Index Models for Sparsely Sampled Functional Data.
Supplementary Material.

May 21, 2014

1 Proof of Theorem 2

We will follow the general argument in the proof of Theorem 3 in Li *et al.* (2010). Write $\mathbf{d} = (d_1, ..., d_p)$ and $\mathbf{d}_c = (d_{c1}, ..., d_{cp})$. Let $G_n(\mathbf{d}_c)$ denote the difference between the criterion values for the candidate and the true dimension vectors:

$$G_n(\mathbf{d}_c) = \log L(\mathbf{d}_c) - \log L(\mathbf{d}) + d_c \log n/[nh_n^{\tilde{d}_c}(d_c)] - d \log n/[nh_n^{\tilde{d}}(d)].$$

Note that the set of candidate dimension vectors is finite. Thus, it is sufficient to show that for each vector $\mathbf{d}_c$ not equal to $\mathbf{d}$ function $G_n(\mathbf{d}_c)$ is positive with probability tending to one. Note that

$$d_c \log n/[nh_n^{\tilde{d}_c}(d_c)] - d \log n/[nh_n^{\tilde{d}}(d)] = O\big(\log n[n^{-4/(\tilde{d}_c+4)} + n^{-4/(\tilde{d}+4)}]\big) = o(1).$$

If $d_{cj} < d_j$ for at least one j, we can choose a positive constant c, such that $\log L(\mathbf{d}_c) - \log L(\mathbf{d}) > c$ with probability tending to one, due to the lack of fit. It follows that $G_n(\mathbf{d}_c) > 0$ with probability tending to one.

Now consider the remaining case of $d_{cj} \geq d_j$ for all j and $d_c > d$. In this case $d_c \log n/[nh_n^{\tilde{d}_c}(d_c)] - d \log n/[nh_n^{\tilde{d}}(d)] > \log n/[nh_n^{\tilde{d}_c}(d_c)]$ for all sufficiently large n. On the other hand,

$$\log L(\mathbf{d}_c) - \log L(\mathbf{d}) = \log(1 + \frac{L(\mathbf{d}_c) - L(\mathbf{d})}{L(\mathbf{d})}) = \log(1 + O_p(1/[nh_n^{\tilde{d}_c}(d_c)])) = O_p(1/[nh_n^{\tilde{d}_c}(d_c)]),$$

by the classical results on local linear smoothing. Consequently, with probability tending to one,

$$G_n(\mathbf{d}_c) > (1/2) \log n/[nh_n^{\tilde{d}_c}] > 0.$$

2 Proof of Theorems 3 and 4

Theorem 3. For simplicity of the exposition we will focus on the case of a single predictor. Due to the additive structure of the SIMFE estimation procedure with respect to the predictors, extension to the general case presents only notational challenges, while the argument itself remains essentially intact. We will also simplify the notation by omitting the superscript containing the predictor index. For example, we will write $\tilde{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{P}}_i$ instead of $\tilde{\boldsymbol{\mu}}_{ij}$ and $\hat{\mathbf{P}}_{ij}$. Define $\delta_{kh} = [\log n/(nh^k)]^{1/2}$. Observe the relationship $||A\hat{\beta}(t) - \beta(t)|| = \|A[\hat{\boldsymbol{\eta}}] - [\boldsymbol{\eta}]\|_F$, where $[\hat{\boldsymbol{\eta}}]$ and $[\boldsymbol{\eta}]$ denote matrixes $[\hat{\boldsymbol{\eta}}_1 ... \hat{\boldsymbol{\eta}}_d]^T$ and $[\boldsymbol{\eta}_1 ... \boldsymbol{\eta}_d]^T$, respectively, and $\|\cdot\|_F$ stands for the Frobenius matrix norm. Hence, we need to show that there exists an invertible matrix A, for which $\|A[\hat{\boldsymbol{\eta}}] - [\boldsymbol{\eta}]\|_F = O_p(h_{opt}^4 + \delta_{dh_{opt}}^2 + n^{-1/2})$.

In the new notation $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}_{i1}, ..., \hat{\mu}_{iq})^T$ and $\hat{\boldsymbol{\eta}} * \hat{\boldsymbol{\mu}}_i = (\hat{\boldsymbol{\eta}}_1^T \hat{\boldsymbol{\mu}}_i, ..., \hat{\boldsymbol{\eta}}_d^T \hat{\boldsymbol{\mu}}_i)^T = \hat{\mathbf{P}}_i$. To be able to conveniently apply existing results, we will slightly modify the trimmed objective function, replacing $\hat{\mathbf{P}}_l$ with $\hat{\mathbf{P}}_l - \hat{\mathbf{P}}_i$ and writing this expression in terms of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$. We will also use $\hat{\boldsymbol{\mu}}_{l:i}$ to denote $\hat{\boldsymbol{\mu}}_l - \hat{\boldsymbol{\mu}}_i$. The modified objective function,

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{n} I_{ni} \rho_{\hat{\boldsymbol{\eta}} i} \left(Y_l - a_i - \mathbf{c}_i^T \hat{\boldsymbol{\eta}} * \hat{\boldsymbol{\mu}}_{l:i} \right)^2 \tilde{K}_h(\hat{\boldsymbol{\eta}} * \hat{\boldsymbol{\mu}}_{l:i}),$$

however, corresponds to exactly the same SIMFE estimator as the original trimmed function. Define $\check{K}_t = I_{ni} \rho_{\hat{\boldsymbol{\eta}} i} \tilde{K}_{h_t}(\hat{\boldsymbol{\eta}} * \hat{\boldsymbol{\mu}}_{l:i})$, $\hat{\boldsymbol{\Delta}}_{l:i} = \mathbf{c}_i \otimes \hat{\boldsymbol{\mu}}_{l:i}$ and $\boldsymbol{\Delta}_{l:i} = \mathbf{c}_i \otimes \tilde{\boldsymbol{\mu}}_{l:i}$. We will use the superscript (t, τ) to indicate that the estimator corresponds to the τ iteration of the algorithm corresponding to the bandwidth h_t. A simple manipulation of formula (23), taking into account the modifications to the objective function, yields

$$\hat{\boldsymbol{\eta}}^{(t,\tau+1)} = \boldsymbol{\eta}_{r_1} + \{ \sum_{i,l=1}^{n} \check{K}_t \hat{\boldsymbol{\Delta}}_{l:i}^{(t,\tau)} (\hat{\boldsymbol{\Delta}}_{l:i}^{(t,\tau)})^T \}^{-1} \sum_{i,l=1}^{n} \check{K}_t \hat{\boldsymbol{\Delta}}_{l:i}^{(t,\tau)} \{ Y_l - a_i^{(t,\tau)} - \boldsymbol{\eta}_{r_1} \hat{\boldsymbol{\Delta}}_{l:i}^{(t,\tau)} \} \qquad (1)$$

Here $\boldsymbol{\eta}_{r_1}$ corresponds to an arbitrary rotation of $[\boldsymbol{\eta}]$, and the above formula holds for each such rotation.

Let M denote the number of time point configurations, at which the predictor functions are observed. We will only consider configurations that are generated with positive probabilities. Denote by A_k, $k = 1, ..., M$, the index set of the observations corresponding to the k-th time point configuration. Note that, using the basis representation for the predictors and the projection functions, equation (10) simplifies to

$$Y_i = \tilde{m}_k \left(\boldsymbol{\eta}_1^T \tilde{\boldsymbol{\mu}}_i, \ldots, \boldsymbol{\eta}_d^T \tilde{\boldsymbol{\mu}}_i \right) + \varepsilon_i^*, \qquad i \in A_k, \qquad (2)$$

where $E(\varepsilon_i^*|\mathbf{W}_i) = 0$. The last equality also implies $E(\varepsilon_i^*|\tilde{\boldsymbol{\mu}}_i) = 0$. This formulation of the SIMFE model allows us to apply some of the theory developed for the MAVE approach. We will first focus on the right-hand side of display (1), for sufficiently large n, with $\hat{\boldsymbol{\Delta}}$ replaced by $\boldsymbol{\Delta}$, and rewrite it as $\boldsymbol{\eta}_{r_1} + \{\sum_{k=1}^M \tilde{\Sigma}_k\}^{-1}\{\sum_{k=1}^M \Sigma_k\}$, where $\tilde{\Sigma}_k = \sum_{i,l\in A_k} \check{K}_t \boldsymbol{\Delta}_{l:i}^{(t,\tau)}(\boldsymbol{\Delta}_{l:i}^{(t,\tau)})^T$ and $\Sigma_k = \sum_{i,l\in A_k} \check{K}_t \boldsymbol{\Delta}_{l:i}^{(t,\tau)}(Y_i - a_j^{(t,\tau)} - \boldsymbol{\eta}_{r_1}\boldsymbol{\Delta}_{l:i}^{(t,\tau)})$. We will apply Lemma A.5 in the supplemental material of Xia (2008) to each Σ_k, and use a natural generalization of Lemma A.4 to handle $\{\sum_{k=1}^M \tilde{\Sigma}_k\}^{-1}$. For each $k \le M$, write π_k for the probability of the k-th time point configuration, and let $\tilde{\boldsymbol{\mu}}^{(k)}$ denote $\tilde{\boldsymbol{\mu}}_i$ for some i in A_k. Define

$$\Phi_n = n^{-1}\sum_{k=1}^M \frac{|A_k|}{n}|\sum_{i\in A_k} \rho(\tilde{f}_i(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}_i))\big(\nabla\tilde{m}_k(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}_i) \otimes \boldsymbol{\nu}(\tilde{\boldsymbol{\mu}}_i)\big)\varepsilon_i^*,$$

$$D_1 = \sum_{k=1}^M \pi_k^2 E\rho(\tilde{f}(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}^{(k)}))\big(\nabla\tilde{m}_k(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}^{(k)}) \otimes \boldsymbol{\nu}(\tilde{\boldsymbol{\mu}}^{(k)})\big)\big(\nabla\tilde{m}_k(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}^{(k)}) \otimes \boldsymbol{\nu}(\tilde{\boldsymbol{\mu}}^{(k)})\big)^T, \quad (3)$$

and

$$D_2 = 2\sum_{k=1}^M \pi_k^2 E\rho(\tilde{f}(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}^{(k)}))\nabla\tilde{m}_k(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}^{(k)})\nabla^T\tilde{m}_k(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}^{(k)}) \otimes \boldsymbol{\omega}(\tilde{\boldsymbol{\mu}}^{(k)})\big)^T. \quad (4)$$

where $\boldsymbol{\nu}(\tilde{\boldsymbol{\mu}}) = \tilde{\boldsymbol{\mu}} - E(\tilde{\boldsymbol{\mu}}|\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}})$ and $\boldsymbol{\omega}(\tilde{\boldsymbol{\mu}}) = E(\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T|\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}) - E(\tilde{\boldsymbol{\mu}}|\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}})E(\tilde{\boldsymbol{\mu}}|\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}})^T$. We will use superscript + to denote the MoorePenrose inverse. By Lemmas A.4 and A.5 in the supplemental material of Xia (2008), there exists a rotation of $[\boldsymbol{\eta}]$, call it $[\boldsymbol{\eta}_{r_2}]$, such that

$$\begin{aligned}\boldsymbol{\eta}_{r_1} + \{\sum_{i,l=1}^n \check{K}_t \boldsymbol{\Delta}_{l:i}^{(t,\tau)}(\boldsymbol{\Delta}_{l:i}^{(t,\tau)})^T\}^{-1}\sum_{i,l=1}^n \check{K}_t \boldsymbol{\Delta}_{l:i}^{(t,\tau)}\{Y_l - a_i^{(t,\tau)} - \boldsymbol{\eta}\boldsymbol{\Delta}_{l:i}^{(t,\tau)}\}\\ = \boldsymbol{\eta}_{r_2} + (I - D_2^+ D_1)^{-1}D_2^+\Phi_n + O_p(h_t^4 + \delta_{dh_t}^2),\end{aligned} \quad (5)$$

provided $\hat{\boldsymbol{\eta}}^{(t,\tau)} - \boldsymbol{\eta}_{r_1} = o_p(h_t)$. Note that $\Phi_n = O_p(n^{-1/2})$. Due to the assumption A3, the unknown parameters $\Delta, \boldsymbol{\mu}_\delta$ and σ^2 are estimated at the usual parametric rate, $n^{-1/2}$, and hence $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(1 + O_p(n^{-1/2}))$. Consequently, equations (1) and (5) imply

$$\hat{\boldsymbol{\eta}}^{(t,\tau+1)} - \boldsymbol{\eta}_{r_2} = O_p(h_t^4 + \delta_{dh_t}^2 + n^{-1/2}), \quad (6)$$

as long as $\hat{\boldsymbol{\eta}}^{(t,\tau)} - \boldsymbol{\eta}_{r_1} = o_p(h_t)$. Note that $\delta_{dh_t}^2 = o(h_t)$, and hence the stochastic bound in display (6) can be written as $o_p(h_t)$. It follows that the final estimator for the bandwidth h_t, which is also the initial estimator for the bandwidth h_{t+1}, satisfies $\hat{\boldsymbol{\eta}}^{(t+1,0)} - \boldsymbol{\eta}_r = o_p(h_{t+1})$, for some rotation of $[\boldsymbol{\eta}]$. Hence, as long as $\hat{\boldsymbol{\eta}}^{(0,0)} - \boldsymbol{\eta} = o_p(h_0)$ holds for the initialization estimator, we can establish, by induction, that the final SIMFE estimator satisfies

$$||[\hat{\boldsymbol{\eta}}] - [\boldsymbol{\eta}_r]||_F = O_p(h_{opt}^4 + \delta_{dh_{opt}}^2 + n^{-1/2}), \quad (7)$$

where $[\boldsymbol{\eta}_r]$ is an appropriate rotation of $[\boldsymbol{\eta}]$. The required bound for the initialization estimator follows from Theorem 4 in Li *et al.* (2010), which establishes that the gOPG estimator converges at the rate $O_p(h_0^2+\delta_{\tilde{p}h_0}^2 h_0^{-1})$. The last stochastic bound is $o_p(h_0)$ by the definitions of h_0 and $\delta_{\tilde{p}h_0}$.

Theorem 4. In the case where $n_i = 1$ for all i, we can repeat the proof of Theorem 3, treating S as an additional univariate predictor. The reduced dimension increases from d to $d+1$. As a result, the convergence rate changes from $O_p\big(n^{-4/(d+4)}\log n + n^{-1/2}\big)$ to $O_p\big(n^{-4/(d+5)}\log n + n^{-1/2}\big)$. In the general case, some of the expressions in the proof of Theorem 3 need to be modified. However, because the sequence $\{n_i\}$ is bounded, the stochastic bounds (6) and (7) remain the same as in the case $n_i = 1$.

3 Proof of Theorem 5

Partition the rows of the matrix $(I-\Omega_i S_i)\Delta(I-\Omega_i S_i)^T + \sigma^2\Omega_i\Omega_i^T$ into p groups of adjacent rows, so that the size of the j-th group is q_j. Partition the columns of the same matrix analogously. This corresponds to a partition of the matrix into p^2 blocks, V_{ijk}, where j is the group index in the row partition, and k is the group index in the column partition. Write $\mathbf{v}_{ijk}$ for the vectorized form of the block V_{ijk}. Recall the definitions of matrices Σ_i and vectors $\boldsymbol{\xi}_i$ in sections 2.4 and 3. Each element of $\boldsymbol{\xi}_i$ has the form $cov(U_{ijl_1}, U_{ikl_2})$ for some indexes j,k,l_1,l_2 that satisfy: $p \geq k \geq j \geq 1$, $d_j \geq l_1 \geq 1$, $d_k \geq l_2 \geq 1$, and $l_2 \geq l_1$ whenever $k = j$. Consequently, each element of $\boldsymbol{\xi}_i$ can be written as $\boldsymbol{\eta}_{jl_1}^T V_{ijk}\boldsymbol{\eta}_{kl_2}$. Thus, there exists a collection of vectors $\boldsymbol{\gamma}_{jl_1kl_2}$, with the indexes satisfying the inequalities given above, such that for each i the elements of $\boldsymbol{\xi}_i$ have the form $\boldsymbol{\gamma}_{jl_1kl_2}^T\mathbf{v}_{ijk}$. We will denote this representation of the vector $\boldsymbol{\xi}_i$ as $\boldsymbol{\gamma} * \mathbf{v}_i$, where $\boldsymbol{\gamma}$ is the vector constructed by stacking the vectors $\boldsymbol{\gamma}_{jl_1kl_2}$, and $\mathbf{v}_i$ is similarly constructed from $\mathbf{v}_{ijk}$. Using the derivations in Section 2.4 and the notation in Appendix D, we see that $\tilde{m}_{\mathbf{t}_i}(\tilde{\mathbf{P}}_i)$ can be written as $\check{m}(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}_i, \boldsymbol{\gamma} * \mathbf{v}_i)$, for some function $\check{m}$, which does not depend on i.

We can now follow the argument in the proof of Theorem 3, with some small modifications. Our initialization estimator is still gOPG, however we use $(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{v}}_i)$, instead of $\hat{\boldsymbol{\mu}}_i$, as the predictor vectors. Here $\hat{\mathbf{v}}_i$ are constructed analogously to $\mathbf{v}_i$, but the unknown parameters Δ and σ^2 are replaced with their estimates. We initialize the bandwidth as $n^{-1/(\check{p}+4)}$, where $\check{p}$ is the dimension of $(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{v}}_i)$. As in the proof of Theorem 3, the parameters Δ and σ^2 are estimated at the parametric rate, $n^{-1/2}$, and the aforementioned gOPG estimator of $(\boldsymbol{\eta}, \boldsymbol{\gamma})$ is

consistent. We no longer partition the observations by time point configuration, but instead treat $\mathbf{v}_i$ as an additional predictor. Equation (2) is replaced with

$$Y_i = \check{m}\left(\boldsymbol{\eta} * \tilde{\boldsymbol{\mu}}_i, \boldsymbol{\gamma} * \mathbf{v}_i\right) + \varepsilon_i^*,$$

and the dimensionality of the corresponding model increases from d to $\tilde{d} = d + d(d+1)/2$. The rest of the proof is the same as that of Theorem 3, with the appropriate modifications, such as replacing $\hat{\boldsymbol{\eta}}$ with $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}})$ and $\hat{\boldsymbol{\mu}}_i$ with $(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{v}}_i)$. Taking into account the increased dimensionality of the problem, stochastic bound in display (7) changes to $O_p(\tilde{h}_{opt}^4 + \delta^2_{\tilde{d}\tilde{h}_{opt}} + n^{-1/2})$, which simplifies to $O_p(n^{-4/(\tilde{d}+4)} \log n + n^{-1/2})$.

References

Li, L., Li, B., and Zhu, L.-X. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association* **105**, 1188–1201.

Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association* **103**, 1631–1640.