf{s}^*}$'s are independent of each other across k, it is seen that $\boldsymbol{\theta}_i$ has an approximate ICA structure. This motivates us making the following assumptions on $\boldsymbol{\theta}_i$.

Condition 1. *Assume that the basis coefficient $\boldsymbol{\theta}_i$ follows the following ICA structure*

$$\boldsymbol{\theta}_i = W_0^{-1}\boldsymbol{z}_i, \tag{12}$$

where $\boldsymbol{z}_i = (Z_{i1}, \dots, Z_{iK})^\top$ is a K dimensional vector with independent mean-zero components, and $W_0 \in \mathbb{R}^{K\times K}$ is the invertible unmixing matrix. We also assume without loss of generality that Z_{ik} has unit variance for all k.

Condition 1 imposes a stronger assumption (12) than the result in Lemma 1. In particular, the $o_{a.s.}(1)$ term therein is ignored. Since K is allowed to diverge with sample size n, Condition 1 is not overly stringent; in fact, by imposing the exact ICA structure, the theory and algorithm can be greatly simplified. The robustness of our method with respect to misspecified ICA structure will be investigated in the numerical sections.

To simplify the presentation, we use $\boldsymbol{\theta}$ and $\boldsymbol{z} = (Z_1, \dots, Z_K)^\top$ to denote generic random variables which are identically distributed to the $\boldsymbol{\theta}_i$'s and $\boldsymbol{z}_i$'s, respectively. The ICA structure ensures that $\mathbf{v}_0(\boldsymbol{\theta}) = W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})$, where $\boldsymbol{u}_0(\cdot) : \mathbb{R}^K \to \mathbb{R}^K$ is the score function of random vector $\boldsymbol{z}$. Since $\boldsymbol{z}$ has independent components, it follows that the ith coordinate of $\boldsymbol{u}_0(\cdot)$ is a univariate function depending only on Z_k for $k = 1, \dots, K$. Denote by $\boldsymbol{u}_0(\boldsymbol{z}) = (u_{0,1}(Z_1), \dots, u_{0,K}(Z_K))^\top$. The ICA assumption reduces the estimation of $\mathbf{v}_0$ to those of W_0 and $\boldsymbol{u}_0$.

In general, ICA models are identifiable only up to a signed permutation. Indeed, since $\boldsymbol{\theta} = W_0^{-1}\boldsymbol{z}$, we can apply the same signed permutation to the components of $\boldsymbol{z}$ and to W_0 without changing the distribution of $\boldsymbol{\theta}$. This is the only source of non-identifiability if and only if at most one component of $\boldsymbol{z}$ is Gaussian. Since any isotropic Gaussian distribution is rotationally invariant, it is easy to see that the condition is necessary. That the condition is also sufficient follows from the fact that the isotropic Gaussian is the only multivariate product distribution that is also rotationally invariant; this result, often called Maxwell's Theorem, originates from James Clerk Maxwell's investigation of gas particles in $\mathbb{R}^3$, see (Feller, 1966, Chapter III.4).

Identifiability is not essential for curve estimation since we are only interested in estimating the score function $\mathbf{v}_0 = W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})$, but it plays an important role in our theoretical analysis.

Under the assumption that $\mathrm{Cov}(\boldsymbol{z}) = I_K$, we see that $W_0\Sigma_{\boldsymbol{\theta}}W_0^\top = I_K$ where $\Sigma_{\boldsymbol{\theta}} = \mathbb{E}\boldsymbol{\theta}\boldsymbol{\theta}^\top$ is the covariance of $\boldsymbol{\theta}$, or, equivalently, $W_0 = U_0\Sigma_{\boldsymbol{\theta}}^{-1/2}$ for some orthogonal matrix U_0. Since $\boldsymbol{z}$ has independent components, we can write $\boldsymbol{z} \mapsto p(\boldsymbol{z}) = \prod_{k=1}^K p_j(Z_j)$ as the density of $\boldsymbol{z}$. Recall that $\boldsymbol{u}_0$ and $\mathbf{v}_0$ are the score functions of $\boldsymbol{\theta}$ and $\boldsymbol{z}$, respectively. By a change of variables, we have that

$$\mathbf{v}_0(\boldsymbol{\theta}) = W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}) \tag{13}$$

$$J\mathbf{v}_0(\boldsymbol{\theta}) = W_0^\top J\boldsymbol{u}_0(W_0^\top\boldsymbol{\theta})W_0, \tag{14}$$

where $J\mathbf{v}_0$ and $J\boldsymbol{u}_0$ are the Jacobian matrices of $\mathbf{v}_0$ and $\boldsymbol{u}_0$, respectively.

The risk function to minimize (c.f. (10)) then can be reparametrized as

$$R(W,\boldsymbol{u}) = \mathbb{E}\big(W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\big)^\top \tilde{\Sigma}\big(W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\big)$$
$$\propto \mathbb{E}(\boldsymbol{u}(W\boldsymbol{\theta})^\top W\tilde{\Sigma}W^\top \boldsymbol{u}(W\boldsymbol{\theta})) + 2\sum_{k=1}^K c(W_{k\cdot})\mathbb{E}(u_k'(W_{k\cdot}^\top \boldsymbol{\theta})) \tag{15}$$

where $W \in \mathbb{R}^{K\times K}$ satisfies $W\Sigma_{\boldsymbol{\theta}}W^\top = I_K$ (or, equivalently, $W = U\Sigma_{\boldsymbol{\theta}}^{-1/2}$ for some orthogonal matrix U), $\boldsymbol{u}(\boldsymbol{z}) = (u_1(z_1),\ldots,u_K(Z_K))^\top$ is the score function of some product density on $\mathbb{R}^K$, $W_{k\cdot}$ denotes the k-th row of the matrix W, and $c(W_{k\cdot}) = W_{k\cdot}^\top \tilde{\Sigma} W_{k\cdot} \geq 0$.

Given n curves and their basis coefficient vectors $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n$, we estimate W_0 and $\boldsymbol{u}_0$ by minimizing the empirical counterpart of (15). To avoid overfitting, we impose constraints on the higher derivatives of the component functions of $u_1(\cdot),\ldots,u_K(\cdot)$. This yields the following empirical risk minimization problem: for $m \geq 2$,

$$\min_{\boldsymbol{u},W} \frac{1}{n}\sum_{i=1}^n \Big\{ \boldsymbol{u}(W\boldsymbol{\theta}_i)^\top W\tilde{\Sigma}W^\top \boldsymbol{u}(W\boldsymbol{\theta}_i) + 2\sum_{k=1}^K c(W_{k\cdot})u_k'(W_{k\cdot}^\top\boldsymbol{\theta}_i)\Big\} \tag{16}$$
$$\text{s.t.} \quad \max_{1\leq k\leq K} \int |u_k^{(m)}(t)|^2\,dt \leq B,$$

where $B \geq 0$ measures the amount of regularization we impose to avoid overfitting.

Our estimator is a solution to (16) over univariate functions $u_1,\ldots,u_K$ as well as matrix W which can be written as $W = U\Sigma_{\boldsymbol{\theta}}^{-1/2}$ for an orthogonal matrix U. An orthogonal matrix U has determinant of either 1 or -1; as we will see, it is convenient to restrict our attention to U such that $\det(U) = 1$. We refer to this class of matrices as special orthogonal matrices and denote it

$$SO(K) := \{U \in \mathbb{R}^{K\times K} : U \text{ orthogonal and } \det(U) = 1\}.$$

We can restrict U to $SO(K)$ without loss of generality as, for any U with a -1 determinant, we can flip the sign of any of its columns to obtain an orthogonal matrix in $SO(K)$. For simplicity of presentation, we hitherto use the term orthogonal matrix to refer to $SO(K)$ unless otherwise stated. The objective function in (16) is not convex, but in Section 4, we provide an alternating descent optimization algorithm which works well in practice.

Given the solution of optimization problem (16), the score function $\mathbf{v}_0$ in (13) can be estimated as $\hat{\mathbf{v}}(\boldsymbol{\theta}) = \hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta})$. Consequently, our final FEmBa estimate for $\tilde{m}_{X_i}^s$ takes the form

$$\hat{m}_{X_i}(t) = \mathbf{s}(t)^\top(\boldsymbol{\theta}_i + \Sigma_{\boldsymbol{\gamma}}\hat{\mathbf{v}}(\boldsymbol{\theta}_i)). \tag{17}$$

3 Finite sample properties of the FEmBa Estimator

We analyze the finite sample properties of the estimator proposed in the last section. As a shorthand, for any vector $\boldsymbol{v} \in \mathbb{R}^K$, we write $\|\boldsymbol{v}\|_{\tilde{\Sigma}}^2 := \boldsymbol{v}^\top\tilde{\Sigma}\boldsymbol{v}$ as the Mahalanobis squared norm with respect to $\tilde{\Sigma}$. Using this notation, the risk function (15) has the simplified form

$$R(W, \boldsymbol{u}) = \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_{\tilde{\Sigma}}^2 \tag{18}$$

$$= \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_{\tilde{\Sigma}}^2 + 2\sum_{k=1}^{K} \|W_{k\cdot}\|_{\tilde{\Sigma}}^2\, \mathbb{E}u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_{\tilde{\Sigma}}^2. \tag{19}$$

Define the empirical risk as

$$\hat{F}(W, \boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \|W^\top \boldsymbol{u}(W\boldsymbol{\theta}_i)\|_{\tilde{\Sigma}}^2 + 2\sum_{k=1}^{K} \|W_{k\cdot}\|_{\tilde{\Sigma}}^2 \frac{1}{n}\sum_{i=1}^{n} u_k'(W_{k\cdot}^\top \boldsymbol{\theta}_i).$$

With these definitions, our optimization (16) can be written as:

$$(\hat{W}, \hat{\boldsymbol{u}}) := \underset{W, u_1, \ldots u_K}{\operatorname{argmin}}\ \hat{F}(W, \boldsymbol{u})$$
$$\text{s.t.}\quad W \in \mathcal{W} \text{ and } u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}, \tag{20}$$

where we move the higher derivative constraint to the constraint set

$$\mathcal{F}_{b,B,m} := \left\{ f\ :\ f = 0 \text{ on } [-b, b]^c,\ f^{(m)} \text{ exists on } \mathbb{R} \text{ and } \int_{-b}^{b} |f^{(m)}|^2 \leq B \right\}, \tag{21}$$
$$\mathcal{W} := \{W \in \mathbb{R}^{K\times K}\ :\ W = U\Sigma_{\boldsymbol{\theta}}^{1/2},\ U \in SO(K)\}.$$

Here, the restriction that $u_k(\cdot)$ be zero outside of an interval $[-b, b]$ is not severe since we can let $u_k(\cdot)$ take on any value outside of the range of the data points without affecting the objective. When the unmixed components $Z_1, \ldots, Z_K$ have sub-Weibull tails, we take b to be some power of $\log n$ (the large deviation bound of $\max_i |Z_i|$).

Informally speaking, our main result for this section (Theorem 2) states that, under tail assumptions on $\boldsymbol{z} = W_0\boldsymbol{\theta}$, the risk of our estimator $R(\hat{W}, \hat{\boldsymbol{u}})$ tends to 0 at a rate of $(\frac{K^{3+\frac{1}{m}} \log^2 K}{n})^{\frac{2m}{2m+3}}$ ignoring additional poly-log terms.

Remark 1. *The estimator defined in* (20) *assumes knowledge of the true covariance matrix* $\Sigma_{\boldsymbol{\theta}}$*. In the more realistic scenario where* $\Sigma_{\boldsymbol{\theta}}$ *is estimated using the empirical covariance matrix* $\hat{\Sigma}_{\boldsymbol{\theta}}$*, our estimation procedure would be to minimize* $\hat{F}(W, \boldsymbol{u})$ *over* $W = V\hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2}$ *for* $V \in SO(K)$ *where we use the empirical covariance* $\hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2}$ *instead of the true covariance* $\Sigma_{\boldsymbol{\theta}}^{-1/2}$*.*

Provided that $\|\hat{\Sigma}_{\boldsymbol{\theta}} - \Sigma_{\boldsymbol{\theta}}\|_2^2 = O_p(K/n)$*, our estimator would still attain the rate given in Theorem 2. To avoid introducing complicated new notation, we omit a formal proof in the estimated* $\hat{\Sigma}_{\boldsymbol{\theta}}$ *case and instead give a detailed description how to adapt the proof for the known* $\Sigma_{\boldsymbol{\theta}}$ *case in Section S4.8.*

3.1 Assumptions

Theoretical analysis of nonparametric density estimation typically assumes that the true underlying density is compactly supported. Since this is unrealistic in a functional data analysis setting, we use a truncation argument to allow the density of $\boldsymbol{\theta}$ to be supported on $\mathbb{R}^K$. Most of our assumptions relate to the tail behavior of the unmixed coefficients $\boldsymbol{z} = W_0\boldsymbol{\theta}$ (and thus also of $\boldsymbol{\theta}$) and are used only in the truncation argument.

Before stating the assumptions, we need the following definition.

Definition 1. *Let $\alpha, \tilde{C}, \sigma > 0$. We say that a random variable X with mean-zero and unit variance is $(\alpha, \tilde{C}, \sigma)$-sub-Weibull if for all $t > 0$, we have*

$$\mathbb{P}(|X| \geq t) \leq \tilde{C} \exp\big(-(t/\sigma)^\alpha\big).$$

We note that if $b = 2\sigma(3\log n)^{1/\alpha}$ and if X is $(\alpha, \tilde{C}, \sigma)$-sub-Weibull, then $\mathbb{P}(|X| > \frac{b}{2}) \leq \frac{\tilde{C}}{n^3}$.

Condition 2. *We make the following assumptions, which are stated in terms of non-negative quantities $c_0, R_0, R_1, \delta_0, c_1, M$.*

A1 Write $c_0 := \max_{k\in[K], j\in[m]} |u_{0k}^{(j)}(0)|$. We assume that B is chosen large enough so that $c_0 \leq \frac{B^{1/2}}{2^{m+1}}$ and that $\int_{-b}^{b} |u_{0k}^{(m)}|^2 \leq \frac{B}{2^{2(m+1)}c_m^2}$ where c_m is a constant depending only on m.

A2 Assume there exists $\delta_0 \geq \frac{1}{2}$ such that $R_0 := \{\mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_{\tilde{\Sigma}}^{2+2\delta_0}\}^{\frac{1}{2+2\delta_0}} < \infty$ and that $R_1 := \max_{k\in[K]}\{\mathbb{E}|u_{0k}'(W_{0k\cdot}^\top \boldsymbol{\theta})|^{1+\delta_0}\}^{\frac{1}{1+\delta_0}} < \infty$.

A3 Assume that the distributions of $Z_1, \dots, Z_K$ are $(\alpha, \tilde{C}, \sigma)$-sub-Weibull. Write $c_1 := \max_{k\in[K]} \mathbb{E}Z_k^4 < \infty$.

A4 Write $p_k(\cdot)$ as the density of Z_k and assume $M := \max_{k\in[K]} \sup_{x\in\mathbb{R}} |p_k(x)| < \infty$.

A5 Write $C_* := \|\Sigma_{\boldsymbol{\theta}}^{-1/2}\tilde{\Sigma}^{1/2}\|_2$ and $c_* := s_{\min}(\Sigma_{\boldsymbol{\theta}}^{-1/2}\tilde{\Sigma}^{1/2})$ where $s_{\min}(\cdot)$ denotes the minimum singular value of a matrix. Assume that $0 < c_* < C_* < \infty$.

We give a brief discussion of these assumptions. Condition A1 states that the parameter B in the estimation procedure is chosen to be large enough so that the true score functions in $\boldsymbol{u}_0$ lie in our constraint set $\mathcal{F}_{b,B,m}$ (or, more precisely, a version of $\boldsymbol{u}_0$ truncated to have support $[-b, b]$). For instance, if $u_{0k}^{(m)}$ is bounded, then we may choose B to be of the same order as b. Conditions A2 and A3 are used in the truncation argument. The quantity $\frac{1}{2}$ in the condition on δ_0 has no significance; it can be set to any value in $(0, 1)$ at the cost of inflating the constants in the risk bound. The Weibull tail decay parameter α is also allowed to take on any value greater than 0, but a smaller α increases the power of the poly-log term in the risk bound. Conditions A4 and A5 are very mild.

Our model is identifiable up to a signed permutation if and only if at most one component of $\boldsymbol{z}$ is Gaussian. Therefore, the rate of convergence in estimating W_0 and $\boldsymbol{u}_0$ depends on a quantity that measures the extent to which the distribution of $\boldsymbol{z}$ deviates from having at least two Gaussian components. We define this quantity below:

Definition 2. *We say that a matrix $H \in \mathbb{R}^{K\times K}$ is skew-symmetric if $H = -H^\top$ (note that if a matrix is skew-symmetric, then its diagonals must be zero). Define*

$$\kappa := \inf_H \mathbb{E}\| - H\boldsymbol{u}_0(\boldsymbol{z}) + (J\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\|_2^2,$$

where the infimum is over all skew-symmetric H satisfying $\|H\|_F = 1$.

It is not obvious that κ should relate to deviation from Gaussianity, but the next Lemma makes the connection clear.

Lemma 2. *Let $\boldsymbol{z} = (Z_1, \ldots, Z_K)^\top$ be independent random variables with mean-zero, unit variance, and score function $\boldsymbol{u}_0(t) = (u_{01}(t), \ldots, u_{0K}(t))^\top$. It holds that $\kappa > 0$ if and only if at most one Z_j is Gaussian.*

To understand the definition of κ, we first review a characterization of special orthogonal matrices from differential geometry. For any orthogonal matrix $V \in SO(K)$, we can write $V = e^H$ for some skew-symmetric matrix H; conversely, e^H is always an orthogonal matrix in $SO(K)$. In particular, the identity matrix is the exponential of the zero matrix (every entry is zero). This correspondence is not unique in that we may have $V = e^H = e^{H'}$ for two different skew-symmetric matrices H, H'. However, using the definition of the matrix exponential, it is easy to show that if $V = e^H$ for some skew-symmetric matrix H such that $\|H\|_F$ is small, then $\|V - I\|_F$ must also be small (see Lemma 9). This characterization of $SO(K)$ also plays a crucial role in our optimization algorithm, as we discuss in Section S1.2.

If $\boldsymbol{u}_0$ is the score function of random vector $\boldsymbol{z}$, then the score function of the rotated random vector $e^H \boldsymbol{z}$ is $\boldsymbol{z} \mapsto e^{-H}\boldsymbol{u}_0(e^H \boldsymbol{z})$. Using the fact that the derivative of $r \mapsto e^{rH}$ is He^{rH} for $r \in \mathbb{R}$, we can differentiate $e^{-rH}\boldsymbol{u}_0(e^{rH}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})$ with respect to the scalar r to see that the derivative at $r = 0$ is exactly $-H\boldsymbol{u}_0(\boldsymbol{z}) + (J\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}$. Therefore, κ can be interpreted as how much the score function changes if we apply an infinitesimally small rotation to $\boldsymbol{z}$ in the direction of e^H.

3.2 Main Result

Let $X(t) = m_*(t) + g(t)$ be a new functional curve that is independent and identically distributed to the training data $X_i(t)$, $i = 1, \ldots, n$. Denote by $\boldsymbol{\theta}$ the basis coefficient vector of $X(t)$ defined analogously to the $\boldsymbol{\theta}_i$'s when a working basis $\mathbf{s}(t)$ is used. Then $\boldsymbol{\theta}$ is independent and identically distributed to the $\boldsymbol{\theta}_i$'s. We have the following result on the estimation accuracy of the FEmBa estimate $\hat{m}^{\mathbf{s}}_X(t)$ as defined in (7) compared to the oracle Tweedie estimate $\tilde{m}^{\mathbf{s}}_X(t)$.

Theorem 2. *Let $\hat{W}$ and $\hat{\boldsymbol{u}}$ be estimators defined as (20) where we choose $b = 2\sigma(3\log n)^{1/\alpha}$ and B and $m \geq 3$ to satisfy A1 in Condition 2. Assume also A2, A3, and A4 in Condition 2.*

There exists universal constants $C, C' > 0$ and a constant $\xi > 0$ possibly depending on $\boldsymbol{u}_0$ such that if $n^{\frac{3}{2m+3}} \geq CKR_0^2 R_1 \tilde{C}$, that $n^{\frac{1}{2m+3}} \geq C\log n$, and that

$$n^{\frac{m-1}{2m-1}} \log^{-1} n \geq C \frac{C_*^6}{c_*} b^{8m} B^2 K^2 (\log K) \left(\frac{K^2 c_1}{\kappa^2} \vee \frac{1}{\xi \kappa c_*^2} \right), \tag{22}$$

then, with probability at least $1 - \frac{C'\tilde{C}}{n}$,

$$\mathbb{E}_{\boldsymbol{\theta}} \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|^2_{\tilde{\Sigma}} \leq \left(\frac{K^{3+\frac{1}{m}} \log^2 K}{n} \right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4 \kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4},$$

where the expectation above is taken with respect to $\boldsymbol{\theta}$.

As a direct consequence, we have with probability at least $1-\frac{C'\tilde{C}}{n}$*, the FEmBa estimate* $\hat{m}_X(t) = \mathbf{s}(t)^\top(\boldsymbol{\theta} + \hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}))$ *satisfies that*

$$\mathbb{E}_{\boldsymbol{\theta}}[\|\hat{m}_X^{\mathbf{s}} - \tilde{m}_X^{\mathbf{s}}\|^2] \leq \left(\frac{K^{3+\frac{1}{m}}\log^2 K}{n}\right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4\kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4}.$$

We relegate the full proof of Theorem 2 to Section S4.3 of the appendix and sketch the main ideas in Section 3.2.1.

Remark 2. *The quantities* C_*, c_*, M, m, κ *do not depend on* n*. The quantity* b *has poly-logarithmic dependence on* n*. Therefore, so long as* B *has poly-logarithmic dependence on* n*, Theorem 2 shows that the rate of convergence is at least as fast as* $O\left((\frac{K^{3+\frac{1}{m}}\log^2 K}{n})^{\frac{2m}{2m+3}} \textit{polylog}(n)\right)$.

Theorem 2 as stated requires m *to be at least 3 but the same result holds for* $m = 2$ *if we replace the Sobolev constraint* $\int_{-b}^{b} |u_k^{(m)}|^2 \leq B$ *with a stronger Hölder constraint* $\|u_k^{(m)}\|_\infty \leq Bb^2$ *in the estimation procedure* (20)*.*

Theorem 2 bounds the expected out-of-sample error. Using similar techniques, we can also provide guarantees on the in-sample error with respect to the n observed curves.

Theorem 3. *Let* $\hat{W}$ *and* $\hat{\boldsymbol{u}}$ *be estimators defined as* (20) *where we choose* $b = 2\sigma(3\log n)^{1/\alpha}$ *and* B *and* $m \geq 3$ *to satisfy Condition 2, A1. Assume also A2, A3, and A4 in Condition 2. Under the same conditions on* n *as in Theorem 2, we have that, with probability at least* $1 - 2\frac{C'\tilde{C}}{n}$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_{\tilde{\Sigma}}^2 \leq \left(\frac{K^{3+\frac{1}{m}}\log^2 K}{n}\right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4\kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4}.$$

As a direct consequence, with probability at least $1 - 2\frac{C'\tilde{C}}{n}$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{m}_{X_i}^{\mathbf{s}} - \tilde{m}_{X_i}^{\mathbf{s}}\|^2 \leq \left(\frac{K^{3+\frac{1}{m}}\log^2 K}{n}\right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4\kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4}.$$

We relegate the full proof of Theorem 3 to Section S4.4 of the appendix.

One of the key steps in proving Theorems 2 and 3 is to show that our risk function is locally strongly convex around $W_0, \boldsymbol{u}_0$. To be precise, define the equivalence class

$$[W_0, \boldsymbol{u}_0] = \{(\tilde{W}_0, \tilde{\boldsymbol{u}}_0) : \text{for a signed perm. matrix } P, \tilde{W}_0 = PW_0,\ \tilde{\boldsymbol{u}}_0(\boldsymbol{z}) = P\boldsymbol{u}_0(P^\top \boldsymbol{z})\}, \quad (23)$$

and all $\tilde{W}_0, \tilde{\boldsymbol{u}}_0$ in the equivalence class yield the same score function $\tilde{W}_0^\top \tilde{\boldsymbol{u}}_0(\tilde{W}_0\boldsymbol{\theta})$. Then, Corollary 5 in the appendix (particularly (S4.57)) shows that if $W, \boldsymbol{u}$ is sufficiently close to some $(\tilde{W}_0, \tilde{\boldsymbol{u}}_0) \in [W_0, \boldsymbol{u}_0]$, then

$$\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_{\tilde{\Sigma}}^2 \geq c_*^2\{\frac{\kappa}{4}\|W\tilde{W}_0^{-1} - I_K\|_F^2 + \frac{1}{4}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \tilde{\boldsymbol{u}}_0(\boldsymbol{z})\|_2^2\},$$

Combining this with Theorem 2, we also show consistency in the estimation of W_0 and of $\boldsymbol{u}_0$:

Corollary 1. *Under the same conditions given in the statement of Theorem 2, with probability at least* $1 - \frac{C'\tilde{C}}{n}$*, there exists* $(\tilde{W}_0, \tilde{\boldsymbol{u}}_0) \in [W_0, \boldsymbol{u}_0]$ *such that*

$$\|\hat{W}\tilde{W}_0^{-1} - I_K\|_F^2 \leq 8\Big(\frac{K^{3+\frac{1}{m}}\log^2 K}{n}\Big)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^6\kappa^3} M^2 2^{6m} b^{6m^2} B^{2m+4},$$

$$\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{z}}(\hat{u}_k(Z_k) - \tilde{u}_{0k}(Z_k))^2 \leq 12\Big(\frac{K^{3+\frac{1}{m}}\log^2 K}{n}\Big)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^6\kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4},$$

where the expectation is with respect to $\boldsymbol{z} = (Z_1, \ldots, Z_K)^\top$*, and* $\tilde{\boldsymbol{u}}_0 = (\tilde{u}_{01}, \ldots, \tilde{u}_{0K})^\top$*.*

We prove Corollary 1 in Section S4.3.1 of the appendix.

Remark 3. *In Theorem 2, we require* n *to be large enough where the lower bound depends on a scalar* $\xi > 0$*, which, as shown in the proof of Theorem 2, is defined as the largest real number such that if* $\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_\Sigma^2 \leq \xi$ *for any* $W, \boldsymbol{u}$*, then*

$$\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_\Sigma^2 \geq c_*^2 \frac{\kappa}{8} \min_P \|PWW_0^{-1} - I_K\|_F^2$$

where the minimum with respect to P *is over all signed permutation matrices. We prove in Proposition 8 that so long as* $\kappa > 0$*,* ξ *exists and is strictly positive. The magnitude of* ξ *depends on smoothness of the underlying density of* $\boldsymbol{z}$ *(and hence of* $\boldsymbol{u}_0$*) but the exact relationship is difficult to characterize. We leave an in-depth investigation to future work.*

Remark 4. *We have not optimized the dependence on* K *in the rate given in Theorem 2. A dependence of* K^2 *is unavoidable since we need to estimate a square matrix* W_0*.*

We have an exponent of $\frac{2m}{2m+3}$ *on* n *instead of* $\frac{2m}{2m+1}$ *which one might expect from estimating univariate functions with* m *degrees of smoothness. This is because the first derivative* u_k' *appears in the optimization objective* (20)*, which has only* $m-1$ *degree of smoothness. This leads one to expect a rate of* $\frac{2(m-1)}{2(m-1)+1}$ *which is still larger than* $\frac{2m}{2m+3}$*. The difference is explained by the fact that our risk is still in terms of* u_k *and hence, we have to bound the error on* u_k *from error bounds on* u_k'*, which leads to additional deterioration in the rate.*

We suspect that the exponent $\frac{2m}{2m+3}$ *may not be optimal but it is not obvious what the optimal rate should be even when* W_0 *is known. Estimating a score function* $p'(x)/p(x)$ *is different from estimating the density* $p(x)$ *because* $p(x)$ *appears in the denominator and hence, the region where* $p(x)$ *is very small can still contribute significantly to the overall error.*

Remark 5. *Since the objective* $\hat{F}(W, \boldsymbol{u})$ *in the estimation procedure* (20) *is not convex, our optimization algorithm (described in Section 4) may not obtain the global optimum. Theorem 2 applies to any* $\hat{W}, \hat{\boldsymbol{u}}$ *such that* $\hat{F}(\hat{W}, \hat{\boldsymbol{u}}) - \hat{F}(W_0, \boldsymbol{u}_0)$ *is no larger than the error bound* $\big(\frac{K^{3+\frac{1}{m}}\log^2 K}{n}\big)^{\frac{2m}{2m+3}}$ *polylog*(n) *given in Theorem 2.*

3.2.1 Proof sketch

We give the full proof of Theorem 2 in Section S4.3 of the appendix and sketch the main ideas here. The proof combines a truncation argument with empirical process theory techniques and a local strong convexity analysis of the risk function.

To be precise, we truncate by working on the event that the unmixed components $Z_1, \dots, Z_K$ lie in the hypercube $[-b/2, b/2]^K$. This event occurs with high probability by assuming that all components of $\boldsymbol{z}_i$ are sub-Weibull. We can further bound the error introduced by the truncation using A2 in Condition 2. On this event, we can use the 0-th order condition that $\hat{F}(\hat{W}, \hat{\boldsymbol{u}}) \leq \hat{F}(W_0, \boldsymbol{u}_0)$ to derive what is known as a "basic inequality" for an empirical process argument:

$$R(\hat{W}, \hat{\boldsymbol{u}}) \leq \left| \frac{1}{n} \sum_{i=1}^{n} \hat{g}(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}_i} \hat{g}(\boldsymbol{\theta}_i) \right| \leq \sup_g \left| \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}_i} g(\boldsymbol{\theta}_i) \right|$$

where $g(\cdot)$ is a function that correspond to a pair $W, \boldsymbol{u}$ and lies inside a function class $\mathcal{G}$ and $\hat{g}$ corresponding to $\hat{W}, \hat{\boldsymbol{u}}$. To bound $R(\hat{W}, \hat{\boldsymbol{u}})$, it thus suffices to analyze the supremum of the empirical process $\sup_{g \in \mathcal{G}} |n^{-1} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}_i} g(\boldsymbol{\theta}_i)|$.

We then apply a peeling argument in which the key is bounding $\sup_g |n^{-1} \sum_{i=1}^{n} g(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}_i} g(\boldsymbol{\theta})|$ where $g(\cdot)$ corresponds to $W, \boldsymbol{u}$ whose risk $\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_\Sigma^2$ is bounded from above by say r^2 for various levels of r. To bound this supremum, we derive a bracketing entropy bound on the function class $\mathcal{G}$ and a bound on the variance of $g(\boldsymbol{\theta})$. In particular, the bound on the variance of $g(\boldsymbol{\theta})$ requires us to show that the risk is locally strongly convex, i.e., if $\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_\Sigma^2$ is small, then $W, \boldsymbol{u}$ is also sufficiently close to $W_0, \boldsymbol{u}_0$ up to a signed permutation matrix transformation.

4 Fitting FEmBa

We now present an algorithm for estimating $\mathbf{v}_0$, needed for calculating the FEmBa estimate $\hat{m}_X^{\mathbf{s}}$. We start by presenting the following empirical risk function, which corresponds to (16) but with the smoothness penalty residing in the objective function, that we minimize to obtain the FEmBa estimate in practice:

$$\begin{aligned} \hat{Q}(\boldsymbol{u}, W) = &\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}(W\boldsymbol{\theta}_i)^\top W \tilde{\Sigma} W^\top \boldsymbol{u}(W\boldsymbol{\theta}_i) \\ &+ \sum_{k=1}^{K} c(W_{k\cdot}) \left\{ \frac{2}{n} \sum_{i=1}^{n} u_k'(W_{k\cdot}^\top \boldsymbol{\theta}_i) + \lambda \int [u_k''(W_{k\cdot}^\top \boldsymbol{\theta})]^2 d\boldsymbol{\theta} \right\}. \end{aligned} \tag{24}$$

Recall that the true unmixing matrix, W_0, has the representation $W_0 = U_0 \Sigma_{\boldsymbol{\theta}}^{-1/2}$, where $\Sigma_{\boldsymbol{\theta}}$ is the covariance of $\boldsymbol{\theta}$, and $U_0 \in SO(K)$. Thus, we can further reparametrize (24) as a function of $\boldsymbol{u}$ and U, and minimize it with respect to $\boldsymbol{u}$ and U under the constraint that $U \in SO(K)$. Let $Q_n(\boldsymbol{u}, U) = \hat{Q}(\boldsymbol{u}, U\Sigma_{\boldsymbol{\theta}}^{-1/2})$. Then the optimization problem we wish to solve is

$$\min_{\boldsymbol{u}, U} Q_n(\boldsymbol{u}, U) \text{ such that } U \in SO(K). \tag{25}$$

We minimize (25) via alternating coordinate descent, updating $\boldsymbol{u}$ while holding U constant, and then updating U conditional on the updated $\boldsymbol{u}$. The covariance matrix $\Sigma_{\boldsymbol{\theta}}$ can be replaced with the sample covariance matrix $\hat{\Sigma}_{\boldsymbol{\theta}}$. We stop the alternating process when the relative decrease in the objective function $Q_n(\boldsymbol{u}, U)$ falls below a pre-specified threshold. To optimize (25) with respect to $\boldsymbol{u}$ for a fixed U, we take a basis representation of $\boldsymbol{u}$ and optimize the basis coefficients; to optimize U for a fixed $\boldsymbol{u}$, we derive a geodesic gradient algorithm (see e.g. Plumbley (2005)) to ensure that U stays in $SO(K)$. We relegate the detailed description of the algorithm to Section S1 of the supplementary file.

Denote the final estimates as $\hat{\boldsymbol{u}}$ and $\hat{U}$. Then the score function $\mathbf{v}_0$ can be estimated as

$$\hat{\mathbf{v}}(\boldsymbol{\theta}) = \hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) \text{ with } \hat{W} = \hat{U}\hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2} \tag{26}$$

The final FEmBa estimate $\hat{m}_X^{\mathbf{s}}(t)$ is then constructed according to (17).

5 Monte Carlo Simulations

We now assess the empirical performance of FEmBa, both in terms of how well it estimates the true score function $\mathbf{v}_0(\boldsymbol{\theta})$, and how well FEmBa estimates of the mean curves estimate the true mean curves, the $m_{*i}(t)$'s. We compare several different approaches in this simulation study that make different assumptions about the data. The first approach assumes that the coordinates of $\boldsymbol{\theta}$ are independent, and is denoted as FEmBa$_{NT}$. The next method assumes that decorrelation is sufficient to produce independent coordinates of $\boldsymbol{\theta}$ (i.e., $\Sigma_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\theta}$ gives a vector with independent components), and is denoted as FEmBa$_T$. For both methods the score function $\mathbf{v}_0(\boldsymbol{\theta}) = W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})$ is estimated by minimizing the empirical risk function (24) with respect to $\boldsymbol{u}$, with $W = I_K$ for FEmBa$_{NT}$ and $W = \hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2}$ for FEmBa$_T$, respectively. Then the method introduced in Section 4 can be used to get the estimate of $\boldsymbol{u}_0$, denoted as $\hat{\boldsymbol{u}}$. Thus, the final estimate of $\mathbf{v}_0$ can be obtained as $\hat{\mathbf{v}}(\boldsymbol{\theta}) = W^\top \hat{\boldsymbol{u}}(W\boldsymbol{\theta})$ with $W = I_K$ for FEmBa$_{NT}$ and $W = \hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2}$ for FEmBa$_T$. We also consider an approach that estimates $\mathbf{v}(\boldsymbol{\theta})$ via direct estimation of the multivariate density, and we denote this method KDE$_M$. This method utilizes a simplified vine-copula approach to obtain a kernel density estimate of the underlying density of $\boldsymbol{\theta}$ (Nagler and Czado, 2016). If we denote this density estimate as $\hat{f}(\boldsymbol{\theta})$, we then estimate the score function as $\hat{\mathbf{v}}(\boldsymbol{\theta}) = \nabla \hat{f}(\boldsymbol{\theta})/\hat{f}(\boldsymbol{\theta})$.

The next two methods, which we refer to as FEmBa$_{fastICA}$ and FEmBa$_{jointICA}$, assume Model (12), and exploit this assumption to estimate the needed score function. FEmBa$_{jointICA}$ uses the algorithm discussed in Section 4 to estimate $\boldsymbol{u}$ and U, and finally estimate $\mathbf{v}_0$ according to (26). FEmBa$_{fastICA}$ estimates the unmixing matrix W_0 separately from $\boldsymbol{u}_0$. Specifically, the FastICA algorithm is applied to the $\boldsymbol{\theta}_i$'s to obtain the estimate $\hat{W}$. Then we minimize (24) with respect to $\boldsymbol{u}$ with W held constant at the FastICA estimate $\hat{W}$. Therefore, the main difference between these two methods is whether W and $\boldsymbol{u}$ are jointly or separately estimated.

All four of these methods are compared to an oracle estimator, denoted ORACLE, which assumes W_0 is known. Since the true score function of $\boldsymbol{\theta}_i$ generally does not admit an explicit expression, for the oracle estimator we use the true unmixing matrix W_0, and then estimate $\boldsymbol{u}_0$ numerically with the method discussed in Section S1.1. Finally we apply (26) to get the oracle estimate of $\mathbf{v}_0$.

We generate the data from the following model so that $\boldsymbol{\theta}_i$ has an ICA structure

$$\boldsymbol{\theta}_i = \tilde{W}_0(\boldsymbol{y}_i\sqrt{\text{SNR}} + \boldsymbol{\gamma}_i) = \boldsymbol{\mu}_{*i} + \boldsymbol{\gamma}_i^*, \quad i = 1, \ldots, n \tag{27}$$

where the $\boldsymbol{y}_i$'s are i.i.d. random vectors with independent coordinates that have mean 0 and variance 1, $\gamma_i \overset{iid}{\sim} N(0, I_K)$, $\tilde{W}_0$ is the mixing matrix, $\boldsymbol{\mu}_{*i} = \tilde{W}_0 \boldsymbol{y}_i \sqrt{\text{SNR}}$, $\boldsymbol{\gamma}_i^* = \tilde{W}_0 \boldsymbol{\gamma}_i$, and the SNR > 0 is the signal-to-noise ratio of the problem. The coordinates of $\boldsymbol{y}_i$ are simulated from an asymmetric mixture of various different distributions including Weibull, Gamma, and Gaussian.

The parameter SNR controls the signal strength in the data, relative to the Gaussian noise. A large signal-to-noise ratio implies that the behavior of the prior $m_{*i}(t)$ is prevalent in the observed data and that little shrinkage is necessary, while a small SNR implies the data is very corrupted by Gaussian noise and that shrinkage is appropriate. We choose $\tilde{W}_0 = \Sigma_0^{1/2} U^\top$, where Σ_0 has an AR(1) structure with correlation parameter 0.3 and $U \in SO(K)$. We also choose a basis, denoted $\mathbf{s}_{\boldsymbol{\theta}}$, in which the curves are perfectly represented. In this notation, $m_{*i}(t) = \mathbf{s}_{\boldsymbol{\theta}}(t)^\top \boldsymbol{\mu}_{*i}$, $g_i(t) = \mathbf{s}_{\boldsymbol{\theta}}(t)^\top \boldsymbol{\gamma}_i^*$, and $X_i(t) = m_{*i}(t) + g_i(t)$. The basis $\mathbf{s}_{\boldsymbol{\theta}}$ is chosen to be a 5-dimensional cubic spline basis. In the following simulations, we generate $J = 50$ functional data sets, each with $n = 1000$ curves observed at $T = 100$ time points $\{t_j\}_{j=1}^T$ over the interval $[0, 1]$. It is seen that our simulations consider functional data observed at discrete but dense time points.

5.1 Score Function Estimation

We first evaluate each approach's ability to estimate $\mathbf{v}_0(\boldsymbol{\theta})$ using $\boldsymbol{\theta}_i$'s fitted from the discretely simulated data $X_i(t_j)$ with true basis $\mathbf{s}_{\boldsymbol{\theta}}$. We calculate $\boldsymbol{\theta}_i$ using the empirical integral as

$$\boldsymbol{\theta}_i = T^{-1} \Sigma_{\mathbf{s}_{\boldsymbol{\theta}}}^{-1} \sum_{j=1}^{T} X_i(t_j) \mathbf{s}_{\boldsymbol{\theta}}(t_j). \tag{28}$$

The robustness analysis with misspecified basis will be conducted in Section 5.3. Using the $\boldsymbol{\theta}_i$'s, we compute $\hat{\boldsymbol{v}}$ using various methods and measure peformance with the empirical risk

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i))^\top \tilde{\Sigma} (\hat{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i)).$$

We present the comparison results in Section S2 of the supplementary file. The numerical results (see Tables 9, 10, and 11 in the supplementary file) show that our proposed approach performs best in all regimes except when SNR $= 0.75$ so that the data is predominantly Gaussian noise. See detailed discussions in the supplementary file.

These results line up closely with existing works in the ICA literature, where FastICA is compared to other approaches (Hastie and Tibshirani, 2002; Bach and Jordan, 2002). In particular, as discussed in Hastie and Tibshirani (2002), "FastICA uses very simple approximations (to the mutual information) based on a single (or a small number of) non-linear contrast functions, which work well for a variety of situations, but not at all well for (the) more complex Gaussian Mixtures." As we have seen, FEmBa also outperforms FastICA when the prior distribution for $\boldsymbol{z}_i$ is some sufficiently complicated mixture, though we are focusing on score function estimation, not estimation of the unmixing matrix. In addition to the complexity of the mixture distribution for $\boldsymbol{z}$, the complexity of the distribution for $\boldsymbol{\theta}_i$ is also related to the SNR. When the SNR is low, the Gaussian noise curve dominates and thus the distribution of $\boldsymbol{\theta}_i$ is close to Gaussian and has low complexity, while when the SNR is high, the prior distribution of $m_{i*}(t)$ dominates and thus the distribution of $\boldsymbol{\theta}_i$'s have high complexity. This explains the relative performance of FEmBa$_{jointICA}$

and $\text{FEmBa}_{fastICA}$.

5.2 Curve Estimation

We now evaluate each method's ability to correct the bias in $X_i(t)$ for estimating the true mean curves $m_{*i}(t)$'s. The data are simulated in the same way as in the last section, and we add two additional comparison methods that estimate the $m_{*i}(t)$'s. One method, SMOOTHED, produces estimates of the $m_{*i}(t)$'s with $\hat{m}_{*i}(t) = \mathbf{s}_{\boldsymbol{\theta}}(t)^\top \boldsymbol{\theta}_i$, the unshrunk curves. The other approach ignores the correlation structure between $X_i(t)$ at different time points, and estimates the $m_{*i}(t)$'s pointwise via the scalar version of Tweedie's formula. In other words, at each time point t, we estimate $m_{*i}(t)$ by applying univariate Tweedie's formula to the scalar data set $\{X_1(t), \cdots, X_N(t)\}$. Note that we are not first projecting the curves onto $\mathbf{s}_{\boldsymbol{\theta}}(t)$ prior to applying Tweedie's formula pointwise; neither do we smooth these pointwise shrinkage estimates afterwards. We denote this approach as KDE_U, which represents how we estimate the *univariate* score function $v_t(X(t))$ at each time point t. As with KDE_M, we first obtain a kernel estimate of the density of $X(t)$ at each time point, and then use this estimate to approximate the corresponding score function $v_t(\cdot)$. Here, we index the score function by t to emphasize that it can vary from one time point to another. We remark that KDE_U was not included in the score function estimation results in the last section because this method does not involve basis projection and hence its score function is different from the one used by other methods.

The two kernel density estimation based methods, KDE_M and KDE_U, generally perform much worse than ICA based methods, with the expectation that KDE_U performs the best when SNR = 0.75, when the data is predominantly Gaussian noise, a scenario close to the ideal setting for applying univariate Tweedie's formula.

As we do not have a consistent estimator for the $m_{*i}(t)$'s, we measure performance by a *relative* risk. Define the following loss function for curve estimation:

$$\hat{R}_m(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}\int (\hat{m}_i(t) - m_{*i}(t))^2 dt,$$

where $\hat{m}_i(t)$ is an estimator of curve $m_{*i}(t)$, and the integral above is evaluated numerically using the discretely observed curves. As we have access to J data sets, we denote the value of $\hat{R}_m$ calculated from data set j as $\hat{R}_m^{(j)}$. Our final measure of performance, then, is the following quantity:

$$\hat{R}_{\text{rel}}(\hat{m}) = \frac{1}{J}\sum_{j=1}^{J}\frac{|\hat{R}_m^{(j)}(\hat{m}) - \hat{R}_m^{(j)}(\hat{m}_{\text{oracle}})|}{\hat{R}_m^{(j)}(\hat{m}_{\text{oracle}})}, \tag{29}$$

where $\hat{R}^{(j)}(\hat{m}_{\text{oracle}})$ is the empirical risk for the oracle Tweedie estimator for data set j. The quantity $\hat{R}_{\text{rel}}$ is used to measure each method's ability to estimate the $m_{*i}(t)$'s relative to that of the oracle estimator. The simulation results are in Tables 2 through 4. The relative performance of the methods in these scenarios is very similar to that of the score function estimation, where $\text{FEmBa}_{jointICA}$ performs consistently best when the mixture distribution of the $\boldsymbol{z}$'s has high complexity, though $\text{FEmBa}_{fastICA}$ and FEmBa_T are more competitive when the SNR is low.

SNR	0.75	1.562	2.375	3.188	4
FEmBa_{NT}	333.93 (8.76)	661.13 (21.33)	495.84 (12.79)	409.52 (8.17)	345.4 (4.82)
FEmBa_{T}	274.46 (9.89)	585 (20.17)	463.17 (12.36)	393.72 (8.09)	335.67 (4.74)
KDE_{M}	396.54 (12.24)	566.39 (18.78)	445.34 (12.43)	375.47 (7.93)	316.6 (4.9)
KDE_{U}	**273.33** (9.59)	573.26 (19.62)	459.25 (12.36)	389.27 (8.07)	332.55 (4.88)
$\text{FEmBa}_{fastICA}$	306.36 (13.95)	544 (20.23)	391.78 (14.82)	310.17 (10.55)	244.63 (7.88)
$\text{FEmBa}_{jointICA}$	293.92 (21.45)	**127.36** (20.18)	**85.38** (11.51)	**95.8** (12.89)	**94.36** (8.77)
SMOOTHED	2711.86 (45.02)	1676.46 (35.06)	869.79 (15.12)	610.04 (8.95)	471.67 (4.85)

Table 2: Relative risk (29) of different curve estimators for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of two Weibull distributions. All error values in the table are multiplied by 10^3.

SNR	0.75	1.562	2.375	3.188	4
FEmBa_{NT}	153.1 (3.48)	175.46 (3.7)	116.71 (1.99)	92.51 (1.67)	79.03 (1.74)
FEmBa_{T}	**83.67** (3.74)	121.09 (3.74)	91.28 (1.88)	78.64 (1.52)	69.54 (1.61)
KDE_{M}	217.5 (7.2)	117.37 (3.56)	80.86 (2)	70.89 (1.57)	61.85 (1.71)
KDE_{U}	87.54 (2.49)	119.73 (3.53)	89.65 (1.76)	76.94 (1.6)	68.68 (1.65)
$\text{FEmBa}_{fastICA}$	108.96 (6.8)	107.54 (3.27)	79.38 (2.25)	69.35 (1.7)	62.16 (1.7)
$\text{FEmBa}_{jointICA}$	142.01 (12.55)	**45.31** (4.32)	**35.46** (3.95)	**55.53** (3.38)	**61.12** (2.06)
SMOOTHED	2414.48 (28.72)	908.25 (7.11)	410.97 (3.25)	257.69 (1.89)	185.14 (1.99)

Table 3: Relative risk (29) of different curve estimators for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of two Gamma distributions, one with shape and scale parameters 1 and .5, respectively, and the other with shape and scale parameters 1 and 8, respectively. All error values in the table are multiplied by 10^3.

SNR	0.75	1.562	2.375	3.188	4
FEmBa$_{NT}$	104.13 (3.95)	189.83 (5.38)	153.8 (2.16)	142.97 (1.68)	131.22 (1.59)
FEmBa$_T$	**56.47** (3.34)	132.06 (4.85)	125.63 (2.07)	127.06 (1.54)	120.68 (1.56)
KDE$_M$	143.32 (5.21)	138.46 (4.87)	123.58 (2.18)	125.16 (1.54)	118.97 (1.61)
KDE$_U$	64.67 (3.58)	127.83 (4.88)	122.58 (2.12)	124.8 (1.52)	119.25 (1.56)
FEmBa$_{fastICA}$	91.76 (7.79)	35.96 (4.19)	11.23 (1.75)	8.97 (1.15)	7.31 (1.36)
FEmBa$_{jointICA}$	176.77 (14.77)	**29.98** (3.21)	**11.23** (1.47)	**6.76** (0.93)	**3.8** (0.83)
SMOOTHED	1933.35 (27.63)	925.92 (9.46)	448.44 (2.85)	308.29 (1.87)	237.21 (1.82)

Table 4: Relative risk (29) of different curve estimators for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of a gamma distribution with shape and scale parameters 1 and .5, respectively, and a Gaussian distribution with mean and standard deviation 10 and 3, respectively. All error values in the table are multiplied by 10^3.

5.3 Robustness of the proposed method

We study the robustness of FEmBa for curve estimation from two perspectives: 1) when the discretely observed curves are contaminated with discrete measurement errors; and 2) when the basis $\mathbf{s}$ chosen for estimating the $\boldsymbol{\theta}_i$'s is misspecified and different from $\mathbf{s}_{\boldsymbol{\theta}}$ that generates the data. In all numerical studies presented in this section, we generate data from the following Model (30), where we observe

$$Y_i(t_j) = X_i(t_j) + \varepsilon_{it_j}, \tag{30}$$

with $i = 1, \ldots, 1000$ and $j = 1, \ldots, 100$. Here, $\varepsilon_{it_j} \overset{iid}{\sim} N(0, 0.03)$. We calculate the $\boldsymbol{\theta}_i$'s identically to (28) with $X_i(t_j)$ replaced with $Y_i(t_j)$. Note that these $\boldsymbol{\theta}_i$'s are generated from a misspecified ICA model, even when the working basis $\mathbf{s}$ is the same as the true basis $\mathbf{s}_{\boldsymbol{\theta}}$, because of the added Gaussian white noise ε_{it_j}'s.

Table 5 summarize the comparison results for curve estimation under the performance measure (29) when the working basis $\mathbf{s}$ is chosen to be $\mathbf{s}_\theta$. It is seen that the additional noise causes the ICA model to be mis-specified and deteriorates the performance of all methods in almost all scenarios, but the relative performance of these methods stays the same as discussed in the last section.

Table 6 presents results when the working basis $\mathbf{s}$ is also different from the true basis $\mathbf{s}_{\boldsymbol{\theta}}$. In particular, the working basis is chosen to be a cubic B-spline basis with $K = 5$, while the underlying true basis $\mathbf{s}_{\boldsymbol{\theta}}$ is a cubic spline with $K = 8$. The results suggest that FEmBa can outperform standard approaches, with FEmBa$_{jointICA}$ outperforming FEmBa$_{fastICA}$, as long as the SNR is not very low.

SNR	0.75	1.562	2.375	3.188	4
FEmBa_{NT}	598.25 (16.14)	663.36 (21.33)	469.18 (13.14)	374.22 (7.7)	311.38 (3.14)
FEmBa_{T}	717.1 (19.35)	704.29 (22.51)	490.5 (13.55)	394.44 (7.98)	325.77 (3.57)
KDE_{M}	851.65 (20.5)	700.74 (22.35)	474.56 (13.19)	375.44 (8.27)	309.53 (3.57)
KDE_{U}	**571.28** (17.92)	623.49 (20.35)	449.35 (12.62)	363.84 (7.68)	305.98 (3.15)
$\text{FEmBa}_{fastICA}$	788.81 (22.8)	660.8 (21.59)	431.65 (13.16)	297.87 (9.6)	239.4 (6.47)
$\text{FEmBa}_{jointICA}$	800.71 (21.72)	**254.17** (11)	**175.88** (5.1)	**163.68** (3.53)	**157.71** (2.44)
SMOOTHED	2802.87 (50.52)	1724.49 (33.49)	887.93 (16)	624.43 (8.98)	479.25 (4.13)

Table 5: Relative risk of different curve estimators for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of two Weibull distributions, and the curves are generated according to Model (30). All error values in the table are multiplied by 10^3.

SNR	0.75	1.562	2.375	3.188	4
FEmBa_{NT}	797.19 (15.49)	899.76 (26.93)	748.42 (14.48)	735.01 (11.3)	761.66 (7.37)
FEmBa_{T}	336.43 (10.64)	772.74 (25.84)	709.14 (14.42)	715.75 (11.33)	748.94 (7.31)
KDE_{M}	396.04 (12.12)	772.01 (25.81)	697.51 (14.15)	702.31 (11.15)	738.3 (7.23)
KDE_{U}	**277.87** (9.31)	776.21 (24.86)	715.77 (14.79)	718.34 (11.55)	750.3 (7.26)
$\text{FEmBa}_{fastICA}$	352.53 (12.33)	750.52 (25.06)	684.32 (14.22)	687.25 (11.34)	723.84 (8.24)
$\text{FEmBa}_{jointICA}$	330.75 (12.13)	**669.41** (23.25)	**625.35** (13.16)	**650.89** (12.01)	**705.97** (9.13)
SMOOTHED	2650.05 (40.9)	1787.56 (39.46)	1085.98 (17.01)	919.25 (12.11)	878.51 (7.66)

Table 6: Relative risk of different curve estimators for several values of SNR. In this scenario, the distribution of Y_{ik} is an asymmetric mixture of two Weibull distributions, and the curves are generated according to Model (30). The true basis that contains the curves is also different than the basis chosen for modeling. All error values in the table are multiplied by 10^3.

6 Exoplanet Radii Estimation

We demonstrate how FEmBa can help answer the important question on whether the radii estimates for the exoplanets discovered by the Kepler missions are accurate. The exoplanet radii estimates derived from the Kepler data are calculated via the transit method, which works as follows. During missions, the Kepler telescope would point at a part of the sky for roughly 90 days and take multiple "images" of that part of the sky, which after processing yield light curves for tracking the relative flux of a star over time. If a planet passed in front of one of the stars Kepler was monitoring during this time period a small, but consistent, drop in the brightness of the star would be observed. This would manifest itself in a periodic dip in the flux values of the light curve. After some additional preprocessing steps one can obtain a light curve that tracks the proportion of light blocked out by the exoplanet as it transits the star, as a function of time, as we saw in Figure 1. One minus the minimum value of this curve is referred to as *transit depth* τ_{td}. The radius of the host star r_s, one can estimate the radius of the exoplanet, denoted r_p with the following formula:

$$r_p = r_s \sqrt{\tau_{td}}. \tag{31}$$

In the above equation, r_s is usually well-estimated, and the challenge is to estimate τ_{td}.

For this analysis we have access to NASA's MAST database, which contains light curve data for every star Kepler observed during its missions. We also have a data set cataloguing all the exoplanets discovered by Kepler, as well as their various properties and the properties of their host stars. After querying the MAST database and preprocessing the queried light curve data, we have 2313 preprocessed light curves for which we calculate transit depths τ_{td}, leading to exoplanet radii estimates by (31). The data was queried and preprocessed with the Lightkurve package (Lightkurve Collaboration et al., 2018). We assume that the observed data is generated according to Model (30), with $m_{*i}(t)$ representing a "perfect" observation of planet i's transit around its host star, $g_i(t)$ representing noise induced by various astronomical conditions and preprocessing steps, and the ε_{ij}'s representing discrete observation error induced by the Kepler telescope itself, that is normally distributed with mean 0 and variance σ^2. If we had observed the $m_{*i}(t)$'s directly, we would have perfect estimates of the transit depth (as one minus the minimum value of $m_{*i}(t)$) and thus the perfect estimate of the exoplanet radii. However, the data is corrupted by various sources of noise, and thus it's possible we need to correct for this noise, via FEmBa. As the database contains multiple transit curves for each exoplanet we can also estimate Σ_γ using the sample estimate based on these repeated measurements. This allow us to apply FEmBa to the transit curves and generate bias-corrected estimates of exoplanet radii. If the corrections are minor, then this suggests that current exoplanet radii estimates are rather accurate, assuming there are no systematic biases in data collection that affect radii estimates as a whole. If the estimates generated from the corrected curves are very different from the initial estimates, then perhaps our current exoplanet radii estimates are not as accurate as we'd hope. The results of the correction analysis are shown in Table 7.

Table 7 suggests that the uncorrected exoplanet radii estimates are quite accurate, and that the FEmBa corrected estimates are almost identical to the uncorrected estimates. Upon reflection this is not surprising, as the Kepler telescope was a space telescope, and thus there was no atmospheric distortion of the light measurements it took. The telescope also observed light emissions of various

Min	Q1	Median	Q3	Max
0.000	0.000	0.000	0.000	0.001

Table 7: Summary statistics for the distribution of the differences between the exoplanet radii estimates derived from bias-corrected curves, and radii estimates derived from uncorrected curves.

objects for unusually long amounts of time, taking many measurements during each time period. This allowed Kepler to observe the same planet transit its host star potentially many times, and for each light curve to be densely observed. All of these factors significantly reduced the noise in the data, leading to minimal selection bias.

6.1 Radii Estimation with Noise

The results of the previous section suggest that the Kepler light curve data is high quality data, with very little noise. However, we can still assess FEmBa's ability to correct these light curves when corrupted with simulated noise, and recover improved estimates of exoplanet radii using simulated data. To accomplish this, we generate data from the following model:

$$X_i(t_l) = m_{*i}(t_l) + g_i(t_l), \quad i = 1, \dots, 2313; l = 1, \dots, 75,$$

where $m_{*i}(t)$ is a smoothed light curve from the raw data; $g_i(t) = \mathbf{s}_{\boldsymbol{\gamma}}(t)^\top \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \sim N\left(0, \frac{1}{\sqrt{SNR}}\Sigma_{\boldsymbol{\gamma}}\right)$ is a random vector of dimension 7; $\mathbf{s}_{\boldsymbol{\gamma}}$ is the basis in which $g_i(t)$ is exactly represented; and SNR again represents the signal-to-noise ratio of the problem. For a given value of SNR we generate $J = 50$ such data sets, with observation i of data set j being denoted as $X_i^{(j)}$. From the $X_i^{(j)}$'s, we use the methodologies presented in the simulation section to calculate the FEmBa estimates of the m_{*i}'s. Unlike in the simulation section, we do not simulate $m_{*i}(t)$'s and thus they unlikely to have an ICA structure. For these simulations we again vary the SNR, to examine performance of the methods as a function of the amount of noise in the data. The measure of performance is the *expected* RMSE of the exoplanet radii estimates, defined as

$$\text{RMSE} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{N}\sum_{i=1}^{N}(\hat{r}_{ij} - r_i)^2\right)^{1/2}, \tag{32}$$

where $\hat{r}_{ij}$ denotes the estimate provided for planets i's radius, obtained from data set j, and r_i is planet i's radius calculated from the raw data and is treated as the underlying truth here. We calculate the RMSE for each method, and each value of the SNR. The results of this analysis are in Table 8. We can see that the bias-corrected curves provide significantly better estimates of the exoplanet radii than the uncorrected curves, and that both ICA based methods provide the best estimates. The fact that FEmBa$_{jointICA}$ does not provide much improvement over FEmBa$_{jointICA}$ may be a result of the lack of an *exact* ICA structure in the data.

SNR	2	4	6	8
FEmBa$_{NT}$	17171.364 (77.016)	13847.819 (62.933)	12246.897 (60.188)	11227.897 (47.835)
FEmBa$_{T}$	13380.170 (97.709)	12223.583 (69.348)	11506.627 (61.836)	11075.360 (55.113)
FEmBa$_{fastICA}$	10374.177 (131.708)	8897.552 (75.206)	**7911.384** (67.045)	7399.863 (55.001)
FEmBa$_{jointICA}$	**10207.020** (100.567)	**8820.057** (71.002)	7937.542 (63.757)	**7383.864** (60.385)
SMOOTHED	35587.135 (119.453)	28341.969 (84.761)	24732.566 (85.886)	22397.175 (81.477)

Table 8: The expected RMSE for exoplanet radii estimates for each considered method, for each considered value of the SNR.

References

Abraham, C., P. A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Abramovich, F., Y. Benjamini, D. L. Donoho, and I. M. Johnstone (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist. 34*(2), 584–653.

Agmon, S. (1966). *Lectures on elliptic boundary value problems*, Volume 369. American Mathematical Soc.

Alonso, A. M., D. Casado, and J. Romo (2012). Supervised classification for functional data: A weighted distance approach. *Comp. Stat. 56*(7), 2334 – 2346.

Bach, F. R. and M. I. Jordan (2002). Kernel independent component analysis. *Journal of machine learning research 3*(Jul), 1–48.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Ann. Statist. 36*(6), 2577–2604.

Birman, M. Š. and M. Z. Solomjak (1967, apr). Piecewise polynomial approximations of functions of the class w_p^α. *Mathematics of the USSR-Sbornik 2*(3), 295–317.

Chen, A. and P. J. Bickel (2006). Efficient independent component analysis. *The Annals of Statistics 34*(6), 2825–2855.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association 70*(350), 311–319.

Fan, Y., N. Foutz, G. M. James, and W. Jank (2014). Functional response additive model estimation with online virtual stock markets. *Ann. Appl. Stat. 8*(4), 2435–2460.

Fan, Y., G. M. James, and P. Radchenko (2015). Functional additive regression. *Ann. Statist. 43*(5), 2296–2325.

Feller, W. (1966). *An introduction to probability theory and its applications, vol 2.* John Wiley & Sons.

Gelman, A. and C. R. Shalizi (2012). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), 8–38.

Goldsmith, J., S. Greven, and C. Crainiceanu (2014). Corrected confidence bands for functional data using principal components. *Biometrics 69*(1), 41–51.

Hastie, T. and R. Tibshirani (2002). Independent components analysis through product density estimation. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, Cambridge, MA, USA, pp. 665–672. MIT Press.

Ikeda, Y., T. Kubokawa, and M. S. Srivastava (2016). Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions. *Computational Statistics and Data Analysis 95*, 95 – 108.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *Ann. Statist. 37*(5A), 2083–2108.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif., pp. 361–379. University of California Press.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical bayes estimation of normal means. *Ann. Statist. 37*(4), 1647–1684.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist. 40*(2), 1024–1060.

Li, B. and Q. Yu (2008). Classification of functional data: A segmentation approach. *Computational Statistics and Data Analysis 52*(10), 4790 – 4800.

Lightkurve Collaboration, J. V. d. M. Cardoso, C. Hedges, M. Gully-Santiago, N. Saunders, A. M. Cody, T. Barclay, O. Hall, S. Sagear, E. Turtelboom, J. Zhang, A. Tzanidakis, K. Mighell, J. Coughlin, K. Bell, Z. Berta-Thompson, P. Williams, J. Dotson, and G. Barentsen (2018, December). Lightkurve: Kepler and TESS time series analysis in Python. Astrophysics Source Code Library.

Nagler, T. and C. Czado (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis 151*, 69–89.

Petrone, S., S. Rizzelli, J. Rousseau, and C. Scricciolo (2014). Empirical bayes methods in classical and bayesian inference. *METRON 72*(2), 201–215.

Plumbley, M. D. (2005). Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras. *Neurocomputing 67*, 161–197. Geometrical Methods in Neural Networks and Learning.

Ramsay, J., J. Ramsay, B. Silverman, and S. S. Media (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. on Math. Statistic. and Prob. 1*, 157–163.

Samworth, R. J. and M. Yuan (2012). Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics 40*(6), 2973 – 3002.

Silverman, B. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist. 24*(1), 1–24.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

Zhang, C.-H. (1997). Empirical bayes and compound estimation of normal means. *Statistica Sinica 7*, 181–193.

Supplementary material to "An empirical Bayes shrinkage method for functional data"

Joshua Derenski, Yingying Fan, Gareth James, and Min Xu

S1 Algorithm for the joint risk optimization (24)

S1.1 Updating u with a given U

With a given U, the matrix $(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)^\top \in \mathbb{R}^{n \times K}$ can be estimated as $U\hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{\theta}$, and W_0 can be estimated as $W = U\hat{\Sigma}_{\boldsymbol{\theta}}^{-1/2}$. For simplicity of presentation, we abuse the notation and still use $\boldsymbol{z}_i$, $i = 1, \ldots, n$ to denote these estimated unmixed independent components vectors in this section. The values of the $\boldsymbol{z}_i$'s and W will be fixed at these given values in this section when estimating $\boldsymbol{u}_0$.

Recall that $\boldsymbol{u}_0$ has coordinates $u_{0,k}$ for $k = 1, \ldots, K$ taking the form of a univariate function. In minimizing (25) with respect to $\boldsymbol{u} = (u_1, \ldots, u_K)^\top$, we represent each u_k using some pre-chosen functional basis. Let $\mathbf{s}_{u_k} \in \mathbb{R}^{l_k}$ be a prechosen l_k-dimensional functional basis with at least the second order derivative existing, and define $u_k(Z_{ik}) = \mathbf{s}_{u_k}(Z_{ik})^\top \boldsymbol{\beta}_k$, where the subscript u_k and k indicate that the basis representation and the coefficient vector $\boldsymbol{\beta}_k$ may vary with k. Using matrix notation we have

$$\boldsymbol{u}(\boldsymbol{z}_i) = \mathscr{S}(\boldsymbol{z}_i)^\top \boldsymbol{\beta}, \tag{S1.1}$$

where $\boldsymbol{z}_i = (Z_{i1}, \ldots, Z_{iK})^\top$ and

$$\mathscr{S}(\boldsymbol{z}_i) = \operatorname{diag}(\mathbf{s}_{u_1}(Z_{i1}), \cdots, \mathbf{s}_{u_k}(Z_{iK})) \in \mathbb{R}^{(\sum_{k=1}^K l_k) \times K}$$

is a block-diagonal matrix, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_k^\top)^\top \in \mathbb{R}^{\sum_{k=1}^K l_k}$. By substituting (S1.1) into (24), we can represent the empirical counterpart to the risk in (24) as a quadratic function of $\boldsymbol{\beta}$, and thus can be easily optimized. The tuning parameter λ can be chosen using the cross validation method with the loss function

$$\ell(\boldsymbol{u}) := \frac{1}{n}\sum_{i=1}^n \boldsymbol{u}(\boldsymbol{z}_i)^\top W\tilde{\Sigma}W^\top \boldsymbol{u}(\boldsymbol{z}_i) + \sum_{k=1}^K c(W_{k\cdot})\left\{\frac{2}{n}\sum_{i=1}^n u_k'(Z_{ik})\right\}.$$

This concludes the alternating step for estimating $\boldsymbol{u}$ with a given U.

S1.2 Updating U with a given u

The value of $\boldsymbol{u}$ will be fixed at the given value throughout this section. We implement gradient descent for searching for the optimal solution for U given $\boldsymbol{u}$. Because of the constraint in (25), we want to guarantee that our gradient descent updates stay within the feasible set. In fact, at the t-th step in this algorithm, we want the corresponding estimate of U, denoted $U^{(t)}$, to be in $SO(K)$. The update step derived via a standard coordinate descent approach will not generate a sequence of $U^{(t)}$'s with this property, because $SO(K)$ is not closed under addition or scalar multiplication.

Thus, we propose to design an algorithm by utilizing the method presented in Plumbley (2005), for which we now provide a brief overview.

Let us begin with some intuition. Notice first that if $U_1 \in SO(K)$ and $U_2 \in SO(K)$, then $U_1U_2 \in SO(K)$. In other words, $SO(K)$ is closed under matrix multiplication. This provides us with a way to "move" from any $U_1 \in SO(K)$ to a $U_2 \in SO(K)$: simply construct $U_3 = U_2U_1^{-1} \in SO(K)$, and then left multiply U_1 by U_3. Now that we have a notion of movement in $SO(K)$, we need to pick the best direction in which to move. As we wish to minimize an objective function over this space, the best direction should be the one of steepest descent, while remaining in $SO(K)$.

As discussed in Plumbley (2005), the *geodesic* gradient descent can be used to implement this move. Specifically, the $(t+1)$-step update takes the form $U^{(t+1)} = \exp(\eta H)U^{(t)}$, where H is a skew-symmetric matrix (i.e., $H^\top = -H$) and $\exp(*)$ denotes the matrix exponential. To give an intuition for this choice of H and for the matrix exponential, let us draw a connection between skew-symmetric matrices, and orthogonal matrices. Denote the space of $K \times K$ skew-symmetric matrices as $\mathfrak{so}(K)$. This space of matrices is known to be a *Lie Algebra*, and in particular it is closed under addition, and multiplication by a scalar. In fact, for any skew-symmetric $H \in \mathfrak{so}(K)$, $\exp(H)$ is an orthogonal matrix. The reverse is also true, that is, taking the matrix logarithm of any $U \in SO(K)$ yields a skew-symmetric matrix. Thus, if one were to convert the step t update $U^{(t)} \in SO(K)$ to a matrix $B \in \mathfrak{so}(K)$ via the matrix logarithm, one could then search for the update $U^{(t+1)}$ by $U^{(t+1)} = \exp(B + \eta H) = U^{(t)}\exp(\eta H)$ with $\eta > 0$ the step size. Thus, constructing the geodesic step in terms of movement in $\mathfrak{so}(K)$ will allow us to guarantee that when we update U, we remain in $SO(K)$. To justify such a step, notice that for the $K \times K$ identity matrix $I_K \in SO(K)$, $\log(I_K)$ is the zero matrix, meaning that $U^{(t+1)} = I_KU^{(t)}$ implies no movement in $SO(K)$. However, if we move from this origin to a non-zero matrix $\eta H \in \mathfrak{so}(K)$, this is equivalent to going from $U^{(t)} \in SO(K)$ to $\exp(\eta H)U^{(t)} \in SO(K)$.

We are now ready to introduce the specifics of determining the H needed for a gradient descent algorithm following the idea in Plumbley (2005). For our update step we need to compute the geodesic gradient of $Q_n(\boldsymbol{u}, U)$ with respect to U, denoted as $\tilde{\nabla}Q_n(\boldsymbol{u}, U)$, derived as follows. Recall that movement in $SO(K)$ can be thought of in terms of movement in $\mathfrak{so}(K)$, by starting at the origin and then moving to $\eta H \in \mathfrak{so}(K)$, and this translates to moving from an element $U^{(t)} \in SO(K)$ to $U^{(t+1)} \in SO(K)$, where $U^{(t+1)} = \exp(\eta H)U^{(t)}$. Following Plumbley (2005), we define the inner product on $\mathfrak{so}(K)$ as $\langle G, H\rangle = \frac{1}{2}\mathrm{tr}(GH)$ for any $G, H \in \mathfrak{so}(K)$. We then define, if it exists, the geodesic gradient $\tilde{\nabla}Q_n(\boldsymbol{u}, U)$ with respect to U as a skew-symmetric matrix such that for any $H \in \mathfrak{so}(K)$ where $\frac{1}{2}\|H\|_F^2 = 1$, it satisfies that

$$\frac{1}{2}\mathrm{tr}(\tilde{\nabla}Q_n(\boldsymbol{u}, U)^\top H) = \lim_{t\to 0}\frac{Q_n(\boldsymbol{u}, \exp(tH)U) - Q_n(\boldsymbol{u}, U)}{t}.$$

We next discuss the calculation of the right hand side above.

Observe that for each $k \in 1, \ldots, K$, the regular gradient of $Q_n(\boldsymbol{u}, U)$ with respect to $U_{k\cdot}$, denoted as $\nabla_{U_{k\cdot}} Q_n(\boldsymbol{u}, U)$, can be calculated as

$$\begin{aligned}
\nabla_{U_{k\cdot}} Q_n(\boldsymbol{u}, U) = \frac{1}{n}\sum_{i=1}^{n}\Big\{ & 2AU_{k\cdot}u_k^2(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i) + 2(U_{k\cdot}^\top AU_{k\cdot})u_k(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i)u_k'(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i)\tilde{\boldsymbol{\theta}}_i \\
& + \sum_{j\neq k} AU_{j*}u_k(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i)u_j(U_{j*}^\top\tilde{\boldsymbol{\theta}}_i) + 2(U_{k\cdot}^\top AU_{j*})u_j(U_{j*}^\top\tilde{\boldsymbol{\theta}}_i)u_k'(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i)\tilde{\boldsymbol{\theta}}_i \\
& + 2AU_{k\cdot}u_k'(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i) + 2(U_{k\cdot}^\top AU_{k\cdot})u_k(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i)u_k''(U_{k\cdot}^\top\tilde{\boldsymbol{\theta}}_i)\tilde{\boldsymbol{\theta}}_i\Big\},
\end{aligned} \tag{S1.2}$$

where

$$\tilde{\boldsymbol{\theta}} = \hat{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\boldsymbol{\theta}, \quad A := \hat{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}\tilde{\Sigma}\hat{\Sigma}_{\boldsymbol{\theta}}^{-\frac{1}{2}}.$$

The gradient of $Q_n(\boldsymbol{u}, U)$ with respect to U, denoted as $\nabla Q_n(\boldsymbol{u}, U)$, then, is a $K \times K$ matrix whose k-th row is $\nabla_{U_{k\cdot}} Q_n(\boldsymbol{u}, U)$. By Plumbley (2005) $\tilde{\nabla} Q_n(\boldsymbol{u}, U)$ can be calculated as

$$\tilde{\nabla} Q_n(\boldsymbol{u}, U) = \nabla Q_n(\boldsymbol{u}, U)U^\top - U\nabla Q_n(\boldsymbol{u}, U)^\top. \tag{S1.3}$$

We then set the updating direction H in our gradient descent algorithm at $U^{(t)}$ as

$$H^{(t)} = \tilde{\nabla} Q_n(\boldsymbol{u}, U^{(t)}).$$

That is, we move from $U^{(t)}$ to $U^{(t+1)}$ by using the following updating rule

$$U^{(t+1)} = \exp(\eta H^{(t)})U^{(t)}.$$

This concludes the updating step of U for a given value of $\boldsymbol{u}$.

S2 Simulation results for Section 5.1

In view of (8), we measure performance with the following empirical risk:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i))^\top\tilde{\Sigma}(\hat{\mathbf{v}}(\boldsymbol{\theta}_i) - \mathbf{v}_0(\boldsymbol{\theta}_i)). \tag{S2.4}$$

Since the true score function $\mathbf{v}_0(\boldsymbol{\theta})$ generally does not admit an explicit expression, we use the values of the score function provided by the oracle approach, discussed in Section 5, as a proxy of the values of the true score function. Such risk is calculated for each of the simulated 50 data sets, and their averages and standard errors are both presented in Tables 9 through 11.

Recall that the Y_{ik}'s, the entries of $\boldsymbol{y}_i$, are simulated from an asymmetric mixture of two distributions. In cases where both components of the mixture distributions are skewed, FEmBa$_{jointICA}$ generally performs best, though FEmBa$_T$ and FEmBa$_{fastICA}$ are competitive when the SNR is quite low. These are consistent with our intuition gained from the univariate Tweedie's formula that when SNR is low, the shrinkage approach does not give much improvement on bias reduction. KDE$_M$ performs poorly compared to ICA based FEmBa methods, which is not surprising because of the well-known curse of dimentionality in multivariate density estimation.

SNR	0.75	1.562	2.375	3.188	4
FEmBa_{NT}	75.17 (2.46)	43.16 (1.4)	30.8 (0.89)	22.32 (0.54)	16.66 (0.29)
FEmBa_{T}	**42.79** (2.5)	38.41 (1.38)	29.44 (0.89)	21.81 (0.54)	16.44 (0.28)
KDE_{M}	60.8 (2.29)	39.59 (1.29)	29.47 (0.87)	21.47 (0.53)	16.03 (0.28)
$\text{FEmBa}_{fastICA}$	47.85 (2.82)	**36.48** (1.33)	**25.36** (0.95)	**17.35** (0.6)	**12.1** (0.38)
$\text{FEmBa}_{jointICA}$	46.11 (3.28)	10.01 (1.08)	5.83 (0.59)	5.87 (0.65)	4.65 (0.42)

Table 9: Score function error for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of two Weibull distributions. All values in the table are multiplied by 10^3.

SNR	0.75	1.562	2.375	3.188	4
FEmBa_{NT}	57.59 (1.71)	18.06 (0.46)	9.43 (0.18)	6.33 (0.15)	4.54 (0.14)
FEmBa_{T}	**22.96** (1.54)	12.78 (0.48)	7.86 (0.2)	5.67 (0.15)	4.21 (0.14)
KDE_{M}	45.64 (1.57)	15.22 (0.45)	8.16 (0.19)	5.67 (0.15)	4.15 (0.13)
$\text{FEmBa}_{fastICA}$	25.52 (1.59)	11.83 (0.43)	6.91 (0.23)	5.04 (0.15)	3.81 (0.13)
$\text{FEmBa}_{jointICA}$	28.62 (2.28)	**6.03** (0.39)	**3.4** (0.32)	**4.13** (0.24)	**3.71** (0.16)

Table 10: Score function estimation for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of two Gamma distributions, one with shape and scale parameters 1 and .5, respectively, and the other with shape and scale parameters 1 and 8, respectively, shifted horizontally to the right by 10. All error values in the table are multiplied by 10^3.

SNR	0.75	1.562	2.375	3.188	4
FEmBa$_{NT}$	48.46 (1.36)	16.53 (0.72)	11.34 (0.29)	9.05 (0.14)	7.32 (0.12)
FEmBa$_{T}$	**14.13** (1.23)	11.08 (0.7)	9.58 (0.29)	8.3 (0.13)	6.93 (0.12)
KDE$_{M}$	31.06 (1.23)	13.21 (0.67)	10.11 (0.29)	8.48 (0.14)	7.03 (0.12)
FEmBa$_{fastICA}$	20.54 (1.82)	4.82 (0.43)	1.72 (0.12)	0.92 (0.06)	0.64 (0.04)
FEmBa$_{jointICA}$	37.53 (2.81)	**4.23** (0.35)	**1.54** (0.12)	**0.84** (0.05)	**0.52** (0.03)

Table 11: Score function estimation for several values of SNR. The distribution of Y_{ik} is an asymmetric mixture of a gamma distribution with shape and scale parameters 1 and .5, respectively, and a Gaussian distribution with mean and standard deviation 10 and 3, respectively. All error values in the table are multiplied by 10^3.

S3 Proofs of Results in Section 2

S3.1 Proof of Lemma 1

Proof. We drop subscript i respecting the observation id in the proof. For example, we write $X_i(t)$ as $X(t)$. The basis $\mathbf{s}^*$ is unknown, so we instead represent $X(t)$ in finite dimensional basis $\mathbf{s} \in \mathbb{R}^K$: $X^{\mathbf{s}}(t) = \mathbf{s}(t)^\top \boldsymbol{\theta}^{\mathbf{s}}$ with $\boldsymbol{\theta}^{\mathbf{s}} = \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) X(t) dt$, and we recall that $\Sigma_{\mathbf{s}} = \int \mathbf{s}(t)\mathbf{s}(t)^\top dt$. We then have

$$\begin{aligned}
\boldsymbol{\theta}^{\mathbf{s}} &= \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) X(t) dt \\
&= \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) \sum_{k=1}^{\infty} s_k^*(t) \theta_k^{\mathbf{s}^*} dt \\
&= \sum_{k=1}^{K} \theta_k^{\mathbf{s}^*} \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) s_k^*(t) dt + \sum_{k=K+1}^{\infty} \theta_k^{\mathbf{s}^*} \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) s_k^*(t) dt.
\end{aligned} \tag{S3.5}$$

As we have assumed that $\lim_{K\to\infty} \sum_{j>K} |\theta_j^{\mathbf{s}^*}| = 0$ almost surely, it follows that

$$\begin{aligned}
&\left\| \sum_{k=K+1}^{\infty} \theta_k^{\mathbf{s}^*} \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) s_k^*(t) dt \right\|_1 = \max_i \left| \sum_{k=K+1}^{\infty} \theta_k^{\mathbf{s}^*} \mathbf{e}_i^\top \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) s_k^*(t) dt \right| \\
&\leq \sum_{k=K+1}^{\infty} |\theta_k^{\mathbf{s}^*}| \max_i \|\mathbf{e}_i^\top \Sigma_{\mathbf{s}}^{-1} \int \mathbf{s}(t) s_k^*(t) dt\|_1 \leq \sum_{k=K+1}^{\infty} |\theta_k^{\mathbf{s}^*}| \max_i \left(\int (\mathbf{e}_i^\top \Sigma_{\mathbf{s}}^{-1} \mathbf{s}(t))^2 dt \int (s_k^*(t))^2 dt \right)^{1/2} \\
&\leq C \sum_{k=K+1}^{\infty} |\theta_k^{\mathbf{s}^*}| \max_i \mathbf{e}_i^\top \Sigma_{\mathbf{s}}^{-1} \mathbf{e}_i.
\end{aligned}$$

Thus, the results in the lemma follows automatically. □

S3.2 Proof of Proposition 1

We wish to minimize $\mathbb{E}(\int(m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu})^2 dt)$ with respect to $\boldsymbol{\mu}$. Let $\| * \|$ denote the L^2 norm. Let us first expand this expression via the law of total expectation:

$$\begin{aligned}
&\operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}(\int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu})^2 dt \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}(\mathbb{E}(\int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu})^2 dt | \boldsymbol{\theta}_i^{\mathbf{s}})).
\end{aligned} \tag{S3.6}$$

Thus, it suffices to minimize $\mathbb{E}[\int(m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu})^2 dt | \boldsymbol{\theta}_i^{\mathbf{s}}]$ for every $\boldsymbol{\theta}_i^{\mathbf{s}}$. We thus consider the following optimization problem:

$$\begin{aligned}
&\operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}(\|m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}\|^2 | \boldsymbol{\theta}_i^{\mathbf{s}}) \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}(\|m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}} + \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}} - \mathbf{s}(t)^\top \boldsymbol{\mu}\|^2 | \boldsymbol{\theta}_i^{\mathbf{s}}) \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \left\{ \mathbb{E}(\|\mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}} - \mathbf{s}(t)^\top \boldsymbol{\mu}\|^2 | \boldsymbol{\theta}_i^{\mathbf{s}}) + \mathbb{E}(\|m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}}\|^2 | \boldsymbol{\theta}_i^{\mathbf{s}}) \right\} \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}(\|\mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}} - \mathbf{s}(t)^\top \boldsymbol{\mu}\|^2 | \boldsymbol{\theta}_i^{\mathbf{s}}),
\end{aligned} \tag{S3.7}$$

where $\mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}}$ is the projection of the curve $m_{*i}(t)$ onto the basis $\mathbf{s}$. Standard calculations then show that the optimization problem on the last line is solved with $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{\mu}_{*i}^{\mathbf{s}} | \boldsymbol{\theta}_i^{\mathbf{s}})$, completing the proof.

S3.3 Proof of Proposition 2

The proof follows automatically from the multivariate version of Tweedie's formula and is thus omitted.

S3.4 Proof of Theorem 1

Proof. In this proof, we will use ∂_j to denote partial derivative with respect to the j-th coordinate of $\boldsymbol{\theta}^{\mathbf{s}}$. First note that by (5),

$$\begin{aligned}
\mathbb{E}\|m_{*i} - \tilde{m}_{X_i}^{\mathbf{s}}\|^2 &= \mathbb{E}[(\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i})^\top \Sigma_{\mathbf{s}} (\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i})] + \mathbb{E} \int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}})^2 dt, \\
\mathbb{E}\|m_{*i} - X_i^{\mathbf{s}}\|^2 &= \mathbb{E}[(\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \boldsymbol{\theta}_i^{\mathbf{s}})^\top \Sigma_{\mathbf{s}} (\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \boldsymbol{\theta}_i^{\mathbf{s}})] + \mathbb{E} \int (m_{*i}(t) - \mathbf{s}(t)^\top \boldsymbol{\mu}_{*i}^{\mathbf{s}})^2 dt.
\end{aligned}$$

Let $S = \Sigma_{\mathbf{s}}^{1/2}$. Recall that $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i} = \boldsymbol{\theta}_i^{\mathbf{s}} + \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i^{\mathbf{s}})$ we have

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i})^\top \Sigma_{\mathbf{s}} (\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i})] = \mathbb{E}\|S(\boldsymbol{\mu}_{*i}^{\mathbf{s}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}_i})\|_2^2 \\
&= \mathbb{E}\|S(\boldsymbol{\theta}_i^{\mathbf{s}} - \boldsymbol{\mu}_{*i}^{\mathbf{s}})\|_2^2 + \mathbb{E}\|S\Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i^{\mathbf{s}})\|_2^2 + 2\mathbb{E}[(\boldsymbol{\theta}_i^{\mathbf{s}} - \boldsymbol{\mu}_{*i}^{\mathbf{s}})^\top S^\top S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i^{\mathbf{s}})].
\end{aligned}$$

We next study $\mathbb{E}[(\boldsymbol{\theta}_i^{\mathbf{s}} - \boldsymbol{\mu}_{*i}^{\mathbf{s}})^\top S^\top S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i^{\mathbf{s}})]$. By definitions, we have $v_j(\boldsymbol{\theta})$, the jth component of $\mathbf{v}_0(\boldsymbol{\theta})$, satisfies that $v_j(\boldsymbol{\theta}) = \partial_j \log f(\boldsymbol{\theta})$, $f(\boldsymbol{\theta}) = \int f(\boldsymbol{\theta}|\boldsymbol{\mu}) f_\mu(\boldsymbol{\mu})$ and $f(\boldsymbol{\theta}|\boldsymbol{\mu}) = \phi_{\Sigma_{\boldsymbol{\gamma}}}(\boldsymbol{\theta} - \boldsymbol{\mu})$, where f_μ

is the marginal density of $\boldsymbol{\mu}_{*i}^{\mathbf{s}}$'s, $\phi_{\Sigma_\gamma}()$ is the density function of multivariate normal distribution $N(\mathbf{0}, \Sigma_{\boldsymbol{\gamma}})$, and $f(\boldsymbol{\theta}|\boldsymbol{\mu})$ is the conditional density of $\boldsymbol{\theta}_i^{\mathbf{s}}$ given $\boldsymbol{\mu}_{*i}^{\mathbf{s}}$. Let $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ be the joint density of $\boldsymbol{\theta}_i^{\mathbf{s}}$ and $\boldsymbol{\mu}_{*i}^{\mathbf{s}}$. Thus,

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\boldsymbol{\mu}) f_\mu(\boldsymbol{\mu}) = -\Sigma_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) f(\boldsymbol{\theta}|\boldsymbol{\mu}) f_\mu(\boldsymbol{\mu}) = -\Sigma_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) f(\boldsymbol{\theta}, \boldsymbol{\mu}).$$

Then it follows from the above result and $\mathbf{v}_0(\boldsymbol{\theta}) = \nabla_\theta \log f(\boldsymbol{\theta})$ that

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{\theta}_i^{\mathbf{s}} - \boldsymbol{\mu}_{*i}^{\mathbf{s}})^\top S^\top S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i^{\mathbf{s}})] &= \int (\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)^\top S^\top S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i, \boldsymbol{\mu}_i) d\boldsymbol{\theta}_i d\boldsymbol{\mu}_i \\
&= -\int (\Sigma_{\boldsymbol{\gamma}} \mathbf{v}(\boldsymbol{\theta}_i))^\top S^\top S \Sigma_{\boldsymbol{\gamma}} \nabla_{\boldsymbol{\theta}_i} f(\boldsymbol{\theta}_i, \boldsymbol{\mu}_i) d\boldsymbol{\theta}_i d\boldsymbol{\mu}_i \\
&= -\int (\Sigma_{\boldsymbol{\gamma}} S^\top S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}(\boldsymbol{\theta}_i))^\top \nabla_{\boldsymbol{\theta}_i} f(\boldsymbol{\theta}_i, \boldsymbol{\mu}_i) d\boldsymbol{\theta}_i d\boldsymbol{\mu}_i \\
&= \int \operatorname{tr}\Big((\Sigma_{\boldsymbol{\gamma}} S^\top S \Sigma_{\boldsymbol{\gamma}} \nabla_{\boldsymbol{\theta}_i}^2 \log f(\boldsymbol{\theta}_i))\Big) f(\boldsymbol{\theta}_i, \boldsymbol{\mu}_i) d\boldsymbol{\theta}_i d\boldsymbol{\mu}_i \\
&= \mathbb{E}[\operatorname{tr}(\Sigma_{\boldsymbol{\gamma}} S^\top S \Sigma_{\boldsymbol{\gamma}} \nabla_{\boldsymbol{\theta}_i}^2 \log f(\boldsymbol{\theta}_i)],
\end{aligned}$$

where in the second to last step we have used integration by parts and the assumption that $f(\boldsymbol{\theta}, \boldsymbol{\mu}) \to 0$ as $\|\boldsymbol{\theta}\|_\infty \to \infty$. Note that it is well know that

$$\mathbb{E}[\nabla_{\boldsymbol{\theta}}^2 \log f(\boldsymbol{\theta})] = -\mathbb{E}[(\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}))^\top] = -\mathbb{E}[\mathbf{v}_0(\boldsymbol{\theta})(\mathbf{v}_0(\boldsymbol{\theta}))^\top]$$

is the fisher information matrix. It follows from

$$\mathbb{E}[\operatorname{tr}(\Sigma_{\boldsymbol{\gamma}} S^\top S \Sigma_{\boldsymbol{\gamma}} \nabla_{\boldsymbol{\theta}_i}^2 \log f(\boldsymbol{\theta}_i)] = -\operatorname{tr}(\mathbb{E}[\Sigma_{\boldsymbol{\gamma}} S^\top S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i)(\mathbf{v}_0(\boldsymbol{\theta}_i))^\top]) = -\mathbb{E}\|S \Sigma_{\boldsymbol{\gamma}} \mathbf{v}_0(\boldsymbol{\theta}_i)\|_2^2$$

that the desired result is proved. This completes the proof. □

S4 Proof of Results in Section 3

Notation for this section: Recall that our data consists of random vectors $\mathcal{D} := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$ and we have $\boldsymbol{z}_i := W_0 \boldsymbol{\theta}_i$. Given estimators $\hat{W}, \hat{\boldsymbol{u}}$ constructed from data, we write

$$\mathbb{E}_{\boldsymbol{\theta}} \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}(W_0\boldsymbol{\theta})\|_2^2 := \mathbb{E}\left[\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}(W_0\boldsymbol{\theta})\|_2^2 \,\middle|\, \mathcal{D}\right]$$

as the L_2 distance between the estimators and the truth.

For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{z} \in \mathbb{R}^n$, we write $(Jf)(\boldsymbol{z}) \in \mathbb{R}^{m \times n}$ as the Jacobian of f at $\boldsymbol{z}$ and write $(D^{(2)} f)(\boldsymbol{z})$ as the second derivative tensor; for $\mathbf{x}, \boldsymbol{y} \in \mathbb{R}^n$, we use $(D^{(2)} f)(\boldsymbol{z})[\mathbf{x}, \boldsymbol{y}] \in \mathbb{R}^m$ to denote a vector where $(D^{(2)} f)(\boldsymbol{z})[\mathbf{x}, \boldsymbol{y}]_k = \mathbf{x}^\top (\nabla^{(2)} f_k)(\boldsymbol{z}) \boldsymbol{y}$. For a matrix W, we write $s_{\min}(W)$ to denote its minimum singular value. We write $\|W\|_2$ as the operator (spectral) norm of W and $\|W\|_F$ as the Frobenius norm.

Throughout the analysis, we use C, C', C_1, C_2 to represent generic universal constants. The actual values of these may change from instance to instance.

S4.1 Variable transformation

To simplify notation, we make, for the entirety of Section S4, the following transformation:

$$W_0 \leftarrow W_0 \tilde{\Sigma}^{1/2}, \quad \boldsymbol{\theta} \leftarrow \tilde{\Sigma}^{-1/2}\boldsymbol{\theta}.$$

We note that $\boldsymbol{z} = W_0\boldsymbol{\theta}$ remains invariant after the transformation. We further reparametrize

$$W \leftarrow W\tilde{\Sigma}^{1/2},$$

and, for the entirety of Section S4, we abuse notation and use the same symbols $W_0, \boldsymbol{\theta}, W$ to represent the new parametrization.

With the new parametrization, the risk function (15) has the simplified form

$$\begin{aligned} R(W, \boldsymbol{u}) &= \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 && \text{(S4.8)} \\ &= \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2\, \mathbb{E}u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2, && \text{(S4.9)} \end{aligned}$$

where $c(W_{k\cdot})$ defined in (15) becomes $\|W_{k\cdot}\|_2^2$ and the covariance of $\boldsymbol{\theta}$ is $\mathbb{E}\boldsymbol{\theta}\boldsymbol{\theta}^\top = \tilde{\Sigma}^{-1/2}\Sigma_{\boldsymbol{\theta}}\tilde{\Sigma}^{-1/2}$. The empirical risk takes the form

$$\hat{F}(W, \boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^n \|W^\top \boldsymbol{u}(W\boldsymbol{\theta}_i)\|_2^2 + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2 \frac{1}{n}\sum_{i=1}^n u_k'(W_{k\cdot}^\top \boldsymbol{\theta}_i).$$

Our estimator remains the same as in (20)

$$\begin{aligned} (\hat{W}, \hat{\boldsymbol{u}}) := & \operatorname*{argmin}_{W, u_1, \ldots u_K} \hat{F}(W, \boldsymbol{u}) \\ & \text{s.t.} \quad W \in \mathcal{W} \text{ and } u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}, \end{aligned} \qquad \text{(S4.10)}$$

where $\mathcal{F}_{b,B,m}$ is defined as in (21) but $\mathcal{W}$ is now defined as

$$\mathcal{W} := \{W \in \mathbb{R}^{K\times K} \,:\, W = U\tilde{\Sigma}^{1/2}\Sigma_{\boldsymbol{\theta}}^{1/2},\, U \in SO(K)\}.$$

We can convert the estimated $\hat{W}, \hat{\boldsymbol{u}}$ back to the original parametrization by simply setting $\hat{W} \leftarrow \hat{W}\tilde{\Sigma}^{-1/2}$ and leaving $\hat{\boldsymbol{u}}$ the same.

S4.2 Proof of Theorem 2

We adopt the notation defined as the beginning of Section S4 and the variable transformation described in Section S4.1. The following quantities appear in the analysis:

$$A_1 := \frac{C_*^2}{c_*} b^{2m+1} B \tag{S4.11}$$

$$A_2 := \left\{ \frac{C_*^4}{c_*^2 \kappa} M 2^{3m} b^{2m^2} B^{m+1} \right\}^{1/2} \tag{S4.12}$$

$$A_3 := \frac{C_* 4}{c_*^2 \kappa} M 2^{3m} b^{3m^2} B^{m+2} \tag{S4.13}$$

$$A_4 := C_*^2 b^{2m} B R_0^2 R_1 \tilde{C}. \tag{S4.14}$$

It is useful to define the following function classes:

$$\begin{aligned} \mathcal{G}_r := \Bigg\{ g \in \mathbb{R}^{(\mathbb{R}^K)} \,:\, g(\boldsymbol{\theta}) = \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\ + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - 2\sum_{k=1}^K \|W_{0k\cdot}\|_2^2 u_{0k}'(W_{0k\cdot}^\top \boldsymbol{\theta}), \\ \text{and } \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \le r^2, \\ \text{where } u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}, W \in \mathcal{W} \Bigg\}, \end{aligned} \tag{S4.15}$$

$$\mathcal{G}_r^{(b)} := \{ g_b(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \in [-\frac{b}{2}, \frac{b}{2}]^K} \,:\, g \in \mathcal{G}_r \}, \tag{S4.16}$$

where $\boldsymbol{z} = W_0\boldsymbol{\theta}$ in (S4.16) and where $\boldsymbol{u}(\boldsymbol{z}) = (u_1(Z_1), \ldots, u_K(Z_K))^\top$ for $\boldsymbol{z} = (Z_1, \ldots, Z_K)^\top \in \mathbb{R}^K$, and $\boldsymbol{u}_0$ and u_{0k} are defined analogously.

S4.3 Proof of Theorem 2

Proof. (of Theorem 2)
Recall that $\boldsymbol{z}_i = W_0\boldsymbol{\theta}_i$; we define the event

$$\mathcal{E}_b = \left\{ \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \left[-\frac{b}{2}, \frac{b}{2} \right]^K \right\}. \tag{S4.17}$$

We write $t := A_3 \big(\frac{K^{3+\frac{1}{m}} \log^2 K}{n} \big)^{\frac{m}{2m+3}}$. Also, as a short-hand, define

$$F(W, \boldsymbol{u}) := \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2 \mathbb{E} u_k'(W_{k\cdot}^\top \boldsymbol{\theta}),$$

$$\hat{F}(W, \boldsymbol{u}) := \frac{1}{n}\sum_{i=1}^n \|W^\top \boldsymbol{u}(W\boldsymbol{\theta}_i)\|_2^2 + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2 \frac{1}{n}\sum_{i=1}^n u_k'(W_{k\cdot}^\top \boldsymbol{\theta}_i).$$

By Lemma 6 and under Condition 2 A1, there exists function $\tilde{\boldsymbol{u}}_0 = (\tilde{u}_{01}, \ldots, \tilde{u}_{0K})^\top$ such that

$\tilde{u}_{0k} \in \mathcal{F}_{b,B,m}$. Morever, on event $\mathcal{E}_b$, we have that $\tilde{\boldsymbol{u}}_0(W_0\boldsymbol{\theta}_i) = \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)$ for all $i \in [n]$. Using the fact that $\hat{\boldsymbol{u}}, \hat{W}$ is the minimizer of $\hat{F}$ over $\mathcal{F}_{b,B,m}$ therefore,

$$\hat{F}(\hat{W}, \hat{\boldsymbol{u}}) \leq \hat{F}(W_0, \tilde{\boldsymbol{u}}_0) = \hat{F}(W_0, \boldsymbol{u}_0). \tag{S4.18}$$

Thus, using the fact that $R(W_0, \boldsymbol{u}_0) = 0$, we have

$$\begin{aligned}
R(\hat{W}, \hat{\boldsymbol{u}}) &= F(\hat{W}, \hat{\boldsymbol{u}}) - F(W_0, \boldsymbol{u}_0) &\text{(S4.19)}\\
&= F(\hat{W}, \hat{\boldsymbol{u}}) - \hat{F}(\hat{W}, \hat{\boldsymbol{u}}) + \underbrace{\hat{F}(\hat{W}, \hat{\boldsymbol{u}}) - \hat{F}(W_0, \boldsymbol{u}_0)}_{\leq 0} \\
&\quad + \hat{F}(W_0, \boldsymbol{u}_0) - F(W_0, \boldsymbol{u}_0) \\
&\leq \big(F(\hat{W}, \hat{\boldsymbol{u}}) - F(W_0, \boldsymbol{u}_0)\big) - \big(\hat{F}(\hat{W}, \hat{\boldsymbol{u}}) - \hat{F}(W_0, \boldsymbol{u}_0)\big) \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}} \hat{g}(\boldsymbol{\theta}) \right|, &\text{(S4.20)}
\end{aligned}$$

where we define the g–function, which is implicitly associated with a pair $(W, \boldsymbol{u})$, by

$$\begin{aligned}
g(\boldsymbol{\theta}) &:= \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&\quad + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2 u_k'(W_{k*}^\top \boldsymbol{\theta}) - \sum_{k=1}^K \|W_{0k\cdot}\|_2^2 u_{0k}'(W_{0k\cdot}^\top \boldsymbol{\theta}), &\text{(S4.21)}\\
g_b(\boldsymbol{\theta}) &:= g(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \in [-\frac{b}{2}, \frac{b}{2}]^K}, \ \text{ where } \boldsymbol{z} = W_0\boldsymbol{\theta},
\end{aligned}$$

and define $\hat{g}$ as the g–function associated with $\hat{W}$ and $\hat{\boldsymbol{u}}$.

We continue from (S4.20). On the event $\mathcal{E}_b$, it holds that $\hat{g}(\boldsymbol{\theta}_i) = \hat{g}(\boldsymbol{\theta}_i)\mathbb{1}_{\boldsymbol{z}_i \in [-\frac{b}{2}, \frac{b}{2}]^K}$ and thus,

$$\begin{aligned}
R(\hat{W}, \hat{\boldsymbol{u}}) &\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i) \mathbb{1}_{\boldsymbol{z}_i \in [-\frac{b}{2}, \frac{b}{2}]^K} - \mathbb{E}_{\boldsymbol{\theta}} \hat{g}(\boldsymbol{\theta}) \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i) \mathbb{1}_{\boldsymbol{z}_i \in [-\frac{b}{2}, \frac{b}{2}]^K} - \mathbb{E}_{\boldsymbol{\theta}} \big[\hat{g}(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \in [-\frac{b}{2}, \frac{b}{2}]^K}\big] \right| \\
&\qquad + \left| \mathbb{E}_{\boldsymbol{\theta}} \big[\hat{g}(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \notin [-\frac{b}{2}, \frac{b}{2}]^K}\big] \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i) \mathbb{1}_{\boldsymbol{z}_i \in [-\frac{b}{2}, \frac{b}{2}]^K} - \mathbb{E}_{\boldsymbol{\theta}} \big[\hat{g}(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \in [-\frac{b}{2}, \frac{b}{2}]^K}\big] \right| + 2^5 \frac{A_4 K^2}{n} \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{g}_b(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}} \hat{g}_b(\boldsymbol{\theta}) \right| + \frac{t^2}{2},
\end{aligned}$$

where the penultimate inequality follows from Proposition 10 and the final inequality follows because, under the assumption that $n^{\frac{3}{2m+3}} \geq 2^7 K R_0^2 R_1 \tilde{C}$, we have that $2^5 \frac{A_4 K^2}{n} \leq \frac{t^2}{2}$.

Let $r_1 > 0$ be defined as in Proposition 3; by taking C in condition (22) suitably large, we may satisfy the $r_1 \leq C$ condition in the statement of Proposition 3. In addition, we take $\xi = \xi_1 \wedge \xi_2 = \xi_2$ (where ξ_1 and ξ_2 are possibly dependent on $\boldsymbol{u}_0$ and defined in Proposition 7 and 8 respectively) and obtain that $r_1 \le r_0 \le 1$ where r_0 is defined in (S4.39). Define $J := \min\{j \in \mathbb{N} : 2^j t \ge r_1\}$ and define

$$\mathcal{S}_j := \big\{W \in \mathcal{W}, \boldsymbol{u} = (u_1, \dots, u_K)^\top \text{ with } u_1, \dots, u_K \in \mathcal{F} : \\ 2^{2(j-1)} t^2 \le \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \le 2^{2j} t^2\big\}. \tag{S4.22}$$

We then have by (S4.20) and Proposition 3 that

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \ge t^2, \mathcal{E}_b\big) \\
&\quad \le \mathbb{P}\Bigg(\exists W \in \mathcal{W}, u_1, \dots, u_K \in \mathcal{F} \text{ s.t. } r_1^2 \ge \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \ge t^2 \\
&\qquad\qquad \text{and } \frac{t^2}{2} + \left|\frac{1}{n}\sum_{i=1}^n g_b(\boldsymbol{\theta}_i) - \mathbb{E} g_b(\boldsymbol{\theta})\right| \ge \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2, \mathcal{E}_b\Bigg) \\
&\qquad + \mathbb{P}(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \ge r_1^2, \mathcal{E}_b) \\
&\le \sum_{j=1}^J \mathbb{P}\left(\exists W, u_1, \dots, u_K \in \mathcal{S}_j \text{ s.t. } \left|\frac{1}{n}\sum_{i=1}^n g_b(\boldsymbol{\theta}_i) - \mathbb{E} g_b(\boldsymbol{\theta})\right| \ge 2^{2(j-2)} t^2, \mathcal{E}_b\right) + \frac{2}{n^4}.
\end{aligned} \tag{S4.23}$$

We apply Corollary 2 to every term in (S4.23); we verify the conditions of Corollary 2 by letting $a = 2^{2(j-2)} t^2$ and $r = 2^j t$. Since $a = r^2/4$ and $r \le r_1 \le 1$, it is immediate that $a \le r^{2(1-\frac{1}{m})}$ and also that $a \le 8 A_2 K^{\frac{1}{2}+\frac{1}{2m}} r^{1-\frac{1}{m}}$ as $A_2, K \ge 1$.

Now, noting that $(1 - \frac{1}{m})(1 - \frac{1}{2(m-1)}) = \frac{2m-3}{2m}$, we have that

$$\begin{aligned}
t &\ge (A_3)^{\frac{2m}{2m+3}} \left(\frac{K^{3+\frac{1}{m}} \log^2 K}{n}\right)^{\frac{m}{2m+3}} \quad (\Rightarrow) \\
t^{\frac{2m+3}{2m}} &\ge \frac{1}{\sqrt{n}} A_3 (K^{\frac{3}{2}+\frac{1}{2m}} \log K) \quad (\Rightarrow) \\
2^{2(j-2)} t^2 &\ge \frac{1}{\sqrt{n}} A_3 (K^{\frac{3}{2}+\frac{1}{2m}} \log K)(2^j t)^{\frac{2m-3}{2m}}
\end{aligned}$$

Therefore, we may apply Corollary 2 to (S4.23) to obtain that there exists universal $C_1, C_2 > 0$ such that

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \ge t^2, \mathcal{E}_b\big) \\
&\quad \le \frac{2}{n^4} + \sum_{j=1}^J C_1 \exp\left(-n \frac{2^{4(j-2)} t^4}{C_2 A_2^2 K^{1+\frac{1}{m}} 2^{2j(1-\frac{1}{m})} t^{2(1-\frac{1}{m})}}\right) \\
&\quad \le \frac{2}{n^4} + \sum_{j=1}^J C_1 \exp\left(-n \frac{2^{2j} t^{\frac{2m+2}{m}}}{2^{-8} C_2 A_2^2 K^{1+\frac{1}{m}}}\right) \\
&\quad \le \frac{2}{n^4} + C_1 \exp\left(-n \frac{t^{\frac{2m+2}{m}}}{2^{-8} C_2 A_2^2 K^{1+\frac{1}{m}}}\right) \left\{1 - \exp\left(-n \frac{t^{\frac{2m+2}{m}}}{2^{-8} C_2 A_2^2 K^{1+\frac{1}{m}}}\right)\right\}^{-1} \\
&\quad \le \frac{2}{n^4} + C_1 \exp\left(-n \frac{t^{\frac{2m+2}{m}}}{2^{-8} C_2 A_2^2 K^{1+\frac{1}{m}}}\right) \le \frac{2}{n^4} + \frac{C_1}{n},
\end{aligned}$$

where the penultimate and the last inequality follows because, by taking C suitably large and noting that $A_3 \geq A_2^2$, we have $n\frac{t^{\frac{2m+2}{m}}}{2^{-8}C_2A_2^2K^{1+\frac{1}{m}}} \geq \log n$ under our condition on n. Therefore,

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top\hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top\boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \geq t^2\big) \\
&\quad\leq \mathbb{P}\big(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top\hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top\boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \geq t^2,\, \mathcal{E}_b\big) + \mathbb{P}(\mathcal{E}_b^c) \leq \frac{2}{n^4} + \frac{C_1}{n} + \frac{\tilde{C}K}{n^2}.
\end{aligned}$$

The conclusion of the Theorem follows as desired. □

Proposition 3. *Define* $r_1^2 := n^{-\frac{m-1}{2m-1}}A_1^{\frac{1}{2m-1}}b^{4m}C_*^4K^2(\log K)(\log n)$. *There exists universal constants* $C, C' > 0$ *such that if* $r_1^2 \leq C$ *and* $n^2\log n \geq C'R_0^2R_1\tilde{C}$, *then*

$$\mathbb{P}\left(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top\hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top\boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \geq r_1^2,\, \mathcal{E}_b\right) \leq \frac{2}{n^4}.$$

Proof. We work on the event $\mathcal{E}_b$. From (S4.20) in the proof of Theorem 2, we have

$$R(\hat{W}, \hat{\boldsymbol{u}}) \leq \left|\frac{1}{n}\sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}}\hat{g}(\boldsymbol{\theta})\right|. \tag{S4.24}$$

Define $r_u^2 := 2b^{2m}C_*^2BK + 2R_0^2$, then we have that for any $W \in \mathcal{W}$ and $u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}$, we have by assumption A2 and Proposition 9 that

$$\mathbb{E}\|W^\top\boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top\boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq 2\mathbb{E}\|W^\top\boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 + 2\mathbb{E}\|W_0^\top\boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq r_u^2.$$

On the event $\mathcal{E}_b$, we have further have that

$$\begin{aligned}
R(\hat{W}, \hat{\boldsymbol{u}}) &\leq \left|\frac{1}{n}\sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i)\mathbb{1}_{\boldsymbol{z}_i \in [-\frac{b}{2}, \frac{b}{2}]^K} - \mathbb{E}_{\boldsymbol{\theta}}\hat{g}(\boldsymbol{\theta})\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i)\mathbb{1}_{\boldsymbol{z}_i \in [-\frac{b}{2}, \frac{b}{2}]^K} - \mathbb{E}_{\boldsymbol{\theta}}\big[\hat{g}(\boldsymbol{\theta})\mathbb{1}_{\boldsymbol{z} \in [-\frac{b}{2}, \frac{b}{2}]^K}\big]\right| \\
&\qquad + \left|\mathbb{E}_{\boldsymbol{\theta}}\big[\hat{g}(\boldsymbol{\theta})\mathbb{1}_{\boldsymbol{z} \notin [-\frac{b}{2}, \frac{b}{2}]}\big]\right|. \\
&\leq \sup_{g_b \in \mathcal{G}_{r_u}^{(b)}}\left|\frac{1}{n}\sum_{i=1}^n g_b(\boldsymbol{\theta}_i) - \mathbb{E}g_b(\boldsymbol{\theta})\right| + 2^5\frac{A_4K^2}{n^3} \\
&\leq \sup_{g_b \in \mathcal{G}_{r_u}^{(b)}}\left|\frac{1}{n}\sum_{i=1}^n g_b(\boldsymbol{\theta}_i) - \mathbb{E}g_b(\boldsymbol{\theta})\right| + \frac{r_1^2}{2},
\end{aligned}$$

where the penultimate inequality follows from Proposition 10 and the last inequality follows because, when $n^2\log n \geq 2^6R_0^2R_1\tilde{C}$, then $2^5\frac{A_4K^2}{n^3} \leq \frac{r_1^2}{2}$.

Let $\epsilon := \frac{r_1^2}{4}$ and let $\mathcal{G}^*$ be an ϵ–L_∞ covering of $\mathcal{G}_{r_u}^b$. We have by Proposition 5 that, for some universal constant $C_1 > 0$,

$$\log |\mathcal{G}^*| \le C_1 K^2 \log \frac{KA_1}{r_1^2} + C_1 A_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} r_1^{-\frac{2}{m-1}}$$

Then, we have that, by Hoeffding's inequality (with Proposition 9) and union bound,

$$\begin{aligned}
&\mathbb{P}(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \ge r_1^2,\, \mathcal{E}_b) \\
&\le \mathbb{P}\bigg(\sup_{g_b \in \mathcal{G}_{r_u}^{(b)}} \bigg|\frac{1}{n}\sum_{i=1}^n g_b(\boldsymbol{\theta}_i) - \mathbb{E} g_b(\boldsymbol{\theta}_1)\bigg| \ge \frac{r_1^2}{2},\, \mathcal{E}_b\bigg) \\
&\le \mathbb{P}\bigg(\sup_{\tilde{g} \in \mathcal{G}^*} \bigg|\frac{1}{n}\sum_{i=1}^n \tilde{g}(\boldsymbol{\theta}_i) - \mathbb{E} \tilde{g}(\boldsymbol{\theta}_1)\bigg| \ge \frac{r_1^2}{4},\, \mathcal{E}_b\bigg) \\
&\le 2\exp\bigg(-2^{-4}\frac{nr_1^4}{b^{4m} C_*^2 B^2 K^2} + C_1 K^2 \log \frac{KA_1}{r_1^2} + C_1 A_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} r_1^{-\frac{2}{m-1}}\bigg).
\end{aligned}$$

Straightforward calculation yields that, with $r_1^2 := n^{-\frac{m-1}{2m-1}} A_1^{\frac{1}{2m-1}} b^{4m} C_*^4 K^2 (\log K)(\log n)$ there exists a universal constant $C_2 > 0$ such that if $r_1 \le C_2$, then

$$\begin{aligned}
2^{-5}\frac{nr_1^4}{b^{4m} C_*^4 B^2 K^2} &\ge 4\log n, \\
2^{-6}\frac{nr_1^4}{b^{4m} C_*^4 B^2 K^2} &\ge C_1 A_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} r_1^{-\frac{2}{m-1}}, \\
2^{-6}\frac{nr_1^4}{b^{4m} C_*^4 B^2 K^2} &\ge C_1 K^2 \log \frac{KA_1}{r_1^2}.
\end{aligned}$$

Hence, we have that

$$\mathbb{P}(\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \ge r_1^2,\, \mathcal{E}_b) \le 2\exp\big(-4\log n\big) \le \frac{2}{n^4}.$$

The proposition follows as desired.

□

S4.3.1 Proof of Corollary 1

Under condition (22) in Theorem 2, it holds that

$$\left(\frac{K^{3+\frac{1}{m}} \log^2 K}{n}\right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4 \kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4} \le r_0^2 \le \xi_2,$$

where r_0 is defined in (S4.39) and ξ_2 is defined in Proposition 8. Thus, on the event that

$$\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \le \left(\frac{K^{3+\frac{1}{m}} \log^2 K}{n}\right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4 \kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4},$$

we also have that $\mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq \xi_2$. Hence, by Proposition 8, we have

$$c_*^2 \frac{\kappa}{8} \|\tilde{W} W_0^{-1} - I_K\|_F^2 \leq \mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2.$$

With condition (22), we may apply Corollary 5 to obtain that

$$\frac{c_*^2}{12} \mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \leq \mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2.$$

The conclusion of Corollary 1 thus immediately follows.

S4.4 Proof of Theorem 3

Proof. (Proof of Theorem 3)

We define

$$t^2 = \left(\frac{K^{3+\frac{1}{m}} \log^2 K}{n} \right)^{\frac{2m}{2m+3}} \frac{C_*^8}{c_*^4 \kappa^2} M^2 2^{6m} b^{6m^2} B^{2m+4}$$

and define $\mathcal{A}$ as the event that the conclusion of Theorem 2 holds. That is,

$$\mathcal{A} := \left\{ \mathbb{E}_{\boldsymbol{\theta}}\|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq t^2 \right\}.$$

Define also $\mathcal{E}_b$ as in (S4.17). We note that $\mathbb{P}(\mathcal{A} \cap \mathcal{E}_b) \geq 1 - \frac{C'\tilde{C}}{n}$ by Theorem 2. We work on event $\mathcal{A}$ for the remainder of this proof.

Using Lemma 6 and Assumption A1, there exist functions $\tilde{\boldsymbol{u}}_0 = (\tilde{u}_{01}, \ldots, \tilde{u}_{0K})$ such that $\tilde{u}_{0k} \in \mathcal{F}_{b,B,m}$ and that, on event $\mathcal{E}_b$, we have $\tilde{\boldsymbol{u}}_0(W_0\boldsymbol{\theta}_i) = \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)$ and likewise for the first derivative. Hence, using the fact that $\hat{\boldsymbol{u}}, \hat{W}$ is the minimizer of the objective $\hat{F}(W, \boldsymbol{u})$ over $\mathcal{F}_{b,B,m}$, we have

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i)\|_2^2 + 2\sum_{k=1}^K \|\hat{W}_{k\cdot}\|_2^2 \hat{u}_k'(\hat{W}_{k\cdot}^\top \boldsymbol{\theta}_i) + \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 \right\} \\
&\qquad - \frac{2}{n} \sum_{i=1}^n \{\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i)\}^\top \{W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\} - \frac{2}{n} \sum_{i=1}^n \sum_{k=1}^K \|\hat{W}_{k\cdot}\|_2^2 \hat{u}_k'(\hat{W}_{k\cdot}\boldsymbol{\theta}_i) \\
&\leq \frac{1}{n} \sum_{i=1}^n \left\{ 2\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 + 2\sum_{k=1}^K \|W_{0k\cdot}\|_2^2 u_{0k}'(W_{0k\cdot}^\top \boldsymbol{\theta}_i) \right\} \\
&\qquad - \frac{2}{n} \sum_{i=1}^n \{\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i)\}^\top \{W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\} - \frac{2}{n} \sum_{i=1}^n \sum_{k=1}^K \|\hat{W}_{k\cdot}\|_2^2 \hat{u}_k'(\hat{W}_{k\cdot}\boldsymbol{\theta}_i) \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \hat{g}^\sharp(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}} \hat{g}^\sharp(\boldsymbol{\theta}_i) \right|,
\end{aligned}$$

where we define a new $g^\sharp$-function, again implicitly associated with a pair $(W, \boldsymbol{u})$, by

$$g^\sharp(\boldsymbol{\theta}) = 2\big(W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}) - W^\top \boldsymbol{u}(W\boldsymbol{\theta})\big)^\top W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}) \\ + 2\sum_{k=1}^K \big\{\|W_{0k\cdot}\|_2^2 u_{0k}'(W_{0k\cdot}^\top \boldsymbol{\theta}) - \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta})\big\},$$

and define $\hat{g}^\sharp$ as the $g^\sharp$-function associated with $\hat{W}$ and $\hat{\boldsymbol{u}}$. It is seen that $\mathbb{E}_{\boldsymbol{\theta}} g^\sharp(\boldsymbol{\theta}) = 0$ because $\mathbb{E}_{\boldsymbol{\theta}}[(W^\top \boldsymbol{u}(W\boldsymbol{\theta}))^\top W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})] = -\sum_{k=1}^K \|W_{k\cdot}\|^2 \mathbb{E}_{\boldsymbol{\theta}} u_k'(W_{k\cdot}\boldsymbol{\theta})$ which leads to $\mathbb{E}_{\boldsymbol{\theta}} g^\sharp(\boldsymbol{\theta}) = R(W_0, \boldsymbol{u}_0) = 0$. Define also

$$g_b^\sharp(\boldsymbol{\theta}) = g^\sharp(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z}\in[-\frac{b}{2},\frac{b}{2}]^K}, \text{ where } \boldsymbol{z} = W_0\boldsymbol{\theta}. \tag{S4.25}$$

Then, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 &\le \left|\frac{1}{n}\sum_{i=1}^n \hat{g}^\sharp(\boldsymbol{\theta}_i)\mathbb{1}_{\boldsymbol{z}_i\in[-\frac{b}{2},\frac{b}{2}]^K} - \mathbb{E}\hat{g}^\sharp(\boldsymbol{\theta})\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^n \hat{g}^\sharp(\boldsymbol{\theta}_i)\mathbb{1}_{\boldsymbol{z}_i\in[-\frac{b}{2},\frac{b}{2}]^K} - \mathbb{E}\big[\hat{g}^\sharp(\boldsymbol{\theta})\mathbb{1}_{\boldsymbol{z}\in[-\frac{b}{2},\frac{b}{2}]^K}\big]\right| \\ &\qquad + \left|\mathbb{E}\big[\hat{g}^\sharp(\boldsymbol{\theta})\mathbb{1}_{\boldsymbol{z}\notin[-\frac{b}{2},\frac{b}{2}]^K}\big]\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^n \hat{g}_b^\sharp(\boldsymbol{\theta}_i) - \mathbb{E}\big[\hat{g}_b^\sharp(\boldsymbol{\theta})\big]\right| + 2^7\frac{A_4K^2}{n} \\ &\le \left|\frac{1}{n}\sum_{i=1}^n \hat{g}_b^\sharp(\boldsymbol{\theta}_i) - \mathbb{E}\big[\hat{g}_b^\sharp(\boldsymbol{\theta})\big]\right| + \frac{t^2}{2}, \end{aligned}$$

where the penultimate inequality follows from Proposition 10 and the final inequality follows from the condition that $n^{\frac{3}{2m+3}} \ge 2^7 R_0 R_1 \tilde{C}$.

Therefore, we have

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 \ge t^2,\, \mathcal{A}\cap\mathcal{E}_b\right) \\ &\le \mathbb{P}\Bigg(\exists W \in \mathcal{W}, u_1, \ldots, u_K \in \mathcal{F}_{b,B,m} \text{ s.t. } \left|\frac{1}{n}\sum_{i=1}^n g_b^\sharp(\boldsymbol{\theta}_i) - \mathbb{E}g_b^\sharp(\boldsymbol{\theta})\right| \ge \frac{t^2}{2}, \\ &\qquad \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \le t^2,\, \mathcal{E}_b\Bigg). \end{aligned} \tag{S4.26}$$

Analogous to (S4.15), we define function class

$$\mathcal{G}_r^\sharp := \Big\{ g^\sharp \,:\, g^\sharp(\boldsymbol{\theta}) = 2\big(W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}) - W^\top \boldsymbol{u}(W\boldsymbol{\theta})\big)^\top W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}) + 2\sum_{k=1}^K \big\{ \|W_{0k\cdot}\|_2^2 u'_{0k}(W_{0k\cdot}^\top \boldsymbol{\theta}) - \|W_{k\cdot}\|_2^2 u'_k(W_{k\cdot}^\top \boldsymbol{\theta}) \big\}, \ \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \le r^2 \Big\}$$

$$\mathcal{G}_r^{\sharp(b)} := \big\{ g_b^\sharp(\boldsymbol{\theta}) = g^\sharp(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z}\in[-\frac{b}{2},\frac{b}{2}]^K} \,:\, g^\sharp \in \mathcal{G}_r^\sharp \big\}.$$

Similar to Corollary (3), we have that, for any $r > 0$, $\epsilon > 0$ and $L \ge 2\epsilon$,

$$H_B(\epsilon, \mathcal{G}_r^{\sharp(b)}, \rho_L) \le CK^2 \log \frac{KA_1}{\epsilon} + CA_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} \epsilon^{-\frac{1}{m-1}}. \tag{S4.27}$$

Since $g(\boldsymbol{\theta})$ and $g^\sharp(\boldsymbol{\theta})$ are almost identical, the proof of (S4.27) proceeds in exactly the same way as that of Corollary (3) and Proposition 5. We omit a full proof for brevity.

Likewise, similar to Corollary 4, we have that for $L = 4b^{2m}BKC_*^2$ and for $r \le r_0$ (where r_0 is defined in (S4.39)), we have

$$\rho_L^2(g_b^\sharp(\boldsymbol{\theta}) - \mathbb{E}g_b^\sharp(\boldsymbol{\theta})) \le C\frac{C_*^4}{c_*^2\kappa} M 2^{3m} b^{2m^2} B^{m+1} K^{1+\frac{1}{m}} r^{2(1-\frac{1}{m})}. \tag{S4.28}$$

The proof of (S4.28) proceeds in exactly the same way as that of Corollary (4) and Proposition 6. We omit a full proof for brevity.

Therefore, Corollary 2 holds for $\mathcal{G}_r^{\sharp(b)}$. We apply Corollary 2 with $a = t^2/2$ and $r = t$; we may verify that all the conditions of Corollary 2 hold in exactly the same way as in the proof of Theorem 2. Continuing from (S4.26), we have that

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 \ge t^2,\, \mathcal{A}\cap\mathcal{E}_b \Big) \\
&\le C\exp\Big(-\frac{nt^4}{2C_2A_2^2K^{1+\frac{1}{m}} t^{2(1-\frac{1}{m})}} \Big) \le C\exp\Big(-\frac{nt^{\frac{2m+2}{m}}}{2C_2A_2^2K^{1+\frac{1}{m}}} \Big) \le \frac{C_1}{n}.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 \ge t^2 \Big) \\
&\le \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \|\hat{W}^\top \hat{\boldsymbol{u}}(\hat{W}\boldsymbol{\theta}_i) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}_i)\|_2^2 \ge t^2,\, \mathcal{A}\cap\mathcal{E}_b \Big) + \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{E}_b^c) \\
&\le 2\frac{C_1}{n} + \frac{\tilde{C}K}{n^2}.
\end{aligned}$$

The conclusion of the Theorem follows as desired.

□

S4.5 Empirical process theory results

We first state a general theorem from empirical process theory. Let $Y_1, \ldots, Y_n$ be iid random objects taking value in some measurable space $\mathcal{Y}$ and let $\mathcal{G}$ be a class of real-valued functions $g(\cdot)$ on $\mathcal{Y}$.

For $L > 0$, define $\bar{g}(\cdot) = g(\cdot) - \mathbb{E}g(\boldsymbol{\theta})$ and

$$\rho_L^2(\bar{g}) = 2L^2 \mathbb{E}\left[e^{\frac{|\bar{g}(\boldsymbol{\theta})|}{L}} - 1 - \frac{|\bar{g}(\boldsymbol{\theta})|}{L}\right]. \tag{S4.29}$$

We note that if $\rho_L^2(\bar{g}) \leq V$, then the random variable $\bar{g}(Y_1)$ satisfies Bernstein's condition (and hence Bernstein's inequality) with variance factor V and scale factor L.

For $u > 0$, define the bracketing entropy $H_B(u, \mathcal{G}, \rho_L)$ as the logarithm of the cardinality of the smallest set $\mathcal{M}$ such that for every $g \in \mathcal{G}$, there exists $(g^U, g^L) \in \mathcal{M}$ such that $g^U \geq g \geq g^L$ and $\rho_L(g^U - g^L) \leq u$. Then, we have

Theorem S1. *(Theorem 5.11 in van de Geer (2000))*

Write $R^2 := \sup_{g \in \mathcal{G}} \rho_L^2(g)$ and suppose there exist $C_0, C_1 > 0$ such that

$$a \leq C_1 R^2 / L, \qquad a \leq 8R$$
$$a \geq \frac{C_0}{\sqrt{n}} \left(\int_0^R H_B^{1/2}(u, \mathcal{G}, \rho_L) du \vee R \right).$$

Then, there exists $C_2 > 0$ dependent only on C_0, C_1 such that

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(Y_i) - \mathbb{E}g(Y_1) \right| \geq a \right) \leq C \exp\left(-\frac{na^2}{C_2 R^2} \right).$$

We apply Theorem S1 to $\mathcal{G}_r^{(b)}$ on our problem.

Corollary 2. *Suppose $r < r_0$ with r_0 defined in* (S4.39)*. Suppose there exists $C_0, C_1 > 0$ such that*

$$a \leq C_1 r^{2(1-\frac{1}{m})}, \qquad a \leq 8A_2 K^{\frac{1}{2}+\frac{1}{2m}} r^{1-\frac{1}{m}}, \tag{S4.30}$$

$$a \geq \frac{C_0}{\sqrt{n}} \left(CA_3 (K^{\frac{3}{2}+\frac{1}{2m}} \log K) r^{(1-\frac{1}{m})(1-\frac{1}{2(m-1)})}. \right), \tag{S4.31}$$

Then, there exists $C_2 > 0$ dependent only on C_0, C_1 such that

$$\mathbb{P}\left(\sup_{g_b \in \mathcal{G}_r^{(b)}} \left| \frac{1}{n} \sum_{i=1}^n g_b(\boldsymbol{\theta}_i) - \mathbb{E}g_b(\boldsymbol{\theta}_1) \right| \geq a \right) \leq C \exp\left(-\frac{na^2}{C_2 A_2^2 K^{1+\frac{1}{m}} r^{2(1-\frac{1}{m})}} \right).$$

Proof. Write $R := \sup_{g_b \in \mathcal{G}_r^{(b)}} \rho_L^2(g_b)$. Since $r < r_0$, we have from Corollary 4 that with

$$L = 4C_*^2 b^{2m} BK,$$

we have

$$R^2 \leq C \frac{C_*^4}{c_*^2 \kappa} M 2^{3m} b^{2m^2} B^{m+1} K r^{2(1-\frac{1}{m})}.$$

We use the short-hand $A_2^2 := \frac{C_*^4}{c_*^2 \kappa} M 2^{3m} b^{2m^2} B^{m+1}$ so that $R^2 = C A_2^2 K^{1+\frac{1}{m}} r^{2(1-\frac{1}{m})}$. We note then that

$$\frac{R^2}{L} \geq C \frac{C_*^2}{c_*^2 \kappa} M 2^{3m} b^{2m(m-1)} B^m K^{\frac{1}{m}} r^{2(1-\frac{1}{m})} \geq C r^{2(1-\frac{1}{m})}.$$

Using Corollary 3 and the fact that $\sqrt{\log x} \leq \log(x+1)$ for all $x > 0$, we have that

$$H_B^{1/2}(u, \mathcal{G}_r^{(b)}, \rho_L) \leq C K \log^{1/2} \frac{A_1 K}{u} + C A_1^{\frac{1}{2(m-1)}} K^{\frac{m}{2(m-1)}} u^{-\frac{1}{2(m-1)}} \tag{S4.32}$$

$$\leq C K \left(\log \frac{A_1 K}{u} + 1 \right) + C A_1^{\frac{1}{2(m-1)}} K^{\frac{m}{2(m-1)}} u^{-\frac{1}{2(m-1)}}, \tag{S4.33}$$

Using the fact that $\int_0^R \log \frac{c}{u} du = R + R \log \frac{c}{R}$ for any $c, R > 0$ we have

$$\begin{aligned} \int_0^R H_B^{1/2}(u, \mathcal{G}_r^{(b)}, \rho_L) du &\leq C K R \left(1 + \log \frac{A_1 K}{R} \right) + C A_1^{\frac{1}{2(m-1)}} K^{\frac{m}{2(m-1)}} R^{1-\frac{1}{2(m-1)}} \\ &\leq C K^{\frac{3}{2}+\frac{1}{2m}} A_2 r^{1-\frac{1}{m}} \left(\log \frac{A_1 K^{\frac{1}{2}-\frac{1}{2m}}}{A_2 r^{1-\frac{1}{m}}} + 1 \right) \\ &\quad + C A_1^{\frac{1}{2(m-1)}} A_2^{\frac{2m-3}{2m-2}} K^{\frac{4m^2-m-3}{4m^2-4m}} r^{(1-\frac{1}{m})(1-\frac{1}{2(m-1)})} \\ &\leq C A_3 (K^{\frac{3}{2}+\frac{1}{2m}} \log K) r^{(1-\frac{1}{m})(1-\frac{1}{2(m-1)})}. \end{aligned}$$

where $A_3 = \frac{C_*^4}{c_*^2 \kappa} M 2^{3m} b^{3m^2} B^{M+2} \geq A_1 A_2 \log(A_1/A_2 + e)$. □

S4.5.1 Bracketing entropy bounds

The following covering number result is due to Birman and Solomjak (Theorem 5.2 in Birman and Solomjak (1967))

Lemma 3. *Let $\mathcal{H} \subset \mathbb{R}^{[-1,1]}$ and $T > 0$ be such that $\int_{-1}^1 h^2 \leq T^2$ and $\int_{-1}^1 |h^{(m)}|^2 \leq T^2$ for all $h \in \mathcal{H}$. Then, we have that, there exists a universal constant $C > 0$ such that, for any $\epsilon > 0$,*

$$\log N(\epsilon, \mathcal{H}, L_\infty) \leq C \epsilon^{-1/m} T^{1/m}.$$

Lemma 3 applies only to functions supported on the unit interval $[-1, 1]$, but it is straightforward to extend the result to functions supported on any interval $[-b, b]$.

Proposition 4. *Let $b \geq 1$ and $T > 0$. Let $\mathcal{H} \subset \mathbb{R}^{[-b,b]}$ such that $\int_{-b}^b h^2 \leq T^2$ and $\int_{-b}^b |h^{(m)}|^2 \leq T^2$ for all $h \in \mathcal{H}$. Then, we have that, there exists a universal constant $C > 0$ such that, for any $\epsilon > 0$,*

$$\log N(\epsilon, \mathcal{H}, L_\infty) \leq C \epsilon^{-1/m} T^{1/m} b.$$

Proof. For any $h \in \mathcal{H}$, define

$$\tilde{h}(x) = \frac{\sqrt{b}}{b^m} h(bx), \qquad \text{for any } x \in [-1,1].$$

Then, we have that $\tilde{h}^{(m)}(x) = \sqrt{b} h^{(m)}(bx)$ and that

$$\int_{-1}^{1} |\tilde{h}^{(m)}|^2 = \int_{-1}^{1} b|h^{(m)}(bx)|^2 dx = \int_{-b}^{b} |h^{(m)}(z)|^2 dz \leq T^2.$$

Similarly, we have that

$$\int_{-1}^{1} \tilde{h}^2 \leq \int_{-1}^{1} \frac{b}{b^{2m}} h(bx)^2 dx \leq \int_{-b}^{b} h(z)^2 dz \leq T^2.$$

Therefore, by Lemma 3, for any $\epsilon > 0$, we have that $\log N(\epsilon, \{\tilde{h} \,:\, h \in \mathcal{H}\}, L_\infty) \leq C\epsilon^{-1/m}$. Since for any $\tilde{h}_1, \tilde{h}_2$ such that $\|\tilde{h}_1 - \tilde{h}_2\|_\infty \leq \epsilon$, we have that

$$\sup_{z \in [-b,b]} |h_1(z) - h_2(z)| \leq \frac{b^m}{\sqrt{b}} \epsilon \leq b^m \epsilon.$$

Therefore, an $b^{-m}\epsilon$–L_∞ covering of $\{\tilde{h} \,:\, h \in \mathcal{H}\}$ is an ϵ covering of $\mathcal{H}$. The claim follows immediately. □

Now recall the definition of $\mathcal{G}_r$ and $\mathcal{G}_r^{(b)}$ from (S4.15). We may extend the covering number results to these function classes.

Proposition 5. *We have that, for any $\epsilon > 0$ and $r > 0$,*

$$\log N(\epsilon, \mathcal{G}_r^{(b)}, L_\infty) \leq CK^2 \log \frac{KA_1}{\epsilon} + CA_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} \epsilon^{-\frac{1}{m-1}},$$

where $A_1 := 24 \frac{C_*^2}{c_*} b^{2m+1} B$.

Before proving Proposition 5, we first derive an immediate corollary:

Corollary 3. *For any $r > 0$, $\epsilon > 0$, and for any $L > 2\epsilon$,*

$$H_B(\epsilon, \mathcal{G}_r^{(b)}, \rho_L) \leq CK^2 \log \frac{KA_1}{\epsilon} + CA_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} \epsilon^{-\frac{1}{m-1}},$$

where ρ_L is defined as (S4.29).

Proof. This follows given any ϵ–L_∞–cover $\mathcal{N}$, we may construct an 4ϵ–ρ_L–bracketing by $\{(g+\epsilon, g-\epsilon) \,:\, g \in \mathcal{N}\}$. To see this, note that $e^x - 1 - x \leq x^2$ for all $x \in (0,1)$; hence, when $L > 2\epsilon$, the ρ_L value of a constant function 2ϵ is at most 4ϵ. □

Proof. (of Proposition 5)
Fix $\epsilon > 0$. Since every $g_b \in \mathcal{G}_r^{(b)}$ is associated with $u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}$ and $W \in \mathcal{W}$, we will

construct a covering of

$$\mathcal{H} := \left\{ u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}, W \in \mathcal{W} \,:\, \mathbb{E}_b \| W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta}) \|_2^2 \le r^2 \right\}.$$

Let $\epsilon_1, \epsilon_2 > 0$. There exists a ϵ_1 Frobenius norm covering $\mathcal{W}^*$ of $\mathcal{W}$ of cardinality $\log |\mathcal{W}^*| \le CK^2 \log \frac{1}{\epsilon_1}$.

By Lemma 5, for any $f \in \mathcal{F}_{b,B,m}$, we have that $\int_{-b}^{b} |f(t)|^2 dt \le 2b^{2m+1}B$. Therefore, by Proposition 4, there exist $\mathcal{F}^*$ an ϵ_2–L_∞ covering of $\mathcal{F}_{b,B,m}$ and $\mathcal{F}^{**}$ an ϵ_2–L_∞ covering of $\mathcal{F}_{b,B,m-1}$ such that

$$\begin{aligned} \log |\mathcal{F}^*| &\le C\epsilon_2^{-\frac{1}{m}} (2b^{2m+1}B)^{\frac{1}{m}} b \le C\epsilon_2^{-\frac{1}{m}} b^4 B^{1/m}, \\ \log |\mathcal{F}^{**}| &\le C\epsilon_2^{-\frac{1}{m-1}} (2b^{2m-1}B)^{\frac{1}{m-1}} b \le C\epsilon_2^{-\frac{1}{m-1}} b^3 B^{\frac{1}{m-1}}. \end{aligned}$$

We will show that, for appropriately chosen ϵ_1, ϵ_2, the set $(\mathcal{F}^{**})^K \times (\mathcal{F}^*)^K \times \mathcal{W}^*$ is an ϵ–L_∞ covering of $\mathcal{G}_r^{(b)}$. To that end, let $g \in \mathcal{G}_r$ and let $W, u_1, \ldots, u_K$ be the corresponding matrix and univariate functions. Let $\tilde{W} \in \mathcal{W}^*$, $\tilde{u}_1, \ldots, \tilde{u}_K \in \mathcal{F}^*$, and $\tilde{u}_1^\sharp, \ldots, \tilde{u}_K^\sharp \in \mathcal{F}^{**}$ such that $\|W - \tilde{W}\|_F \le \epsilon_1$, that $\|u_k - \tilde{u}_k\|_\infty \le \epsilon_2$, and that $\|u_k' - \tilde{u}_k^\sharp\|_\infty \le \epsilon_2$ for all $k \in [K]$. We note that $\tilde{u}_k^\sharp$ need not be the derivative of $\tilde{u}_k$.

We write

$$\begin{aligned} \tilde{g}(\cdot) &= \|\tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\cdot)\|_2^2 - \|W_0^\top \boldsymbol{u}_0(W_0\cdot)\|_2^2 \\ &\quad + 2\sum_{k=1}^K \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(\tilde{W}_{k\cdot}^\top \cdot) - 2\sum_{k=1}^K \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top \cdot). \end{aligned}$$

Let $\boldsymbol{\theta} \in \mathbb{R}^K$ be such that $z = W_0\boldsymbol{\theta} \in [-b/2, b/2]^K$. Then, we have that

$$\begin{aligned} g(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{\theta}) &= \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|\tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2^2 \qquad \text{(Term 1)} \\ &\quad + 2\sum_{k=1}^K \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - 2\sum_{k=1}^K \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(\tilde{W}_{k\cdot}^\top \boldsymbol{\theta}) \qquad \text{(Term 2)}. \end{aligned} \tag{S4.34}$$

We bound Term 1 and Term 2 separately.

Bounding Term 1 of (S4.34):

By triangle inequality and Proposition 9, we have that

$$\begin{aligned}
&\left| \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|\tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2^2 \right| \\
&\quad \le \left\{ \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2 + \|\tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2 \right\} \|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - \tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2 \\
&\quad \le 2b^m C_* B^{1/2} K^{1/2} \|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - \tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2 \\
&\quad \le 2b^m C_* B^{1/2} K^{1/2} \Big\{ \|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - \tilde{W}^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2 \quad \text{(Term 1A)} \\
&\qquad\qquad + \|\tilde{W}^\top \boldsymbol{u}(W\boldsymbol{\theta}) - \tilde{W}^\top \tilde{\boldsymbol{u}}(W\boldsymbol{\theta})\|_2 \quad \text{(Term 1B)} \\
&\qquad\qquad + \|\tilde{W}^\top \tilde{\boldsymbol{u}}(W\boldsymbol{\theta}) - \tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2 \Big\}. \quad \text{(Term 1C)}
\end{aligned}$$

We then have by Proposition 9 that

$$\text{(Term 1A)} \le \|W - \tilde{W}\|_2 \|\boldsymbol{u}(W\boldsymbol{\theta})\|_2 \le b^m B^{1/2} K^{1/2} \epsilon_1. \tag{S4.35}$$

and that

$$\text{(Term 1B)} \le \|\tilde{W}\|_2 \|\boldsymbol{u}(W\boldsymbol{\theta}) - \tilde{\boldsymbol{u}}(W\boldsymbol{\theta})\|_2 \le C_* K^{1/2} \epsilon_2. \tag{S4.36}$$

We now turn to Term 1C. Let $k \in [K]$; by mean value theorem,

$$\begin{aligned}
|\tilde{u}_k(W_{k\cdot}^\top \boldsymbol{\theta}) - \tilde{u}_k(\tilde{W}_{k\cdot}^\top \boldsymbol{\theta})| &\le \big(\sup_t \tilde{\boldsymbol{u}}_k'(t)\big) |(W_{k\cdot} - \tilde{W}_{k\cdot})^\top \boldsymbol{\theta}| \\
&\le b^m B^{1/2} \|W_{k\cdot} - \tilde{W}_{k\cdot}\|_2 \|\boldsymbol{\theta}\|_2.
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
\|\tilde{\boldsymbol{u}}(W\boldsymbol{\theta}) - \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2^2 &\le b^{2m} B \sum_{k=1}^K \|W_{k\cdot} - \tilde{W}_{k\cdot}\|_2^2 \|\boldsymbol{\theta}\|_2^2 \\
&\le c_*^{-2} b^{2m} B K b^2 \epsilon_1^2.
\end{aligned}$$

We then obtain

$$\begin{aligned}
\text{(Term 1C)} &\le \|\tilde{W}\|_2 \|\tilde{\boldsymbol{u}}(W\boldsymbol{\theta}) - \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2 &\text{(S4.37)} \\
&\le \frac{C_*}{c_*} b^m B^{1/2} K^{1/2} b \epsilon_1. &\text{(S4.38)}
\end{aligned}$$

Putting the bounds on Term 1A, 1B, and 1C together, we have that

$$\left| \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|\tilde{W}^\top \tilde{\boldsymbol{u}}(\tilde{W}\boldsymbol{\theta})\|_2^2 \right| \le 6 \frac{C_*^2}{c_*} b^{2m+1} B K (\epsilon_1 + \epsilon_2).$$

Bounding Term 2 of (S4.34):

For any $k \in [K]$, we have that

$$\begin{aligned}\|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(\tilde{W}_{k\cdot}^\top \boldsymbol{\theta}) = \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - \|\tilde{W}_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) \quad &\text{(Term 2A)} \\ + \|\tilde{W}_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(W_{k\cdot}^\top \boldsymbol{\theta}) \quad &\text{(Term 2B)} \\ + \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(W_{k\cdot}^\top \boldsymbol{\theta}) - \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(\tilde{W}_{k\cdot}^\top \boldsymbol{\theta}) \quad &\text{(Term 2C).}\end{aligned}$$

Then, we have that

$$\begin{aligned}\text{Term 2A} &\leq \Big| \|W_{k\cdot}\|_2^2 - \|\tilde{W}_{k\cdot}\|_2^2 \Big| b^m B^{1/2} \\ &\leq \big\{ \|W_{k\cdot}\|_2 + \|\tilde{W}_{k\cdot}\|_2 \big\} \|W_{k\cdot} - \tilde{W}_{k\cdot}\|_2 b^m B^{1/2} \\ &\leq 2 C_* b^m B^{1/2} \epsilon_1.\end{aligned}$$

and that

$$\text{Term 2B} \leq C_*^2 \big| u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - \tilde{u}_k^\sharp(W_{k\cdot}^\top \boldsymbol{\theta}) \big| \leq C_*^2 \epsilon_2.$$

For Term 2C, similar to our bound on Term 1C (S4.38), we have by Lemma 5 applied on $\mathcal{F}_{b,B,m-1}$ that

$$\begin{aligned}\text{Term 2C} &\leq C_*^2 \Big\{ \sup_{t \in [-b,b]} |\tilde{u}_k^{\sharp\prime}(t)| \Big\} \|\tilde{W}_{k\cdot} - W_{k\cdot}\|_2 \|\boldsymbol{\theta}\|_2 \\ &\leq C_*^2 \Big\{ \sup_{t \in [-b,b]} |\tilde{u}_k^{\sharp\prime}(t)| \Big\} K^{1/2} c_*^{-1} b \|\tilde{W}_{k\cdot} - W_{k\cdot}\|_2 \\ &\leq \frac{C_*^2}{c_*} b^{m+1} B^{1/2} K^{1/2} \|\tilde{W}_{k\cdot} - W_{k\cdot}\|_2.\end{aligned}$$

We note that $\sum_{k=1}^K \|\tilde{W}_{k\cdot} - W_{k\cdot}\|_2 \leq K^{1/2} \epsilon_1$.

Putting our bounds on Term 2A, 2B, 2C together, we have

$$\Big| 2 \sum_{k=1}^K \big\{ \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - \|\tilde{W}_{k\cdot}\|_2^2 \tilde{u}_k^\sharp(\tilde{W}_{k\cdot}^\top \boldsymbol{\theta}) \big\} \Big| \leq 6 \frac{C_*^2}{c_*} K (b^{m+1} B^{1/2} \epsilon_1 + \epsilon_2).$$

Bounding $|g(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{\theta})|$:

Combining our bounds on Term 1 and Term 2 of (S4.34) and using the fact that $B, A_1 \geq 1$, we have we have that

$$|g(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{\theta})| \leq 12 \frac{C_*^2}{c_*} b^{2m+1} B K (\epsilon_1 + \epsilon_2).$$

By choosing $\epsilon_1 = \epsilon_2 = \epsilon \big(24 \frac{C_*^2}{c_*} b^{2m+1} B K \big)^{-1}$, we have that

$$|g(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{\theta})| \leq \epsilon.$$

Therefore, $\mathcal{W}^* \times (\mathcal{F}^*)^{\otimes K} \times (\mathcal{F}^{**})^{\otimes K}$ is an ϵ–L_∞–covering of $\mathcal{G}_r^{(b)}$ for any $r > 0$. Since

$$\begin{aligned}
&\log |\mathcal{W}^* \times (\mathcal{F}^*)^{\otimes K} \times (\mathcal{F}^{**})^{\otimes K}| \\
&\leq CK^2 \log \frac{1}{\epsilon_1} + 2CK\epsilon_2^{-\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} \\
&\leq CK^2 \log \frac{KA_1}{\epsilon} + CA_1^{\frac{1}{m-1}} b^4 B^{\frac{1}{m-1}} K^{\frac{m}{m-1}} \epsilon^{-\frac{1}{m-1}}.
\end{aligned}$$

where $A_1 := \frac{C_*^2}{c_*} b^{2m+1} B$. The conclusion of the proposition immediately follows. □

S4.5.2 Variance bound

Corollary 4. *Let $r > 0$ and suppose that $r \leq r_0$ where r_0 is defined as* (S4.39). *Then, for $L := 4b^{2m}BKC_*^2$, we have*

$$\rho_L^2(\bar{g}_b) \leq C \frac{C_*^4}{c_*^2 \kappa} M 2^{3m} b^{2m^2} B^{m+1} K^{1+\frac{1}{m}} r^{2(1-\frac{1}{m})}.$$

Proof. We note that for $L \geq \sup_{g_b \in \mathcal{G}_r^{(b)}} \sup_{\boldsymbol{\theta}} |g_b(\boldsymbol{\theta})|$ by Proposition 9 and thus, we have

$$\rho_L^2(\bar{g}_b) \leq 4\mathbb{E}\big(g_b(\boldsymbol{\theta}) - \mathbb{E}g_b(\boldsymbol{\theta})\big)^2 \leq 4\mathbb{E}\big(g_b(\boldsymbol{\theta})^2\big).$$

The desired conclusion then follows from Proposition 6. □

Define also the constant

$$r_0 := \left\{\frac{c_*^2}{12}\right\}^{1/2} \wedge \left\{2^{-11} \frac{\kappa^2}{K^2 c_1 b^{2m} B}\right\}^{1/2} \wedge \left\{\xi_1 \frac{\kappa}{8} c_*^2\right\}^{1/2} \wedge \xi_2^{1/2}, \qquad \text{(S4.39)}$$

where ξ_1, ξ_2 are defined in Proposition 7 and 8 and are possibly dependent on $\boldsymbol{u}_0$.

Proposition 6. *Let $r > 0$ and suppose $r \leq r_0$. Then, for any $g_b \in \mathcal{G}_r^{(b)}$,*

$$\mathbb{E}g_b(\boldsymbol{\theta})^2 \leq C \frac{C_*^4}{c_*^2 \kappa} M 2^{3m} b^{2m^2} B^{m+1} K^{1+\frac{1}{m}} r^{2(1-\frac{1}{m})}.$$

Proof. Recall that $g_b(\cdot)$ is defined with respect to some $W, \boldsymbol{u}$ and the fact that $g_b \in \mathcal{G}_r^{(b)}$ implies that $\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq r^2$. Since $r^2 \leq \xi_2$ and , we have by Proposition 8 that, for some $\tilde{W}, \tilde{\boldsymbol{u}} \in [W, \boldsymbol{u}]$ (see Definition in (23)),

$$\|\tilde{W} W_0^{-1} - I_K\|_F^2 \leq \frac{8}{\kappa c_*^2} \mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq \frac{8}{\kappa c_*^2} r^2. \qquad \text{(S4.40)}$$

Since $r^2 \leq 2^{-11} \frac{\kappa^2}{K^2 c_1 b^{2m} B} \wedge \xi_1 \frac{\kappa}{8} c_*^2$, we also have that

$$\|\tilde{W} W_0^{-1} - I_K\|_F^2 \leq 2^{-8} \frac{\kappa}{K^2 c_1 b^{2m} B} \wedge \xi_1.$$

Therefore, we have by Corollary 5 that

$$\sum_{k=1}^{K} \mathbb{E}\big(\tilde{u}_k(Z_k) - u_{0k}(Z_k)\big)^2 \leq \frac{12}{c_*^2}\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq \frac{12}{c_*^2} r^2. \tag{S4.41}$$

We assume without loss of generality that $\tilde{W} = W$ and $\tilde{\boldsymbol{u}} = \boldsymbol{u}$.
Now, using the definition of $g_b(\boldsymbol{\theta})$, we have that

$$\begin{aligned}
\mathbb{E}g_b(\boldsymbol{\theta})^2 \leq 4\mathbb{E}\Big\{\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2\Big\}^2 \mathbb{1}_{\boldsymbol{z}\in[-\frac{b}{2},\frac{b}{2}]^K} \quad &\text{(Term 1)}\\
+ 4\mathbb{E}\Big\{\sum_{k=1}^{K} \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top\boldsymbol{\theta}) - \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top\boldsymbol{\theta})\Big\}^2 \mathbb{1}_{\boldsymbol{z}\in[-\frac{b}{2},\frac{b}{2}]^K} \quad &\text{(Term 2)}.
\end{aligned}$$

To bound Term 1, we have by Proposition 9 that

$$\begin{aligned}
\text{Term 1} &\leq 4\mathbb{E}\bigg[\Big\{\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2 + \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2\Big\}^2 \mathbb{1}_{\boldsymbol{z}\in[-\frac{b}{2},\frac{b}{2}]^K}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|^2\bigg]\\
&\leq 16b^{2m}C_*^2 BK\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|^2 \leq 16b^{2m}C_*^2BKr^2.
\end{aligned}$$

To bound Term 2, we have that

$$\begin{aligned}
&\sum_{k=1}^{K} \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top\boldsymbol{\theta}) - \|W_{0k\cdot}\|_2^2 u_{0k}'(W_{0k\cdot}^\top\boldsymbol{\theta})\\
&\quad= \sum_{k=1}^{K} \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top\boldsymbol{\theta}) - \|W_{0k\cdot}\|_2^2 \boldsymbol{u}_k'(W_{k\cdot}^\top\boldsymbol{\theta}) \qquad \text{(2A)}\\
&\quad\quad+ \sum_{k=1}^{K} \|W_{0k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top\boldsymbol{\theta}) - \|W_{0k\cdot}\|_2^2 \boldsymbol{u}_k'(W_{0k\cdot}^\top\boldsymbol{\theta}) \qquad \text{(2B)}\\
&\quad\quad+ \sum_{k=1}^{K} \|W_{0k\cdot}\|_2^2 u_k'(W_{0k\cdot}^\top\boldsymbol{\theta}) - \|W_{0k\cdot}\|_2^2 u_{0k}'(W_{0k\cdot}^\top\boldsymbol{\theta}) \qquad \text{(2C)}.
\end{aligned}$$

Using the fact that $(\|W_{k\cdot}\|_2 + \|W_{0k*}\|_2)^2 \leq C_*^2$, using Lemma 5, and (S4.40), we have that

$$\begin{aligned}
\mathbb{E}(2\text{A})^2 &\leq \Big\{\sum_{k=1}^{K}\big\{\|W_{k\cdot}\|_2^2 - \|W_{0k\cdot}\|_2^2\big\} \sup_{t\in[-b,b]} |u_k'(t)|\Big\}^2\\
&\leq \Big\{\sum_{k=1}^{K}\big\{\|W_{k\cdot}\|_2 + \|W_{0k\cdot}\|_2\big\}\|W_{k\cdot} - W_{0k\cdot}\|_2 \sup_{t\in[-b,b]} |u_k'(t)|\Big\}^2\\
&\leq 4C_*^2 b^{2m} B\Big\{\sum_{k=1}^{K} \|W_{k\cdot} - W_{0k\cdot}\|_2\Big\}^2 \leq 4C_*^2 b^{2m} BK\Big(\frac{8}{\kappa c_*^2} r^2\Big).
\end{aligned}$$

Next, using the fact that $|u_k'(W_{k\cdot}^\top\boldsymbol{\theta}) - u_k'(W_{0k\cdot}^\top\boldsymbol{\theta})| \leq \big\{\sup_{t\in[-b,b]} |u_k''(t)|\big\}|(W_{k\cdot} - W_{0k\cdot})^\top\boldsymbol{\theta}|$, Lemma 5, and (S4.40), we have that

$$\begin{aligned}
\mathbb{E}(2\mathrm{B})^2 &\le \mathbb{E}\Big\{\sum_{k=1}^{K} C_*^2\{u_k'(W_{k\cdot}^\top\boldsymbol{\theta}) - u_k'(W_{0k*}^\top\boldsymbol{\theta})\}\Big\}^2 \\
&\le \mathbb{E}\Big\{\sum_{k=1}^{K} C_*^2\Big\{\sup_{t\in[-b,b]}|u_k''(t)|\Big\}|(W_{k\cdot} - W_{0k\cdot})^\top\boldsymbol{\theta}|\Big\}^2 \\
&\le C_*^4 b^{2m} B\,\mathbb{E}\Big\{\sum_{k=1}^{K}|\boldsymbol{e}_k^\top(WW_0^{-1} - I_K)\boldsymbol{z}|\Big\}^2 \\
&\le C_*^4 b^{2m} B\,\mathbb{E}\big\{K\|(WW_0^{-1} - I_K)\boldsymbol{z}\|_2^2\big\} \\
&\le C_*^4 b^{2m} BK\,\mathbb{E}\big\{\mathrm{tr}((WW_0^{-1} - I_K)\boldsymbol{z}\boldsymbol{z}^\top(WW_0^{-1} - I_K))\big\} \\
&\le C_*^4 b^{2m} BK\|WW_0^{-1} - I_K\|_F^2 \le C_*^4 b^{2m} BK\left(\frac{8}{\kappa c_*^2} r^2\right),
\end{aligned}$$

where the fourth inequality follows from Cauchy–Schwarz and the penultimate inequality follows from the fact that $\mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top$ is identity.

Then, we have

$$\begin{aligned}
\mathbb{E}(2\mathrm{C})^2 &\le \mathbb{E}\Big\{\sum_{k=1}^{K}\|W_{0k\cdot}\|_2^2\{u_k'(W_{0k\cdot}^\top\boldsymbol{\theta}) - u_{0k}'(W_{0k\cdot}^\top\boldsymbol{\theta})\}\Big\}^2 \\
&\le C_*^4\mathbb{E}\Big\{\sum_{k=1}^{K}\{u_k'(Z_k) - u_{0k}'(Z_k)\}\Big\}^2 \\
&\le C_*^4 K\sum_{k=1}^{K}\mathbb{E}(u_k'(Z_k) - u_{0k}'(Z_k))^2 \\
&\le C_*^4 K\sum_{k=1}^{K}\big\{2^{3m}B^{m+1}b^{2m^2}M\,\mathbb{E}(u_k(Z_k) - u_{0k}(Z_k))^2\big\}^{1-\frac{1}{m}} \\
&\le C_*^4 2^{3m}B^{m+1}b^{2m^2}MK^2\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(u_k(Z_k) - u_{0k}(Z_k))^2\right)^{1-\frac{1}{m}} \\
&\le C_*^4 2^{3m}B^{m+1}b^{2m^2}MK^{1+\frac{1}{m}}\left(\frac{12}{c_*^2}r^2\right)^{1-\frac{1}{m}},
\end{aligned}$$

where the fourth inequality follows from Corollary 6 with $f = u_k - u_{0k}$ and the fifth inequality follows from Jensen's inequality.

Combining these bounds, we have

$$\begin{aligned}\text{Term 2} &\leq 4\mathbb{E}\big\{(2\text{A}) + (2\text{B}) + (2\text{C})\big\}^2 \\ &\leq 12(\mathbb{E}(2\text{A})^2 + \mathbb{E}(2\text{B})^2 + \mathbb{E}(2\text{C})^2) \\ &\leq 32\frac{C_*^2}{c_*^2\kappa}b^{2m}BKr^2 + 8\frac{C_*^4}{c_*^2\kappa}b^{2m}BKr^2 + 12\frac{C_*^4}{c_*^2}2^{3m}B^{m+1}b^{2m^2}MK^{1+\frac{1}{m}}r^{2(1-\frac{1}{m})} \\ &\leq C\frac{C_*^4}{c_*^2\kappa}b^{2m^2}B^{m+1}2^{3m}MK^{1+\frac{1}{m}}r^{2(1-\frac{1}{m})}.\end{aligned}$$

Combining the bounds on Term 1 and Term 2, we have

$$\mathbb{E}g_b(\boldsymbol{\theta})^2 \leq C\frac{C_*^4}{c_*^2\kappa}M2^{3m}b^{2m^2}B^{m+1}K^{1+\frac{1}{m}}r^{2(1-\frac{1}{m})}.$$

The proposition thus follows.

□

S4.6 Local strong convexity of the risk

Theorem S2. *Suppose $\boldsymbol{z}$ is a mean-zero random vector in $\mathbb{R}^K$ with twice continuously differentiable score function $\boldsymbol{u}_0$. Suppose $\boldsymbol{z}$ has independent components and no two components of $\boldsymbol{z}$ are univariate Gaussians with the same variance. Let $\kappa = \inf_H \mathbb{E}\| - H\boldsymbol{u}_0(\boldsymbol{z}) + (J\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\|_2^2$ where the infimum is over all skew-symmetric H satisfying $\|H\|_F = 1$ and note that $\kappa > 0$ by Lemma 2.*

Let $\boldsymbol{u} : \mathbb{R}^K \to \mathbb{R}^K$ be of the form $\boldsymbol{u}(\boldsymbol{z}) = (u_1(Z_1), \ldots, u_K(Z_K))$. Let $\xi_1 > 0$ be defined in Proposition 7 whose value is possibly dependent on $\boldsymbol{u}_0$.

Suppose there exists some $\epsilon_0 \in (0, 1/2)$ such that, for some $0 < A < \infty$ and for any non-negative sequences $\{\nu_k\}_{k\in[K]}$ and $\{\nu'_k\}_{k\in[K]}$ that satisfy $\sum_{k=1}^K \nu_k \leq 1$ and $\sum_{k=1}^K \nu'_k \leq 1$, we have

$$\mathbb{E}\bigg\{\sup_{V\,:\,\|V-I\|_F\leq\epsilon_0}\sum_{k=1}^K\Big(u_k(V_{k\cdot}^\top\boldsymbol{z})^2 + 5\|\boldsymbol{z}\|_2^2u'_k(V_{k\cdot}^\top\boldsymbol{z})^2\nu_k + \|\boldsymbol{z}\|_2^4u''_k(V_{k\cdot}^\top\boldsymbol{z})^2\nu'_k\Big)\bigg\} \leq A.$$

If $\|V - I\|_F^2 \leq 2^{-5}\kappa A^{-1} \wedge \epsilon_0 \wedge \xi_1$, then we have that

$$\mathbb{E}\|V^\top\boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \geq \frac{1}{12}\mathbb{E}\|\boldsymbol{u}_0(\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z})\|_2^2.$$

Moreover, if in addition $\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \leq 2^{-6}\kappa$ and $\mathbb{E}\|\boldsymbol{u}'(\boldsymbol{z}) - \boldsymbol{u}'_0(\boldsymbol{z})\|_2^2 \leq 2^{-6}(\mathbb{E}\|\boldsymbol{z}\|_2^2)^{-1}\kappa$, then, we have that

$$\mathbb{E}\|V^\top\boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \geq \frac{\kappa}{4}\|V - I\|_F^2 + \frac{1}{4}\mathbb{E}\|\boldsymbol{u}_0(\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z})\|_2^2.$$

Proof. By Lemma 9, since $\|V - I\|_F^2 \leq 2^{-5}\kappa A^{-1}$, we have that $\|H\|_F^2 \leq 2^{-4}\kappa A^{-1}$.

As a short hand, we write

$$\mathbb{E}\|e^{-H}\boldsymbol{u}(e^H\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 = \mathbb{E}\|T_1 + T_2 + T_3\|_2^2,$$

where the terms T_1, T_2, T_3 are defined as

$$
\begin{aligned}
T_1 &= e^{-H}\boldsymbol{u}(e^{H}\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z}) - \big\{-H\boldsymbol{u}(\boldsymbol{z}) + (D\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\big\} \\
T_2 &= \big\{-H\boldsymbol{u}(\boldsymbol{z}) + (D\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\big\} \\
T_3 &= \boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}).
\end{aligned}
$$

Then, we have that

$$
\begin{aligned}
\mathbb{E}\|T_1 + T_2 + T_3\|_2^2 &\geq \frac{1}{2}\mathbb{E}\|T_2 + T_3\|_2^2 - \mathbb{E}\|T_1\|_2^2 \\
&= \frac{1}{2}\mathbb{E}\|T_2\|_2^2 + \frac{1}{2}\mathbb{E}\|T_3\|_2^2 + \mathbb{E}T_2^\top T_3 - \mathbb{E}\|T_1\|_2^2 \\
&\geq \frac{1}{2}\mathbb{E}\|T_3\|_2^2 + \mathbb{E}T_2^\top T_3 - \mathbb{E}\|T_1\|_2^2. \qquad \text{(S4.42)}
\end{aligned}
$$

We bound each of the terms in turn.

Bounding $\mathbb{E}\|T_1\|_2^2$. Fix $z \in \mathbb{R}^K$ and define $\phi(r, \boldsymbol{z}) = e^{-rH}\boldsymbol{u}(e^{rH}\boldsymbol{z}) - u(\boldsymbol{z})$ for $r \in [0, 1]$. Then, from the fact that the derivative of $r \mapsto e^{rH}$ is He^{rH}, we have

$$
\begin{aligned}
\phi'(r, \boldsymbol{z}) &= -He^{-rH}\boldsymbol{u}(e^{rH}\boldsymbol{z}) + e^{-rH}(J\boldsymbol{u})(e^{rH}\boldsymbol{z})He^{rH}z \\
\phi''(r, \boldsymbol{z}) &= -H^2e^{-rH}\boldsymbol{u}(e^{rH}\boldsymbol{z}) - 2He^{-rH}(D\boldsymbol{u})(e^{rH}\boldsymbol{z})He^{rH}z \\
&\quad + e^{-rH}\boldsymbol{w} + e^{-rH}(J\boldsymbol{u})(e^{rH}\boldsymbol{z})H^2e^{rH}\boldsymbol{z},
\end{aligned}
$$

where we use $\boldsymbol{w} \in \mathbb{R}^K$ to denote a vector whose k-th coordinate is

$$
\boldsymbol{w}_k = u_k''\big((e^{rH}\boldsymbol{z})_k\big)(He^{rH}\boldsymbol{z})_k^2.
$$

We note that $\phi'(0, \boldsymbol{z}) = -H\boldsymbol{u}(\boldsymbol{z}) + (J\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}$ so that $T_1 = \phi(r, \boldsymbol{z}) - \phi(0, \boldsymbol{z}) - \phi'(0, \boldsymbol{z})$. Hence, by mean value theorem, there exists $r_z \in [0, r]$ such that $T_1 = \phi''(r_z, \boldsymbol{z})$. We thus bound $\|\phi''(r, \boldsymbol{z})\|_2^2$ uniformly over $r \in [0, 1]$.

To that end, we define $\bar{H} = \frac{H}{\|H\|_F}$ and have that, using the fact that $\|\bar{H}^2 e^{-rH}\|_{\text{op}} \leq 1$,

$$
\begin{aligned}
\|-H^2e^{-rH}\boldsymbol{u}(e^{rH}\boldsymbol{z})\|_2^2 &\leq \|H\|_F^4\|-\bar{H}^2e^{-rH}\boldsymbol{u}(e^{rH}\boldsymbol{z})\|_2^2 \\
&\leq \|H\|_F^4\|\boldsymbol{u}(e^{rH}\boldsymbol{z})\|_2^2, \qquad \text{(S4.43)}
\end{aligned}
$$

and that, using the fact that $\|\bar{H}e^{rH}\|_{\text{op}} \le 1$ and Cauchy–Schwarz inequality,

$$\begin{aligned}
\|2He^{-rH}(J\boldsymbol{u})(e^{rH}\boldsymbol{z})He^{rH}z\|_2^2 &= 4\|H\|_F^4\|\bar{H}e^{-rH}(J\boldsymbol{u})(e^{rH}\boldsymbol{z})\bar{H}e^{rH}z\|_2^2 \\
&\le 4\|H\|_F^4\|(J\boldsymbol{u})(e^{rH}\boldsymbol{z})\bar{H}e^{rH}z\|_2^2 \\
&= 4\|H\|_F^4\sum_{k=1}^K u_k'((e^{rH}\boldsymbol{z})_k)^2\{(\bar{H}_{k\cdot}^\top(e^{rH}\boldsymbol{z})\}^2 \\
&\le 4\|H\|_F^4\sum_{k=1}^K u_k'((e^{rH}\boldsymbol{z})_k)^2\|e^{rH}z\|_2^2, \\
&\le 4\|H\|_F^4\|z\|_2^2\sum_{k=1}^K u_k'((e^{rH}\boldsymbol{z})_k)^2\|\bar{H}_{k\cdot}\|_2^2. \qquad \text{(S4.44)}
\end{aligned}$$

We also have

$$\begin{aligned}
\|e^{-rH}\boldsymbol{w}\|_2^2 \le \|\boldsymbol{w}\|_2^2 &= \sum_{k=1}^K u_k''((e^{rH}\boldsymbol{z})_k)^2(He^{rH}\boldsymbol{z})_k^4 \\
&\le \|H\|_F^4\sum_{k=1}^K u_k''((e^{rH}\boldsymbol{z})_k)^2\|\bar{H}_{k\cdot}\|_2^4\|e^{rH}z\|_2^4 \\
&\le \|H\|_F^4\|z\|_2^4\sum_{k=1}^K u_k''((e^{rH}\boldsymbol{z})_k)^2\|\bar{H}_{k\cdot}\|_2^4, \qquad \text{(S4.45)}
\end{aligned}$$

and that, using the fact that $\|\bar{H}^2e^{rH}\|_{k\cdot} \le \|\bar{H}^2e^{rH}\|_{\text{op}} \le 1$,

$$\begin{aligned}
\|e^{-rH}(J\boldsymbol{u})(e^{rH}\boldsymbol{z})H^2e^{rH}\boldsymbol{z}\|_2^2 &= \|H\|_F^4\|(J\boldsymbol{u})(e^{rH}\boldsymbol{z})\bar{H}^2e^{rH}\boldsymbol{z}\|_2^2 \\
&= \|H\|_F^4\sum_{k=1}^K u_k'((e^{rH}\boldsymbol{z})_k)^2\{(\bar{H}^2)_{k\cdot}^\top(e^{rH}\boldsymbol{z})\}^2 \\
&\le \|H\|_F^4\sum_{k=1}^K u_k'((e^{rH}\boldsymbol{z})_k)^2\|(\bar{H}^2)_{k\cdot}\|_2^2\|e^{rH}\boldsymbol{z}\|_2^2 \qquad \text{(S4.46)} \\
&\le \|H\|_F^4\|\boldsymbol{z}\|_2^2\sum_{k=1}^K u_k'((e^{rH}\boldsymbol{z})_k)^2\|(\bar{H}^2)_{k\cdot}\|_2^2. \qquad \text{(S4.47)}
\end{aligned}$$

We write $\nu_k = \frac{4}{5}\|\bar{H}_{k\cdot}\|_2^2 + \frac{1}{5}\|(\bar{H}^2)_{k\cdot}\|_2^2$ and observe that $\sum_{k=1}^K \nu_k = \frac{4}{5}\|\bar{H}\|_F^2 + \frac{1}{5}\|\bar{H}^2\|_F^2 \le 1$. We also write $\nu_k' = \|\bar{H}_{k\cdot}\|_2^4$ and observe that $\sum_{k=1}^K \nu_k' \le (\sum_{k=1}^K \|\bar{H}_{k\cdot}\|_2^2)^2 \le 1$. Combining this and (S4.43), (S4.44), (S4.45), (S4.47), we have that, for any $r \in [0,1]$,

$$\|\phi''(r,\boldsymbol{z})\|_2^2 \le \|H\|_F^4\left\{\sup_{V:\,\|V-I\|_F\le c_1}\sum_{k=1}^K\left(u_k(V_{k\cdot}^\top\boldsymbol{z})^2 + 5\|z\|_2^2u_k'(V_{k\cdot}^\top\boldsymbol{z})^2\nu_k + \|z\|_2^4u_k''(V_{k\cdot}^\top\boldsymbol{z})^2\nu_k'\right)\right\}. \qquad \text{(S4.48)}$$

Therefore, using the fact that $\|H\|_F^2 A \le 2^{-4}\kappa$,

$$\begin{aligned}
\mathbb{E}\|T_1\|_2^2 &= \mathbb{E}\|e^{-H}\boldsymbol{u}(e^H\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z}) - \{-H\boldsymbol{u}(\boldsymbol{z}) + (D\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\}\|_2^2 \\
&= \mathbb{E}\|\phi''(r_{\boldsymbol{z}}, \boldsymbol{z})\|_2^2 \\
&= \|H\|_F^4 \mathbb{E}\Big\{\sup_{V:\|V-I\|_F\le c_1} \sum_{k=1}^K \Big(u_k(V_{k\cdot}^\top \boldsymbol{z})^2 + 5\|\boldsymbol{z}\|_2^2 u_k'(V_{k\cdot}^\top \boldsymbol{z})^2\nu_k + \|\boldsymbol{z}\|_2^4 u_k''(V_{k\cdot}^\top \boldsymbol{z})^2\nu_k'\Big)\Big\} \\
&\le \|H\|_F^4 A \le 2^{-4}\kappa\|H\|_F^2.
\end{aligned}$$

Bounding $\mathbb{E}T_2^\top T_3$: We again write $\bar{H} = \frac{H}{\|H\|_F}$. We observe that, since H is skew–symmetric, $H_{kk} = 0$ for any k and thus $H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z})$ and $H_{k\cdot}^\top \boldsymbol{z}$ are functions of only $\{Z_j\}_{j\ne k}$. Moreover, since $\mathbb{E}\boldsymbol{u}_0(\boldsymbol{z}) = 0$ and $\mathbb{E}\boldsymbol{z} = 0$, we have

$$\mathbb{E}\big[\big((D\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\big)^\top\big(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\big)\big] \tag{S4.49}$$

$$= \sum_{k=1}^K \mathbb{E}\big[u_k'(Z_k)(H_{k\cdot}^\top Z)\big(u_k(Z_k) - u_{0k}(Z_k)\big)\big] \tag{S4.50}$$

$$= \sum_{k=1}^K \mathbb{E}\big[u_k'(Z_k)\big(u_k(Z_k) - u_{0k}(Z_k)\big)\big]\mathbb{E}[H_{k\cdot}^\top Z] = 0, \tag{S4.51}$$

and that

$$\mathbb{E}\big[\boldsymbol{u}_0(\boldsymbol{z})^\top H^\top\big(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\big)\big] \tag{S4.52}$$

$$= \sum_{k=1}^K \mathbb{E}\big[(u_k(Z_k) - u_{0k}(Z_k))(H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z}))\big] \tag{S4.53}$$

$$= \sum_{k=1}^K \mathbb{E}\big[u_k(Z_k) - u_{0k}(Z_k)\big]\mathbb{E}\big[H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z})\big] = 0. \tag{S4.54}$$

Using (S4.51) and (S4.54), we have

$$\begin{aligned}
\mathbb{E}T_2^\top T_3 &= \mathbb{E}(-H\boldsymbol{u}(\boldsymbol{z}) + (J\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z})^\top(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})) \\
&= -\mathbb{E}\big[\boldsymbol{u}(\boldsymbol{z})^\top H^\top(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}))\big] \\
&= -\mathbb{E}\big[(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}))^\top H^\top(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}))\big] \\
&= -\|H\|_F^2\mathbb{E}\big[(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}))^\top \bar{H}^\top(\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}))\big] \\
&\ge -\|H\|_F^2\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \ge -\frac{1}{4}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2.
\end{aligned}$$

Combining the bounds: We combine all the bounds with (S4.42). We have by Proposition 7 and Lemma 9 that

$$\begin{aligned}
\mathbb{E}\|V^\top\boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 &\ge \frac{1}{4}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 - 2^{-4}\kappa\|H\|_F^2 \\
&\ge \frac{1}{4}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 - 2\mathbb{E}\|V^\top\boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2.
\end{aligned}$$

The first claim of the Theorem directly follows.

Proving the second claim of the Theorem: For the second claim of the Theorem, we write

$$\mathbb{E}\|e^{-H}\boldsymbol{u}(e^{H}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 = \mathbb{E}\|T_1 + T_{2a} + T_{2b} + T_3\|_2^2,$$

where the terms T_1, T_{2a}, T_{2b}, T_4 are

$$\begin{aligned}
T_1 &= e^{-H}\boldsymbol{u}(e^{H}\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z}) - \big\{-H\boldsymbol{u}(\boldsymbol{z}) + (D\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\big\}\\
T_{2a} &= \big\{-H\boldsymbol{u}(\boldsymbol{z}) + (D\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\big\} - \big\{-H\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\big\}\\
T_{2b} &= \big\{-H\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\big\}\\
T_3 &= \boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}).
\end{aligned}$$

Using this, we have that

$$\begin{aligned}
\mathbb{E}\|T_1 + T_{2a} + T_{2b} + T_4\|_2^2 &\geq \frac{1}{2}\mathbb{E}\|T_{2b} + T_3\|_2^2 - \mathbb{E}\|T_1 + T_{2a}\|_F^2\\
&\geq \frac{1}{2}\mathbb{E}\|T_{2b}\|_2^2 + \frac{1}{2}\mathbb{E}\|T_3\|_2^2 + \mathbb{E}T_{2b}^\top T_3 - 2\mathbb{E}\|T_1\|_2^2 - 2\mathbb{E}\|T_{2a}\|_2^2.
\end{aligned}$$

Bounding $\mathbb{E}\|T_{2a}\|_2^2$.

We again write $\bar{H} := \frac{H}{\|H\|_F}$. Then, using the fact that $\bar{H}$ is skew-symmetric and hence $\bar{H}_{k\cdot}^\top \boldsymbol{z}$ is independent of Z_k, we have

$$\begin{aligned}
\mathbb{E}\|T_2\|_2^2 &\leq 2\mathbb{E}\|H(\boldsymbol{u}_0(\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z}))\|_2^2 + 2\mathbb{E}\|(J\boldsymbol{u}_0 - J\boldsymbol{u})(\boldsymbol{z})H\boldsymbol{z}\|_2^2\\
&\leq 2\|H\|_F^2\mathbb{E}\|\bar{H}(\boldsymbol{u}_0(\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z}))\|_2^2 + 2\|H\|_F^2\sum_{k=1}^{K}\mathbb{E}\big[\big\{(u_k'(Z_k) - u_{0k}'(Z_k))\bar{H}_{k\cdot}^\top \boldsymbol{z}\big\}^2\big]\\
&\leq 2\|H\|_F^2\mathbb{E}\|\boldsymbol{u}_0(\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z})\|_2^2 + 2\|H\|_F^2\sum_{k=1}^{K}\mathbb{E}\big[(u_k'(Z_k) - u_{0k}'(Z_k))^2\big]\mathbb{E}[(\bar{H}_{k\cdot}^\top \boldsymbol{z})^2]\\
&\leq 2\|H\|_F^2\mathbb{E}\|\boldsymbol{u}_0(\boldsymbol{z}) - \boldsymbol{u}(\boldsymbol{z})\|_2^2 + 2\|H\|_F^2\sum_{k=1}^{K}\mathbb{E}\big[(u_k'(Z_k) - u_{0k}'(Z_k))^2\big]\mathbb{E}\|\boldsymbol{z}\|_2^2\\
&\leq 2^{-4}\kappa\|H\|_F^2,
\end{aligned}$$

where the last inequality follows because $\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \leq 2^{-6}\kappa$ and $\mathbb{E}\|\boldsymbol{u}'(\boldsymbol{z}) - \boldsymbol{u}_0'(\boldsymbol{z})\|_2^2 \leq 2^{-6}(\mathbb{E}\|\boldsymbol{z}\|_2^2)^{-1}\kappa$.

Bounding $\mathbb{E}\|T_{2b}\|_2^2$.

We have that

$$
\begin{aligned}
\mathbb{E}\|T_3\|_2^2 &= \mathbb{E}\|\{-H\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\}\|_2^2 \\
&= \|H\|_F^2 \mathbb{E}\|\{-\bar{H}\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})\bar{H}\boldsymbol{z}\}\|_2^2 \\
&\geq \kappa \|H\|_F^2.
\end{aligned}
$$

Bounding $\mathbb{E}T_{2b}^\top T_3$.

Using the same argument for bounding $\mathbb{E}T_2^\top T_3$ (see (S4.51) and (S4.54)), we have

$$
\begin{aligned}
&\mathbb{E}\{-H\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\}^\top\{\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\} \\
&= \sum_{k=1}^K \mathbb{E}\{-(u_k(Z_k) - u_{0k}(Z_k))H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z})\} + \mathbb{E}\{u_{0k}'(Z_k)H_{k\cdot}^\top \boldsymbol{z}(u_k(Z_k) - u_{0k}(Z_k))\} \\
&= \sum_{k=1}^K \mathbb{E}\{-(u_k(Z_k) - u_{0k}(Z_k))\}\mathbb{E}\{H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z})\} + \mathbb{E}\{u_{0k}'(Z_k)(u_k(Z_k) - u_{0k}(Z_k))\}\mathbb{E}\{H_{k\cdot}^\top \boldsymbol{z}\} = 0.
\end{aligned}
$$

Combining the bounds: Putting all the previous bounds together, we have

$$
\begin{aligned}
\mathbb{E}\|T_1 + T_{2a} + T_{2b} + T_3\|_2^2 &\geq \frac{1}{2}\mathbb{E}\|T_{2b}\|_2^2 + \frac{1}{2}\mathbb{E}\|T_3\|_2^2 + \mathbb{E}T_{2b}^\top T_3 - 2\mathbb{E}\|T_1\|_2^2 - 2\mathbb{E}\|T_{2a}\|_2^2 \\
&\geq \frac{\kappa}{2}\|H\|_F^2 + \frac{1}{2}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 - \frac{\kappa}{4}\|H\|_F^2 \\
&\geq \frac{\kappa}{4}\|H\|_F^2 + \frac{1}{2}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2,
\end{aligned}
$$

as desired.

□

We state two corollaries that are immediately applicable to risk analysis of ICA estimation. Recall that we have $\mathcal{W} := \{W \in \mathbb{R}^{K\times K} : W = U\Sigma_{\boldsymbol{\theta}}^{-1/2}\tilde{\Sigma}^{1/2}, \text{ for } U \in SO(K)\}$.

Corollary 5. *Let $\boldsymbol{\theta}$ be a random vector taking value in $\mathbb{R}^K$ such that $\boldsymbol{z} := W_0\boldsymbol{\theta}$ has independent components with mean zero for some $W_0 \in \mathcal{W}$. Let $\boldsymbol{u}_0 : \mathbb{R}^K \to \mathbb{R}^K$ be the score function of Z. Define κ as Definition 2.*

Let $u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}$ and let $W \in \mathcal{W}$. Then, there exists $\xi_1 \in (0, 1/2)$ depending only on $\boldsymbol{u}_0$ such that if

$$
\|WW_0^{-1} - I\|_F^2 \leq 2^{-8}\frac{\kappa}{K^2 c_1 b^{2m} B} \wedge \xi_1, \tag{S4.55}
$$

then

$$
\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \geq \frac{c_*^2}{12}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2.
$$

Moreover, if we additionally have

$$
\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \leq 2^{-6}\kappa, \qquad \text{and} \qquad \mathbb{E}\|\boldsymbol{u}'(\boldsymbol{z}) - \boldsymbol{u}_0'(\boldsymbol{z})\|_2^2 \leq 2^{-6}\frac{\kappa}{K}. \tag{S4.56}
$$

Then, it holds that

$$\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \geq c_*^2 \left\{ \frac{\kappa}{4}\|WW_0^{-1} - I_K\|_F^2 + \frac{1}{4}\mathbb{E}\|\boldsymbol{u}(\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \right\}. \tag{S4.57}$$

Proof. (of Corollary 5)

Writing $\boldsymbol{z} := W_0\boldsymbol{\theta}$, we have that

$$\begin{aligned}
&\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&\quad = \mathbb{E}\|W_0^\top \{W_0^{-\top} W^\top \boldsymbol{u}(WW_0^{-1}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\}\|_2^2 \\
&\quad \geq c_*^2\, \mathbb{E}\|\tilde{V}^\top \boldsymbol{u}(\tilde{V}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2,
\end{aligned}$$

where we write $\tilde{V} := WW_0^{-1}$. Since W and W_0 are both in $\mathcal{W}$, $\tilde{V}$ is an orthogonal matrix.

Since $u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}$, it holds by Lemma 5 that $|u_k| \leq b^m B^{1/2}$, that $|u_k'| \leq b^{m-1}B^{1/2}$, and that $|u_k''| \leq b^{m-2}B^{1/2}$. Therefore, for any sequence $\{\nu_k\}_{k=1}^K$ and $\{\nu_k'\}_{k=1}^K$ where $\sum_k \nu_k, \sum_k \nu_k' \leq 1$, we have

$$\begin{aligned}
&\mathbb{E}\left\{ \sup_V \sum_{k=1}^K \left(u_k(V_{k\cdot}^\top \boldsymbol{z})^2 + 5\|\boldsymbol{z}\|_2^2 u_k'(V_{k\cdot}^\top \boldsymbol{z})^2 \nu_k + \|\boldsymbol{z}\|_2^4 u_k''(V_{k\cdot}^\top \boldsymbol{z})^2 \nu_k' \right) \right\} \\
&\leq \mathbb{E}\left\{ Kb^{2m}B + 5\|\boldsymbol{z}\|_2^2 b^{2m-2}B \sum_{k=1}^K \nu_k + \|\boldsymbol{z}\|_2^4 b^{2m-4} B \sum_{k=1}^K \nu_k' \right\} \\
&\leq Kb^{2m}B + 5Kb^{2m}B + \mathbb{E}(\|\boldsymbol{z}\|_2^4) b^{2m}B \\
&\leq 6Kb^{2m}B + K^2 c_1 b^{2m} B \leq 7K^2 c_1 b^{2m} B,
\end{aligned}$$

where the second inequality follows because $\mathbb{E}(\|\boldsymbol{z}\|_2^4) \leq K\mathbb{E}\sum_{k=1}^K Z_k^4 \leq K^2 c_1$.

Thus, by applying Theorem S2 with $A = 7K^2 c_1 b^{2m} B$ and taking ϵ_0 to be equal to ξ_1, the Corollary follows. ☐

Proposition 7. *Let $\boldsymbol{z} = (Z_1, \ldots, Z_K)$ be independent random variables with differentiable densities. Let $\boldsymbol{u}_0 : \mathbb{R}^K \to \mathbb{R}^K$ be the score function of Z.*

Then, there exists $\xi_1 \in (0, 1/2)$ dependent only on $\boldsymbol{u}_0$ such that, for all orthogonal matrix V satisfying $\|V - I\|_F^2 \leq \xi_1$, for all function $\boldsymbol{u} : \mathbb{R}^K \to \mathbb{R}^K$ of the form $\boldsymbol{u}(\boldsymbol{z}) = (u_1(Z_1), \ldots, u_K(Z_K))$,

$$\mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \geq \frac{\kappa}{8}\|V - I\|_F^2.$$

Proof. By Lemma 9, there exists a skew-symmetric matrix H such that $V = e^H$. Thus, we have that

$$\begin{aligned}
\mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 &= \mathbb{E}\|\boldsymbol{u}(V\boldsymbol{z}) - V\boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \\
&= \mathbb{E}\|\boldsymbol{u}(e^H\boldsymbol{z}) - e^H \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \\
&\geq \mathbb{E}\|\boldsymbol{u}^*(e^H\boldsymbol{z}) - e^H \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \\
&= \mathbb{E}\|e^{-H}\boldsymbol{u}^*(e^H\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2,
\end{aligned}$$

where $u_k^*(t) = \mathbb{E}[(e^H)_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z}) \,|\, (e^H)_{k\cdot}^\top \boldsymbol{z} = t]$. As a short-hand, write $f(H) = \mathbb{E}\|e^{-H}\boldsymbol{u}^*(e^H\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2$ so that $f(0) = 0$. We then have that $f(\cdot)$ is twice continuously differentiable and thus,

$$f(H) = f(0) + \mathrm{tr}\big((\nabla f)(0)^\top H\big) + (D^{(2)}f)(0)[H,H] + \eta,$$

where $\eta = o(\|H\|_F^2)$ and we write $(D^{(2)}f)(0)[H,H]$ to denote $\sum_{(j,k),(j',k')}(D^{(2)}f)(0)_{(j,k),(j'k')}H_{jk}H_{j'k'}$.

Suppose $(\nabla f)(0) = 0$ and $(D^{(2)}f)(0)[H,H] \geq \kappa\|H\|_F^2$. Since $\eta = o(\|H\|_F^2)$, there exists $\xi_1 > 0$ such that, if $\|V - I\|_F^2 \leq 4\|H\|_F^2 \leq \xi_1$ (first inequality following from Lemma 9), then $|\eta| \leq \frac{\kappa}{2}\|H\|_F^2$ and thus

$$f(H) \geq \frac{\kappa}{2}\|H\|_F^2 \geq \frac{\kappa}{8}\|V - I\|_F^2.$$

It remains to show that $(\nabla f)(0) = 0$ and $(D^{(2)}f)(0)[H,H] \geq \kappa\|H\|_F^2$. To this end, define $\bar{H} = \frac{H}{\|H\|_F}$ and, for $r \geq 0$, $t \in \mathbb{R}$, and $k \in [K]$,

$$\tilde{u}_k(r,t) = \mathbb{E}[(e_{k\cdot}^{r\bar{H}})^\top \boldsymbol{u}_0(\boldsymbol{z}) \,|\, (e^{r\bar{H}})_{k\cdot}^\top \boldsymbol{z} = t].$$

We note that $\tilde{u}_k(0,t) = u_{0k}(t)$ and $\tilde{u}_k(\|H\|_F, t) = u_k^*(t)$. For a vector $\boldsymbol{z} \in \mathbb{R}^K$, we define $\tilde{\boldsymbol{u}}(r,\boldsymbol{z}) = (\tilde{u}_1(r,Z_1),\ldots,\tilde{u}_K(r,Z_K))$. We now define

$$\phi(r,\boldsymbol{z}) = e^{-r\bar{H}}\tilde{\boldsymbol{u}}(r, e^{r\bar{H}}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z}).$$

Using the fact that the derivative of $r \mapsto e^{r\bar{H}}$ is $\bar{H}e^{r\bar{H}}$, we have

$$\begin{aligned}\partial_r\phi(r,\boldsymbol{z}) = -\bar{H}e^{-r\bar{H}}\tilde{\boldsymbol{u}}(r, e^{r\bar{H}}\boldsymbol{z}) + e^{-r\bar{H}}(D_z\tilde{\boldsymbol{u}})(r, e^{r\bar{H}}\boldsymbol{z})\bar{H}e^{r\bar{H}}z \\ + e^{-r\bar{H}}(\partial_r\tilde{\boldsymbol{u}})(r, e^{r\bar{H}}\boldsymbol{z}).\end{aligned}$$

In particular, writing $\boldsymbol{\psi}(\boldsymbol{z}) := (\partial_r\tilde{\boldsymbol{u}})(0,\boldsymbol{z})$, we have

$$\partial_r\phi(r,\boldsymbol{z})\big|_{r=0} = -\bar{H}\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})\bar{H}z + \boldsymbol{\psi}(\boldsymbol{z}).$$

Thus, by Fatou's lemma, we have that

$$\begin{aligned}&\liminf_{r\to 0}\frac{1}{r^2}\mathbb{E}\big\|e^{-r\bar{H}}\tilde{\boldsymbol{u}}(r, e^{r\bar{H}}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\big\|^2 \\ &\quad\leq \mathbb{E}\left[\liminf_{r\to 0}\left\|\frac{e^{-r\bar{H}}\tilde{\boldsymbol{u}}(r, e^{r\bar{H}}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})}{r}\right\|^2\right] \\ &\quad= \mathbb{E}\big\|-\bar{H}\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})\bar{H}\boldsymbol{z} + \boldsymbol{\psi}(\boldsymbol{z})\big\|_2^2 \\ &\quad\geq \mathbb{E}\big\|-\bar{H}\boldsymbol{u}_0(\boldsymbol{z}) + (D\boldsymbol{u}_0)(\boldsymbol{z})\bar{H}\boldsymbol{z}\big\|_2^2 + \mathbb{E}\|\boldsymbol{\psi}(\boldsymbol{z})\|_2^2 \\ &\qquad\qquad + 2\mathbb{E}\big[(-\bar{H}\boldsymbol{u}_0(\boldsymbol{z}) + (J\boldsymbol{u}_0)(\boldsymbol{z})\bar{H}\boldsymbol{z})^\top\boldsymbol{\psi}(\boldsymbol{z})\big] \\ &\quad\geq \kappa + \mathbb{E}\|\boldsymbol{\psi}(\boldsymbol{z})\|_2^2 \geq \kappa,\end{aligned}$$

where the penultimate inequality follows because $\bar{H}$ is skew-symmetric and $\mathbb{E}\boldsymbol{u}_0(\boldsymbol{z}) = 0$ and thus,

$$\mathbb{E}[\boldsymbol{u}_0(\boldsymbol{z})^\top\bar{H}^\top\boldsymbol{\psi}(\boldsymbol{z})] = \sum_{k\neq\ell}\bar{H}_{\ell k}\mathbb{E}[u_{0k}(Z_k)\boldsymbol{\psi}_\ell(Z_\ell)] = 0.$$

Likewise, we have that $\mathbb{E}\big[\boldsymbol{z}^\top \bar{H}^\top (D\boldsymbol{u}_0)(\boldsymbol{z})\boldsymbol{\psi}(\boldsymbol{z})\big] = 0$. To summarize, we have

$$\liminf_{r\to 0} \frac{1}{r^2}\mathbb{E}\big\|e^{-r\bar{H}}\tilde{\boldsymbol{u}}(r, e^{r\bar{H}}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\big\|^2 \geq \kappa,$$

which directly implies that $\liminf_{r\to 0} \frac{f(r\bar{H})}{r^2} \geq \kappa$. This yields that $\mathrm{tr}\big\{(\nabla f)(0)^\top \bar{H}\big\} = 0$ and that $(D^{(2)} f)(0)[\bar{H}, \bar{H}] \geq \kappa$. Since $\bar{H}$ is an arbitrary skew-symmetric matrix satisfying $\|\bar{H}\|_F = 1$, the proposition follows.

□

Proposition 8. *Let $\boldsymbol{z} = (Z_1, \ldots, Z_K)$ be independent random variables with differentiable densities and let $\boldsymbol{\theta} = W_0^{-1}\boldsymbol{z}$. Let $\boldsymbol{u}_0 : \mathbb{R}^K \to \mathbb{R}^K$ be the score function of Z. Suppose no two components of Z are Gaussians with the same variance.*

Then, there exists $\xi_2 \in (0, 1)$ dependent only on $\boldsymbol{u}_0$ (and less than ξ_1 in Proposition 7) such that if $\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq \xi_2$, then

$$\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \geq c_*^2 \frac{\kappa}{8}\|\tilde{W} W_0^{-1} - I\|_2^2,$$

for some $\tilde{W}$ in the equivalence class $[W, \boldsymbol{u}]$.

Proof. We assume without loss of generality that W minimizes $\|\tilde{W} W_0^{-1} - I_K\|_F$ among all $\tilde{W}$ in $[W, \boldsymbol{u}]$. In other words, $\|W W_0^{-1} - I_K\|_F = \min_{P \text{ signed perm.}} \|P W W_0^{-1} - I_K\|_F$.

Write $V = W W_0^{-1}$ so that

$$\begin{aligned}\mathbb{E}\|W^\top \boldsymbol{u}(W\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 &= \mathbb{E}\big\|W_0^\top\big\{V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\big\}\big\|_2^2 \\ &\geq c_*^2 \mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2.\end{aligned}$$

Define $\xi_2 := \inf\{\mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \,:\, \boldsymbol{u}, V, \text{ s.t. } \|PV - I\|_F^2 \geq \xi_1\ \forall P \text{ signed perm.}\}$ where $\xi_1 > 0$ is defined in Proposition 7 whose value is possibly dependent on $\boldsymbol{u}_0$. Since the function $V \mapsto \inf_{\boldsymbol{u}} \mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2$ is continuous and $\{V \in SO(K) \,:\, \|PV - I\|_F^2 \geq \xi_1\ \forall P \text{ signed perm.}\}$ is compact, there exists some $\check{V}$ such that $\|P\check{V} - I\|_F^2 \geq \xi_1$ for all signed permutation matrix P that attains the infimum in the definition of ξ_2. That is, $\inf_{\boldsymbol{u}} \mathbb{E}\|\check{V}^\top \boldsymbol{u}(\check{V}\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 = \xi_2$. Thus, it must be that $\xi_2 > 0$ by Maxwell's theorem.

Hence, we have that if $\mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \leq \xi_2$, then there exists a signed permutation matrix P such that $\|PV - I\|_F^2 \leq \xi_1$ which implies by Proposition 7 that

$$\mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{z}) - \boldsymbol{u}_0(\boldsymbol{z})\|_2^2 \geq \frac{\kappa}{8}\|PV - I\|_F^2.$$

The Proposition follows as desired.

□

S4.6.1 Proof of Lemma 2

For readers' convenience, we first restate Lemma 2 in the main paper.

Lemma 4. *Let $\boldsymbol{z} = (Z_1, \ldots, Z_K)$ be independent centered random variables with unit variance where Z_k has density $p_k(\cdot)$. Let $u_{0k} := \frac{p_k'(z_k)}{p_k(z_k)}$ be the score functions. Then, we have that*

$$\kappa := \inf_{\substack{H \text{ skew-symm.} \\ \|H\|_F = 1}} \mathbb{E}\| - H\boldsymbol{u}_0(\boldsymbol{z}) + (J\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\|_2^2 = 0$$

if and only if there exists a pair (j, k) such that p_j and p_k are standard Gaussian.

Proof. First suppose that

$$\inf_{\substack{H \text{ skew-symm.} \\ \|H\|_F = 1}} \mathbb{E}\| - H\boldsymbol{u}_0(\boldsymbol{z}) + (J\boldsymbol{u}_0)(\boldsymbol{z})H\boldsymbol{z}\|_2^2 = 0$$

Since $\{H : H \text{ skew-symm.}, \|H\|_F = 1\}$ is a compact set, there must exists H such that $\sum_{k=1}^K \mathbb{E}\big(-H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z}) + u_{0k}'(Z_k)H_{k\cdot}^\top \boldsymbol{z}\big)^2 = 0$. Let $k \in [K]$ be such that the k-th row of H is non-zero. Since

$$\mathbb{E}\big(-H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z}) + u_{0k}'(Z_k)H_{k\cdot}^\top \boldsymbol{z}\big)^2 = 0,$$

it must be that $-H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z}) + u_{0k}'(Z_k)H_{k\cdot}^\top \boldsymbol{z} = 0$ for almost every $\boldsymbol{z} \in \mathbb{R}^K$. The fact that $H_{kk} = 0$ implies that $u_{0k}'(Z_k)$ must be a negative constant (negativity follows from the fact that the positive linear function $x \mapsto ax$ for $a > 0$ does not correspond to the score function of any density) which we denote by $-a$ for some $a > 0$, so that $u_{0k}(Z_k) = -aZ_k$, and that, for all $j \neq k$ such that $H_{kj} \neq 0$, $u_{0j}(Z_j) = aZ_j$ a.e. This implies that $(Z_k, Z_j) \sim N(0, a^{-1}I_2)$. Since Z_j has unit variance, it must be that $a = 1$.

Now suppose there exists $j \neq k$ such that p_j, p_k are both $N(0, 1)$. Then, we have that $u_{0j}(Z_j) = -Z_j$ and $u_{0k}(Z_k) = -Z_k$. Define H as a matrix of all zeros except that $H_{jk} = 1/\sqrt{2}$ and $H_{kj} = -1/\sqrt{2}$, then we have that H is skew symmetric, that $\|H\|_F = 1$, and that

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\big(-H_{k\cdot}^\top \boldsymbol{u}_0(\boldsymbol{z}) + u_{0k}'(Z_k)H_{k\cdot}^\top \boldsymbol{z}\big)^2 \\
&\quad = \mathbb{E}\left(-\frac{Z_j}{\sqrt{2}} + \frac{Z_j}{\sqrt{2}}\right)^2 + \mathbb{E}\left(-\frac{Z_k}{\sqrt{2}} + \frac{Z_k}{\sqrt{2}}\right)^2 = 0.
\end{aligned}$$

□

S4.7 Other auxiliary results

S4.7.1 Bound on maximum value

Lemma 5. *Let $f \in \mathcal{F}_{b,B,m}$ and let $k \in \{0, 1, \ldots, m-1\}$. We then have that*

$$\sup_{t \in [-b,b]} |f^{(k)}(t)| \leq b^{m-k} B^{1/2}.$$

Proof. Let $f \in \mathcal{F}_{b,B,m}$. Using the fact that $f = 0$ outside of $[-b, b]$ and Cauchy–Schwartz inequality, we have that, for any $x \le 0$,

$$\begin{aligned}|f^{(m-1)}(x)| &= \left|\int_{-b}^{x} f^{(m)}(t)dt\right| \\ &\le \left\{\int_{-b}^{x} |f^{(m)}(t)|^2 dt \int_{-b}^{x} 1 dt\right\}^{1/2} \\ &\le B^{1/2} b^{1/2} \le bB^{1/2}.\end{aligned}$$

For $x \ge 0$, we may apply the same argument except we integrate from x to b. This establishes the lemma for the $k = m-1$ case. For the $k = m-2$ case, we use the uniform bound on $f^{(m-1)}$ to obtain, for any $x \le 0$,

$$|f^{(m-2)}(x)| = \left|\int_{-b}^{x} f^{(m-1)}(t)dt\right| \le b^2 B^{1/2}.$$

The cases where $k = m-3, \ldots, 0$ follow similarly. □

Proposition 9. *Let $u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}$ and let $W \in \mathcal{W}$. Then,*

$$\sup_{\boldsymbol{\theta}} \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 \le b^{2m} C_*^2 BK \quad \text{and} \quad \sup_{\boldsymbol{\theta}} \|\boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 \le b^{2m} BK.$$

Moreover, we have that

$$\sup_{g_b \in \mathcal{G}_r^{(b)}} \sup_{\boldsymbol{\theta}} |g_b(\boldsymbol{\theta})| \le 4b^{2m} C_*^2 BK.$$

Proof. By Lemma 5, we have, for any $\boldsymbol{\theta} \in \mathbb{R}^K$,

$$\|\boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 = \sum_{k=1}^{K} |u_k(W_{k\cdot}\boldsymbol{\theta})|^2 \le b^{2m} BK.$$

Then, directly, we also have that $\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 \le C_*^2 b^{2m} BK$ as desired for the first part of the Proposition.

Now consider the second part of the proposition. We have that

$$\begin{aligned}g(\boldsymbol{\theta}) = \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\ + 2\sum_{k=1}^{K} \|W_{k\cdot}\|_2^2 \boldsymbol{u}_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - 2\sum_{k=1}^{K} \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top \boldsymbol{\theta}).\end{aligned}$$

By Lemma 6, there exists $\tilde{u}_{01}, \ldots, \tilde{u}_{0K} \in \mathcal{F}_{b,B,m}$ such that $\tilde{u}_{0k}(t) = u_{0k}(t)$ for all $t \in [-b/2, b/2]$. Hence, using the first result of this Proposition, we have that

$$\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \mathbb{1}_{\{\boldsymbol{z} \in [-\frac{b}{2}, \frac{b}{2}]^K\}} = \|W_0^\top \tilde{\boldsymbol{u}}_0(W_0\boldsymbol{\theta})\|_2^2 \le b^{2m} C_*^2 BK. \tag{S4.58}$$

Similarly, we have by Lemma 5 that

$$|u_{0k}'(W_{0k*}^\top \boldsymbol{\theta})| \mathbb{1}_{\{z \in [-\frac{b}{2}, \frac{b}{2}]^K\}} = |\tilde{u}_{0k}'(z_k)| \le b^m B^{1/2}. \tag{S4.59}$$

Combining (S4.58) and (S4.59), we have that

$$|g(\boldsymbol{\theta})|\mathbb{1}_{\{z\in[-\frac{b}{2},\frac{b}{2}]^K\}} \le 2(b^{2m}C_*^2BK + KC_*^2b^mB^{1/2}) \le 4b^{2m}C_*^2BK.$$

The Proposition follows as desired.

□

S4.7.2 Smooth extension of a function

Lemma 6. *Let $c_0 := \max_{k\in[K]}\max_{j\in[m-1]}|u_{0k}^{(j)}(0)|$. For any $b \ge 2$, the restriction $u_{0k}\mathbb{1}_{[-\frac{b}{2},\frac{b}{2}]}$ has an extension $\tilde{u}_{0k}$ where*

1. *$u_{0k}^{(j)} = \tilde{u}_{0k}^{(j)}$ on $[-b/2, b/2]$ for all $j = 0, 1, \ldots, m$,*
2. *$\tilde{u}_{0k}$ is m–times differentiable on $\mathbb{R}$,*
3. *$\tilde{u}_{0k} = 0$ outside of $[-b, b]$, and*
4. *there exists $c_m > 0$ depending only on m such that $\left\{\int_{-b}^{b}|\tilde{u}_{0k}^{(m)}|^2\right\}^{1/2} \le 2^m c_m\left\{\int_{-b}^{b}|u_{0k}^{(m)}|^2\right\}^{1/2} + 2^m c_0$.*

In particular, if $\int_{-b}^{b}|u_{0k}^{(m)}|^2 \le \frac{B}{2^{2(m+1)}c_m^2}$ and $c_0 \le \frac{B^{1/2}}{2^{m+1}}$, then $\int_{-b}^{b}|\tilde{u}_{0k}^{(m)}|^2 \le B$.

Proof. We use a standard mollifier argument. Define the smooth cutoff function

$$\varphi_b(t) = \begin{cases} e^{-\frac{1}{1-(\frac{2t}{b}-1)^2}+1} & \text{if } b/2 \le t \le b \\ e^{-\frac{1}{1-(\frac{2t}{b}+1)^2}+1} & \text{if } -b/2 \ge t \ge -b \\ 1 & \text{if } |t| \le b/2 \\ 0 & \text{if } |t| \ge b. \end{cases}$$

We may then take the derivative to verify that φ_b is infinitely differentiable on $\mathbb{R}$ and that, there exists $c_m > 0$ depending only on m such that

$$|\varphi_b^{(j)}| \le c_m b^{-j} \quad \text{for all } j \in [m] \tag{S4.60}$$

We then construct $\tilde{u}_{0k} = u_{0k}\varphi_b$. By construction, we fulfill the first three requirements in the statement of the Lemma. To verify the last requirement, we have

$$\tilde{u}_{0k}^{(m)} = \sum_{j=0}^{m}\binom{m}{j}\varphi_b^{(j)}u_{0k}^{(m-j)}.$$

For a function f on $[-b, b]$, we write $\|f\|_{L_2(b)} = \{\int_{-b}^{b} f'(t)^2dt\}^{1/2}$. We observe that for any $j \ge 1$, by repeated applications of Lemma 7,

$$\begin{aligned}\|u_{0k}^{(m-j)}\|_{L_2(b)} &\le b\|u_{0k}^{(m-j+1)}\|_{L_2(b)} + |u_{0k}^{(m-j)}(0)| \\ &\le b^j\|u_{0k}^{(m)}\|_{L_2(b)} + \left(\sum_{s=0}^{j-1} b^s\right)c_0.\end{aligned}$$

Therefore, we have that

$$
\begin{aligned}
\|\tilde{u}_{0k}^{(m)}\|_{L_2(b)} &\le 2^m \max_{j\in\{0,1,\dots,m\}} \|\varphi_b^{(j)} u_{0k}^{(m-j)}\|_{L_2(b)} \\
&\le 2^m \left(\|\varphi_b u_{0k}^{(m)}\|_{L_2(b)} \vee \max_{j\in[m]} c_m b^{-j} \left\{ b^j \|u_{0k}^{(m)}\|_{L_2(b)} + \left(\sum_{s=0}^{j-1} b^s\right) c_0 \right\} \right) \\
&\le 2^m c_m \|u_{0k}^{(m)}\|_{L_2(b)} + 2^m c_0.
\end{aligned}
$$

The lemma follows as desired.

□

The following variation of Poincare inequality is standard.

Lemma 7. *Let $b > 0$ and let $f : [-b, b] \to \mathbb{R}$ be absolutely continuous. Write $\|f\|_{L_2(b)} = \{\int_{-b}^{b} |f'(x)|^2 dx\}^{1/2}$; we then have*

$$
\|f\|_{L_2(b)} \le b\|f'\|_{L_2(b)} + |f(0)|.
$$

Proof. Since f is absolutely continuous, we have by Fubini's theorem that

$$
\begin{aligned}
\int_{-b}^{b} (f(x) - f(0))^2 dx &= \int_{-b}^{b} \left\{ \int_0^x f'(z) dz \right\}^2 dx \\
&\le \int_{-b}^{b} \left\{ \int_0^{|x|} |f'(z)|^2 dz \cdot |x| \right\} dx \\
&\le \int_{-b}^{b} |f'(z)|^2 dz \int_{-b}^{b} |x| \, dx \\
&= b^2 \int_{-b}^{b} |f'(z)|^2 dz.
\end{aligned}
$$

Thus, $\|f - f(0)\|_{L_2(b)} \le \|f'\|_{L_2(b)}$ and the desired claim immediately follows. □

S4.7.3 Tail bound

Proposition 10. *Let $g(\boldsymbol{\theta})$ and $g^\sharp(\boldsymbol{\theta})$ be defined as* (S4.21) *and* (S4.25) *respectively. Under assumption A2 and A3, we have that*

$$
\begin{aligned}
|\mathbb{E} g(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \notin [-\frac{b}{2}, \frac{b}{2}]^K}| &\le 2^5 \frac{A_4 K^2}{n}, \\
|\mathbb{E} g^\sharp(\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{z} \notin [-\frac{b}{2}, \frac{b}{2}]^K}| &\le 2^7 \frac{A_4 K^2}{n},
\end{aligned}
$$

where $A_4 := C_*^2 b^{2m} B R_0^2 R_1 \tilde{C}$.

Proof. Recall that

$$
g(\boldsymbol{\theta}) = \|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^2 - \|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 + 2\sum_{k=1}^{K} \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta}) - 2\sum_{k=1}^{K} \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top \boldsymbol{\theta}).
$$

Therefore, we have

$$|g(\boldsymbol{\theta})|^{1+\delta_0} \le 4\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^{2+2\delta_0} + 4\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^{2+2\delta_0} \\ + 8\Big\{\sum_{k=1}^K \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta})\Big\}^{1+\delta_0} + 8\Big\{\sum_{k=1}^K \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top \boldsymbol{\theta})\Big\}^{1+\delta_0}.$$

By Proposition 9 and assumption A2, we have that

$$\mathbb{E}|g(\boldsymbol{\theta})|^{1+\delta_0} \le 4(b^{2m}C_*^2 BK)^{1+\delta_0} + 4(R_0^2)^{1+\delta_0} + 8(C_* b^m B^{1/2} K)^{1+\delta_0} + 8(C_* K R_1)^{1+\delta_0}.$$

Since $\delta_0 \ge 1/2$, we have that

$$\begin{aligned} |\mathbb{E} g(\boldsymbol{\theta})\mathbb{1}_{\boldsymbol{z}\notin[-\frac{b}{2},\frac{b}{2}]^K}| &\le \big\{\mathbb{E}|g(\boldsymbol{\theta})|^{1+\delta_0}\big\}^{\frac{1}{1+\delta_0}} \mathbb{P}(\boldsymbol{z}\notin[-b/2,b/2]^K)^{\frac{\delta_0}{1+\delta_0}} \\ &\le 2^5 b^{2m} C_*^2 B R_0^2 R_1 \tilde{C} K \left(\frac{K}{n^3}\right)^{\frac{\delta_0}{1+\delta_0}} \\ &\le 2^5 A_4 \frac{K^2}{n}. \end{aligned}$$

Similarly, we have that, for any $\boldsymbol{\theta} \in \mathbb{R}^K$,

$$\begin{aligned} |g^\sharp(\boldsymbol{\theta})|^{1+\delta_0} &\le 8\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^{2(1+\delta_0)} + 8\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^{1+\delta_0}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^{1+\delta_0} \\ &\quad + 8\Big\{\sum_{k=1}^K \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta})\Big\}^{1+\delta_0} + 8\Big\{\sum_{k=1}^K \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top \boldsymbol{\theta})\Big\}^{1+\delta_0} \\ &\le 12\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^{2(1+\delta_0)} + 4\|W^\top \boldsymbol{u}(W\boldsymbol{\theta})\|_2^{2(1+\delta_0)} \\ &\quad + 8\Big\{\sum_{k=1}^K \|W_{k\cdot}\|_2^2 u_k'(W_{k\cdot}^\top \boldsymbol{\theta})\Big\}^{1+\delta_0} + 8\Big\{\sum_{k=1}^K \|W_{0k*}\|_2^2 u_{0k}'(W_{0k*}^\top \boldsymbol{\theta})\Big\}^{1+\delta_0}. \end{aligned}$$

We can thus bound $g^\sharp$ in exactly the same way and the Proposition thus follows. □

S4.8 Unknown covariance case

For notational simplicity, we assume $\tilde{\Sigma} = I_K$; the general $\tilde{\Sigma}$ case follows in exactly the same way. We also assume without loss of generality that $\Sigma_{\boldsymbol{\theta}} = I_K$ and write $\hat{\Sigma} \equiv \hat{\Sigma}_{\boldsymbol{\theta}}$; we can prewhiten the data to ensure that this assumption holds. Then, W_0 is an orthogonal matrix and our estimator can be written as

$$\hat{V}, \hat{\boldsymbol{u}} := \mathrm{argmin}\big\{\hat{F}(V\hat{\Sigma}^{-1/2}, \boldsymbol{u}) \,:\, V \in SO(K),\, u_1, \ldots, u_K \in \mathcal{F}_{b,B,m}\big\}.$$

For simplicity, we assume that $\hat{\Sigma}^{-1/2}$ is estimated on a heldout dataset and hence independent of the dataset $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ used to estimate $\hat{V}$ and $\hat{\boldsymbol{u}}$; in practice, we find this to not be necessary. We then have that

$$
\begin{aligned}
R(\hat{V}\hat{\Sigma}^{-1/2}, \boldsymbol{u}) &= \mathbb{E}_{\boldsymbol{\theta}}\|(\hat{V}\hat{\Sigma}^{-1/2})^\top \hat{\boldsymbol{u}}(\hat{V}\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&= \mathbb{E}_{\boldsymbol{\theta}}\|(\hat{V}\hat{\Sigma}^{-1/2})^\top \hat{\boldsymbol{u}}(\hat{V}\hat{\Sigma}^{-1/2}\boldsymbol{\theta})\|_2^2 - 2\sum_{k=1}^K \mathbb{E}_{\boldsymbol{\theta}}\big\{\|(\hat{V}\hat{\Sigma}^{-1/2})_{k\cdot}\|_2^2 \hat{\boldsymbol{u}}_k'((\hat{V}\hat{\Sigma}^{-1/2})_{k\cdot}^\top \boldsymbol{\theta})\big\} \\
&\quad + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&= F(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2. \qquad (\star)
\end{aligned}
$$

By working on the event $\mathcal{E}_b := \{\boldsymbol{z}_i \in [-b/2, b/2]^K, \forall i \in [n]\}$ and using the same reasoning as (S4.18) in the proof of Theorem 2, we have that $\hat{F}(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) \le \hat{F}(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0)$. Therefore,

$$
\begin{aligned}
(\star) &= F(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) - \hat{F}(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) + \underbrace{\hat{F}(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) - \hat{F}(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0)}_{\le 0} \\
&\quad + \hat{F}(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) - F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) \\
&\quad + F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&\le \big(F(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) - F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0)\big) - \big(\hat{F}(\hat{V}\hat{\Sigma}^{-1/2}, \hat{\boldsymbol{u}}) - \hat{F}(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0)\big) \\
&\quad + F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&\le \left|\frac{1}{n}\sum_{i=1}^n \hat{g}(\boldsymbol{\theta}_i) - \mathbb{E}\hat{g}(\boldsymbol{\theta}_1)\right| \\
&\quad + F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2, \qquad \text{(S4.61)}
\end{aligned}
$$

where we associate a pair $(V, \boldsymbol{u})$ with a modified g-function:

$$
\begin{aligned}
g(\boldsymbol{\theta}_i) &= \|(V\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}_i)\|_2^2 - \|(W_0\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}_0(W_0\hat{\Sigma}^{-1/2}\boldsymbol{\theta}_i)\|_2^2 \\
&\quad + 2\sum_{k=1}^K \|(V\hat{\Sigma}^{-1/2})_{k\cdot}\|_2^2 u_k'((V\hat{\Sigma}^{-1/2})_{k\cdot}^\top \boldsymbol{\theta}_i) - 2\sum_{k=1}^K \|(W_0\hat{\Sigma}^{-1/2})_{k\cdot}\|_2^2 u_{0k}'((W_0\hat{\Sigma}^{-1/2})_{k\cdot}^\top \boldsymbol{\theta}_i),
\end{aligned}
$$

and we define $\hat{g}$ as the g-function associated with $\hat{V}, \hat{\boldsymbol{u}}$.

To bound the extra approximation error term $F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2$ in (S4.61), we use integration by parts to get

$$
\begin{aligned}
&F(W_0\hat{\Sigma}^{-1/2}, \boldsymbol{u}_0) + \mathbb{E}\|W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \\
&= \mathbb{E}\|(W_0\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}_0(W_0\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2.
\end{aligned}
$$

Using Lemma 8 (see also Remark 6) and an additional integrability condition on u_{0k}', we see that this is of order $\frac{K^3}{n}$ which is negligible compared to the rate of $(\frac{K^{3+\frac{1}{m}}\log^2 K}{n})^{2m/(2m+3)}$ given in Theorem 2.

The statement and proof of the covering number bound (Proposition 5) require no modification. The bound on $\mathbb{E}g_b(\boldsymbol{\theta})^2$ (Proposition 6) also holds but the proof requires an additional step where we use Lemma 8 and the fact that $r^2 \geq O(K^3/n)$ to show that

$$\begin{aligned} &\mathbb{E}\|(V\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq r^2 \\ (\Rightarrow) \qquad &\mathbb{E}\|V^\top \boldsymbol{u}(V\boldsymbol{\theta}) - W_0^\top \boldsymbol{u}_0(W_0\boldsymbol{\theta})\|_2^2 \leq 4r^2. \end{aligned}$$

The remainder of the proof of Proposition 6 goes through without issue.

The following Lemma is used in the discussion in this section.

Lemma 8. *Let $V \in SO(K)$ and $\boldsymbol{u} : \mathbb{R}^K \to \mathbb{R}^K$ be of the form $\boldsymbol{u}(\boldsymbol{z}) = (u_1(Z_1), \ldots, u_K(Z_K))$. If there exists $B_1 > 0$ such that*

$$B_1 \geq \left\{\mathbb{E}\|\boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta})\|_2^2\right\} \vee \left\{\sum_{k=1}^K \mathbb{E}\left[\sup_{\lambda\in[0,1]} u_k'(V(\lambda I_K + (1-\lambda)\hat{\Sigma}^{-1/2})\boldsymbol{\theta})^2\|\boldsymbol{\theta}\|_2^2\right]\right\},$$

then

$$\mathbb{E}\|(V\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - V^\top \boldsymbol{u}(V\boldsymbol{\theta})\|_2^2 \leq 2\|\hat{\Sigma}^{-1/2} - I\|_2^2 B_1.$$

Remark 6. *In Lemma 8, if $u_1, \ldots, u_K$ are in $\mathcal{F}_{b,B,m}$, then B_1 is of order K^2 (suppressing dependence on b, B, etc). If $\|\hat{\Sigma}^{-1/2} - I\|_2^2 = O_p(K/n)$, then we have that $\mathbb{E}\|(V\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - V^\top \boldsymbol{u}(V\boldsymbol{\theta})\|_2^2$ is order K^3/n.*

Proof. (of Lemma 8)
We have that

$$\begin{aligned} &\mathbb{E}\|(V\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - V^\top \boldsymbol{u}(V\boldsymbol{\theta})\|_2^2 \\ &\leq 2\mathbb{E}\|(V\hat{\Sigma}^{-1/2})^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - V^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta})\|_2^2 \quad \text{(Term 1)} \\ &\qquad + 2\mathbb{E}\|V^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - V^\top \boldsymbol{u}(V\boldsymbol{\theta})\|_2^2. \quad \text{(Term 2)} \end{aligned}$$

We bound the two terms separately. For the first term, we observe that

$$\begin{aligned} \text{(Term 1)} &= 2\mathbb{E}\|(\hat{\Sigma}^{-1/2} - I_K)\big\{V^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta})\big\}\|_2^2 \\ &\leq 2\|\hat{\Sigma}^{-1/2} - I_K\|_2^2 \mathbb{E}\|V^\top \boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta})\|_2^2 \leq 2\|\hat{\Sigma}^{-1/2} - I_K\|_2^2 B_1. \end{aligned}$$

We now turn to the second term. We write $E = \hat{\Sigma}^{-1/2} - I_K$ and $\bar{E} := E/\|E\|_2$. By mean value theorem, there exists $\lambda \in [0,1]$, dependent on $\boldsymbol{\theta}$, such that

$$\begin{aligned} &\|\boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - \boldsymbol{u}(V\boldsymbol{\theta})\|_2^2 \\ &= \|(J\boldsymbol{u})(V\{I_K + \lambda E\}\boldsymbol{\theta})VE\boldsymbol{\theta}\|_2^2 \\ &= \|E\|_2^2 \sum_{k=1}^K \big(u_k'(V\{I_K + \lambda E\}\boldsymbol{\theta})(V\bar{E})_{k\cdot}^\top \boldsymbol{\theta}\big)^2 \\ &\leq \|E\|_2^2 \sum_{k=1}^K u_k'(V\{I_K + \lambda E\}\boldsymbol{\theta})^2 \|\boldsymbol{\theta}\|_2^2. \end{aligned}$$

Therefore, we may bound Term 2 as

$$\begin{aligned}(\text{Term 2}) &= \mathbb{E}\|\boldsymbol{u}(V\hat{\Sigma}^{-1/2}\boldsymbol{\theta}) - \boldsymbol{u}(V\boldsymbol{\theta})\|_2^2 \\ &\leq 2\|E\|_2^2 \sum_{k=1}^{K} \mathbb{E}\left[\sup_{\lambda\in[0,1]} u_k'(V\{I_K + \lambda E\}\boldsymbol{\theta})^2 \|\boldsymbol{\theta}\|_2^2\right] \leq 2\|\hat{\Sigma}^{-1/2} - I_K\|_2^2 B_1.\end{aligned}$$

The conclusion of the Lemma follows as desired. □

S4.9 Technical lemmas

S4.9.1 Orthogonal matrices

For an orthogonal matrix $V \in SO(K)$, we may have $V = e^H = e^{H'}$ for two different matrices H, H'. The next Lemma shows that if V is sufficiently close to identity, then its matrix exponential representation is essentially unique if we restrict ourselves to skew-symmetric matrices with small Frobenius norm.

Lemma 9. *Let $V \in SO(K)$ suppose $\|V - I\|_F \leq \frac{1}{2}$. Then, there exists a skew-symmetric matrix H such that $V = e^H$ and*

$$\frac{1}{2}\|H\|_F \leq \|V - I\|_F \leq 2\|H\|_F.$$

Proof. Since $\|V - I\|_2 \leq \|V - I\|_F < \frac{1}{2}$, we may define the convergent series

$$H := \log V = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(V - I)^k}{k}$$

and we note that $e^H = \sum_{k=0}^{\infty} \frac{H^k}{k!} = V$.

Using the fact that $\|V - I\|_F \leq \frac{1}{2}$, we have

$$\begin{aligned}\|H\|_F &\leq \sum_{k=1}^{\infty} \frac{\|V - I\|_F^k}{k} \\ &\leq \|V - I\|_F \left\{\sum_{k=0}^{\infty} \frac{\|V - I\|_F^k}{k+1}\right\} \\ &\leq 2\|V - I\|_F.\end{aligned}$$

In particular, we have that $\|H\|_F < 1$. Using the fact that $e^x - 1 \leq 2x$ when $x \leq 1$, we also have

$$\begin{aligned}\|V - I\|_F &= \|e^H - I\|_F \\ &\leq \|H\|_F + \frac{\|H\|_F^2}{2} + \frac{\|H\|_F^3}{3!} + \dots \\ &\leq e^{\|H\|_F} - 1 \leq 2\|H\|_F.\end{aligned}$$

In summary, we have that

$$\frac{1}{2}\|H\|_F \leq \|V - I\|_F \leq 2\|H\|_F, \tag{S4.62}$$

as desired. To see that H is skew symmetric, note that $(V - I)(V^\top - I) = (V^\top - I)(V - I)$ so that $HH^\top = H^\top H$. Hence, $e^{H+H^\top} = e^H e^{H^\top} = VV^\top = I$. Therefore, we conclude that H is a skew symmetric matrix.

□

S4.9.2 Bounds on the first derivative under higher derivative integral constraints

The key result of this Section is Proposition 11, which bounds $\int |f'(x)|^2 p(x)dx$ in terms of $\int |f(x)|^2 p(x)dx$ for a sufficiently smooth density $p(\cdot)$ and a univariate function $f(\cdot)$ satisfying a Sobolev condition $\int |f''(x)|^2 dx \leq B$. Our analysis extends result by Agmon which considers the case where $p(\cdot)$ is uniform; see Theorem 3.1 and its Corollary in Agmon (1966) and also Lemma 10.8 in van de Geer (2000).

We start with the following Lemma:

Lemma 10. *Let $a < b$ be real numbers and let f be a twice continuously differentiable function on $[a, b]$; let $q_1(\cdot)$, $q_2(\cdot)$ be densities on $[a, b]$. Then,*

$$\int_a^b |f'(x)|^2 q_1(x)dx \leq 128\left(\sup_{x\in(a,b)} q_2(x)\right)\int_a^b |f(x)|^2 q_2(x)dx + 2(b-a)\int_a^b |f''(x)|^2 dx.$$

Proof. First assume that $[a, b] = [0, 1]$ and write $M_0 = \sup_{x\in[0,1]} q_2(x)$. Let $\alpha \in (0, 1/2)$ and define $s_\alpha, t_\alpha \in [0, 1]$ such that $\int_0^{s_\alpha} q_2(x)dx = \int_{t_\alpha}^1 q_2(x)dx = \alpha$. We observe that

$$t_\alpha - s_\alpha = \frac{1}{M_0}\int_{s_\alpha}^{t_\alpha} M_0 dx \geq \frac{1}{M_0}\int_{s_\alpha}^{t_\alpha} q_2(x)dx = \frac{1-2\alpha}{M_0}.$$

For any $x_1 \in (0, s_\alpha)$ and $x_2 \in (t_\alpha, 1)$, there exists $\eta \in (0, 1)$ such that $f'(\eta) = \frac{f(x_2)-f(x_1)}{x_2-x_1}$, which further implies that for any $x \in (0, 1)$,

$$|f'(x)| = \left|f'(\eta) + \int_\eta^x f''(t)dt\right| \leq M_0\frac{|f(x_2)| + |f(x_1)|}{1-2\alpha} + \int_0^1 |f''(t)|dt, \tag{S4.63}$$

where the inequality follows because $x_2 - x_1 \geq t_\alpha - s_\alpha \geq \frac{1-2\alpha}{M_0}$. Since x_1, x_2 are arbitrary elements of $(0, s_\alpha)$ and $(t_\alpha, 1)$ respectively, we may integrate them in (S4.63) to obtain

$$\alpha^2|f'(x)| \leq \frac{\alpha M_0}{1-2\alpha}\int_{(0,s_\alpha)\cup(t_\alpha,1)} |f(t)|q_2(t)dt + \alpha^2\int_0^1 |f''(t)|dt \tag{S4.64}$$

$$(\Rightarrow) \quad |f'(x)| \leq \frac{M_0}{\alpha(1-2\alpha)}\int_0^1 |f(t)|q_2(t)dt + \int_0^1 |f''(t)|dt. \tag{S4.65}$$

Squaring both sides and applying Jensen's inequality twice, we obtain

$$|f'(x)|^2 \leq 2\bigg(\frac{M_0}{\alpha(1-2\alpha)}\bigg)^2 \int_0^1 |f(t)|^2 q_2(t)dt + 2\int_0^1 |f''(t)|^2 dt.$$

Using the fact that $\alpha(1-2\alpha)$ is minimized at $\alpha = 1/4$ with the value of $1/8$, we integrate to obtain

$$\int_0^1 |f'(x)|^2 q_1(x)dx \leq 128 M_0^2 \int_0^1 |f(x)|^2 q_2(x)dx + 2\int_0^1 |f''(x)|^2 dx.$$

For the general case, we apply the change of variable $z = (b-a)x + a$ so that $z \in [a,b]$. We note that, writing $g(z) = f(x) = f(\frac{z-a}{b-a})$,

$$g'(z) = f'(x)\frac{1}{b-a} \quad \text{and} \quad g''(z) = f''(x)\frac{1}{(b-a)^2}.$$

Therefore, we have

$$\begin{aligned}
(b-a)^2 \int_a^b |g'(z)|^2 q_1\bigg(\frac{z-a}{b-a}\bigg)\frac{1}{b-a}dz &\leq 128 M_0^2 \int_a^b |g(z)|^2 q_2\bigg(\frac{z-a}{b-a}\bigg)\frac{1}{b-a}dz \\
&\quad + 2(b-a)^4 \int_a^b |g''(z)|^2 \frac{dz}{b-a} \\
(\Rightarrow) \int_a^b |g'(z)|^2 q_1\bigg(\frac{z-a}{b-a}\bigg)\frac{1}{b-a}dz &\leq 128\frac{M_0^2}{(b-a)^2} \int_a^b |g(z)|^2 q_2\bigg(\frac{z-a}{b-a}\bigg)\frac{1}{b-a}dz \\
&\quad + 2(b-a)\int_a^b |g''(z)|^2 dz
\end{aligned}$$

Since any density $p(\cdot)$ on $[a,b]$ can be written as $p(z) = q\big(\frac{z-a}{b-a}\big)\frac{1}{b-a}$ for some density $q(\cdot)$ on $[0,1]$, and $q(\cdot) \leq M_0$ implies that $p(\cdot) \leq \frac{M_0}{b-a}$, the desired conclusion follows.

□

Lemma 11. *Let $a < b$ be real numbers and let p be a density on (a,b). We assume there exists $M_0, \delta_0 > 0$ such that*

$$\text{for any interval } (s,t) \text{ such that } t-s \leq \delta_0, \text{ we have that } \sup_{x\in(s,t)} p(x) \leq M_0\frac{1}{t-s}\int_s^t p(x)dx. \tag{S4.66}$$

Then, for any $\epsilon \in (0, \delta_0^2(b-a)^{-2}]$,

$$\int_a^b |f'(x)|^2 p(x)dx \leq 128 M_0^2 \frac{1}{\epsilon(b-a)^2}\int_a^b |f(x)|^2 p(x)dx + \bigg(\sup_{x\in(a,b)} p(x)\bigg)\epsilon(b-a)^2 \int_a^b |f''(x)|^2 dx.$$

Proof. Define a partition $a = a_1 < a_2 < \ldots < a_n = b$ where $a_{i+1} - a_i \leq \sqrt{\epsilon}(b-a)$.

Fix $i \in [n-1]$ and define $\tilde{q}(x) = \frac{p(x)}{\int_{a_i}^{a_{i+1}} p(x)dx}$ so that $\tilde{q}(\cdot)$ is a density on (a_i, a_{i+1}). We note that since $(a_{i+1}-a_i)=\sqrt{\epsilon}(b-a)\le\delta_0$, we have

$$\sup_{x\in(a_i,a_{i+1})}\tilde q(x)\le\frac{\sup_{x\in(a_i,a_{i+1})}p(x)}{\frac{1}{a_{i+1}-a_i}\int_{a_i}^{a_{i+1}}p(x)dx}\frac{1}{a_{i+1}-a_i}\le\frac{M_0}{a_{i+1}-a_i}. \tag{S4.67}$$

Hence, by applying Lemma 10 with $q_1=q_2=\tilde q$, we have that

$$\int_{a_i}^{a_{i+1}}|f'(x)|^2\tilde q(x)dx\le 128M_0^2\frac{1}{\epsilon(b-a)^2}\int_{a_i}^{a_{i+1}}|f(x)|^2\tilde q(x)dx+\epsilon(b-a)^2\int_{a_i}^{a_{i+1}}|f''(x)|^2\frac{dx}{a_{i+1}-a_i}.$$

Multiplying both slides by $\int_{a_i}^{a_{i+1}}p(x)dx$ and noting that $\frac{1}{a_{i+1}-a_i}\int_{a_i}^{a_{i+1}}p(x)dx\le\sup_{x\in(a,b)}p(x)$, we have that

$$\int_{a_i}^{a_{i+1}}|f'(x)|^2p(x)dx\le 128M_0^2\frac{1}{\epsilon(b-a)^2}\int_{a_i}^{a_{i+1}}|f(x)|^2p(x)dx+\left(\sup_{x\in(a,b)}p(x)\right)\epsilon(b-a)^2\int_{a_i}^{a_{i+1}}|f''(x)|^2dx.$$

Summing this inequality for all $i\in[n-1]$ yields the first conclusion.

□

Proposition 11. *Let $a<b$ be real numbers and let p be a density on (a,b). We assume there exists $M_0,M_1\ge 1$ and $\delta_0\in(0,1]$ such that*

1. *$\sup_{x\in(a,b)}p(x)\le M_1$ and*
2. *for any interval (s,t) such that $t-s\le\delta_0$, we have that $\sup_{x\in(s,t)}p(x)\le M_0\frac{1}{t-s}\int_s^t p(x)dx$.*

Then, for all $\epsilon\in(0,\delta_0^2(b-a)^{-2}]$,

$$\int_a^b|f'(x)|^2p(x)dx\lesssim M_0^{2m}\frac{1}{\epsilon(b-a)^2}\int_a^b|f(x)|^2p(x)dx+M_1M_0\epsilon^{m-1}(b-a)^{2(m-1)}\int_a^b|f^{(m)}(x)|^2dx.$$

As a direct consequence, if $\int_a^b|f^{(m)}(x)|^2dx\le B$ and if $\int_a^b|f(x)|^2p(x)dx\le\delta_0^{2m}$, then, by letting $\epsilon=\left(\int_a^b|f(x)|^2p(x)dx\right)^{\frac{1}{m}}(b-a)^{-2}$, we have that

$$\int_a^b|f'(x)|^2p(x)dx\lesssim M_0^{2m}M_1B\left(\int_a^b|f(x)|^2p(x)dx\right)^{1-\frac{1}{m}}$$

Proof. Let (s,t) be an interval such that $|t-s|\le\delta_0$ and define $\tilde q(x)=\frac{p(x)}{\int_s^t p(x)dx}$ as a density on (s,t). By the same reasoning as (S4.67), we have that $\sup_{x\in(s,t)}\tilde q(x)\le\frac{M_0}{t-s}$.

Therefore, by Lemma 10, we have

$$\frac{1}{t-s}\int_s^t|f''(x)|^2dx\le 128M_0^2\frac{1}{(t-s)^2}\int_s^t|f'(x)|^2\tilde q(x)dx+2(t-s)^2\int_s^t|f^{(3)}(x)|^2\frac{dx}{t-s}.$$

We note that if the density $p(\cdot)$ on (a,b) satisfies condition (S4.66), then $\tilde q(\cdot)$ also satisfies condition (S4.66) since, for any (s', t') whose width is at most δ_0, we have

$$\sup_{x \in (s',t')} \tilde{q}(x) \leq \sup_{x \in (s',t')} \frac{p(x)}{\int_s^t p(x)dx} \leq \frac{M_0}{t'-s'} \frac{\int_{s'}^{t'} p(x)dx}{\int_s^t p(x)dx} \leq \frac{M_0}{t'-s'}.$$

Let $\epsilon_0 = (256M_0^3)^{-1}$ so that $128\epsilon_0 M_0^3 \leq \frac{1}{2}$; we note also that $\epsilon_0 \leq 1 \leq \delta_0^2 \frac{1}{(t-s)^2}$ by the assumption that $t - s \leq \delta_0$. Hence, by applying Lemma 11 with $\epsilon = \epsilon_0$, we obtain

$$\int_s^t |f'(x)|^2 \tilde{q}(x)dx \leq 128\frac{M_0^2}{\epsilon_0}\frac{1}{(t-s)^2}\int_s^t |f(x)|^2\tilde{q}(x)dx + \frac{M_0}{t-s}\epsilon_0(t-s)^2\int_s^t |f''(x)|dx$$

$$\int_s^t |f'(x)|^2 \tilde{q}(x)dx \leq 128\frac{M_0^2}{\epsilon_0}\frac{1}{(t-s)^2}\int_s^t |f(x)|^2\tilde{q}(x)dx$$
$$+ M_0\epsilon_0(t-s)\left\{128M_0^2\frac{1}{t-s}\int_s^t |f'(x)|^2\tilde{q}(x)dx + 2(t-s)^2\int_s^t |f^{(3)}(x)|^2dx\right\}.$$

$$(\Rightarrow)\int_s^t |f'(x)|^2 \tilde{q}(x)dx \leq 256\frac{M_0^2}{\epsilon_0}\frac{1}{(t-s)^2}\int_s^t |f(x)|^2\tilde{q}(x)dx + 4M_0\epsilon_0(t-s)^3\int_s^t |f^{(3)}(x)|^2dx.$$

Using the fact that $\frac{1}{t-s}\int_s^t p(x)dx \leq M_1$ and writing $C_1 = 256M_0^2/\epsilon_0$ and $C_2 = 4M_0\epsilon_0$, we then have that

$$\int_s^t |f'(x)|^2 p(x)dx \leq C_1\frac{1}{(t-s)^2}\int_s^t |f(x)|^2 p(x)dx + C_2(t-s)^4\int_s^t |f^{(3)}(x)|^2dx. \tag{S4.68}$$

We note that $C_1 \leq 2^{16}M_0^5$ and that $C_2 \leq 4M_1M_0$.

Now fix $\epsilon \in (0, \frac{\delta_0^2}{(b-a)^2})$ and choose a partition $a = a_1 < a_2 < \dots a_n = b$ where $a_{i+1} - a_i \leq \sqrt{\epsilon}(b-a)$. We may now apply (S4.68) on each (a_i, a_{i+1}) in an argument identical to that of Lemma 11 to obtain the conclusion of the proposition for $m = 3$.

The argument for any $m > 3$ is similar. For the sake of concisely, we only sketch out the analysis for $m = 4$. Fix a sub-interval (s, t) of (a, b) where $t - s \leq \delta_0$. Let $\tilde{q}(x) = \frac{p(x)}{\int_s^t p(x)}$ and note that $\sup_{x\in(s,t)} \tilde{q}(x) \leq M_0$ and that $\tilde{q}$ satisfies condition (S4.66).

By applying Lemma 10 on $f^{(2)}$, we obtain

$$\frac{1}{t-s}\int_s^t |f^{(3)}(x)|^2dx \leq 128M_0\int_s^t |f^{(2)}(x)|^2\tilde{q}(x)dx + 2(t-s)\int_s^t |f^{(4)}(x)|^2dx. \tag{S4.69}$$

Now, by applying the conclusion of this proposition in the $m = 3$ case to f' with $\epsilon = \epsilon_0$ where ϵ_0 is a real number in $(0, 1)$ whose value is given later, we have

$$\int_s^t |f^{(2)}(x)|^2\tilde{q}(x)dx \leq C_1\frac{1}{\epsilon_0(t-s)^2}\int_s^t |f'(x)|^2\tilde{q}(x)dx + M_1C_2\epsilon_0^2(t-s)^4\int_s^t |f^{(4)}(x)|^2dx. \tag{S4.70}$$

By plugging (S4.70) into (S4.69), we have

$$\frac{1}{t-s}\int_s^t |f^{(3)}(x)|^2dx \leq 128M_0C_1\frac{1}{\epsilon_0(t-s)^2}\int_s^t |f'(x)|^2\tilde{q}(x)dx$$
$$+\left\{128M_0M_1C_2\epsilon_0^2(t-s)^4 + 2(t-s)\right\}\int_s^t |f^{(4)}(x)|^2dx. \quad \text{(S4.71)}$$

Now, we apply the conclusion of this proposition in the $m=3$ case to f with $\epsilon=\epsilon_1$ where ϵ_1 is a real number in $(0,1)$ whose value is given later,

$$\int_s^t |f'(x)|^2\tilde{q}(x)dx \leq C_1\frac{1}{\epsilon_1(t-s)^2}\int_s^t |f(x)|^2\tilde{q}(x)dx + M_1C_2\epsilon_1^2(t-s)^4\int_s^t |f^{(3)}(x)|^2dx. \quad \text{(S4.72)}$$

By choosing ϵ_0, ϵ_1 such that $128M_0M_1C_2\epsilon_0^2 = 1$ and $128M_0^{1/2}C_1C_2^{1/2}\epsilon_1^2 = \frac{1}{2}$, and by plugging (S4.71) into (S4.72), we have that

$$\int_s^t |f'(x)|^2\tilde{q}(x)dx \leq C_1'\frac{1}{(t-s)^2}\int_s^t |f(x)|^2\tilde{q}(x)dx + 3(t-s)^5\int_s^t |f^{(3)}(x)|^2dx,$$

where $C_1' = C_1/\epsilon_1 \lesssim M_0^8$. Now, by sub-dividing the interval (a,b) into segments of length $\sqrt{\epsilon}(b-a)$ and applying the same argument for the $m=3$, we obtain the desired conclusion for $m=4$. □

Corollary 6. *Let $b \geq 2$, let $m \geq 2$ be an integer, and suppose $p(\cdot)$ is a density on $(-b,b)$ of the form $p(x) = e^{u(x)}$ where $u(x)$ satisfies*

1. $\int_{-b}^b |u^{(m)}(x)|^2dx \leq B$ *and*
2. $|u^{(j)}(0)| \leq B^{1/2}$ *for all* $j \in [m]$.

Suppose also $\sup_x p(x) \leq M$ *for some* $M > 0$. *Then, for any interval (s,t) such that $|s-t| \leq 1$,*

$$\sup_{x\in(s,t)} p(x) \leq eB^{1/2}b^m\frac{1}{t-s}\int_s^t p(x)dx.$$

As a direct corollary, if $f(\cdot)$ is a function such that $\int_{-b}^b |f^{(m)}(x)|^2p(x)dx \leq B$ and $\int_{-b}^b |f(x)|^2p(x)dx \leq 1$, then

$$\int_{-b}^b |f'(x)|^2p(x)dx \lesssim 2^{3m}B^{m+1}b^{2m^2}M\left(\int_{-b}^b |f(x)|^2p(x)dx\right)^{1-\frac{1}{m}}.$$

Proof. We observe that, for any $x \in (-b,b)$,

$$|u^{(m-1)}(x)| = \left|\int_0^x u^{(m)}(t)dt - u^{(m)}(0)\right| \leq bB^{1/2} + B^{1/2} = (b+1)B^{1/2}$$
$$|u^{(m-2)}(x)| \leq b(b+1)B^{1/2} + B^{1/2} = (b^2+b+1)B^{1/2}$$
$$\dots$$
$$|u'(x)| \leq (b^{m-1} + b^{m-2} + \ldots + b + 1)B^{1/2} \leq b^mB^{1/2},$$

where the last inequality follows from the assumption that $b \geq 2$.

Write $\epsilon = b^{-m}B^{-1/2}$. Let $(s,t) \subset (-b,b)$ such that $|s-t| \leq 1$ and let $s^* \in [s,t]$ be the point such that $p(s^*) = \sup_{x\in(s,t)} p(x)$. Then, by mean value theorem,

$$\begin{aligned}
\frac{\frac{1}{t-s}\int_s^t p(x)dx}{\sup_{x\in(s,t)} p(x)} &= \frac{1}{t-s}\int_s^t e^{u(x)-u(s^*)}dx \\
&\geq \int_{(s^*-\epsilon,s^*+\epsilon)\cap(s,t)} e^{u(x)-u(s^*)}dx \\
&\geq \epsilon \inf_{x\in(s^*-\epsilon,s^*+\epsilon)} e^{(u(x)-u(s^*))} \\
&\geq \epsilon e^{-|b^m B^{1/2}|\epsilon} \\
&\geq b^{-m}B^{-1/2}e^{-1}.
\end{aligned}$$

The conclusion then follows directly from Proposition 11 by letting $M_1 = M$, $M_0 = eb^m B^{1/2}$.

$\square$