

Li, Yingxing; Härdle, Wolfgang Karl; Huang, Chen

Working Paper

Smooth principal component analysis for high dimensional data

SFB 649 Discussion Paper, No. 2017-024

Provided in Cooperation with:

Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Li, Yingxing; Härdle, Wolfgang Karl; Huang, Chen (2017) : Smooth principal component analysis for high dimensional data, SFB 649 Discussion Paper, No. 2017-024, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:

<https://hdl.handle.net/10419/169214>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Smooth Principal Component Analysis for High Dimensional Data

Yingxing Li ^{*}
Wolfgang K. Härdle ^{*2}
Chen Huang ^{*2}



^{*} Xiamen University, P. R. China

^{*2} Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche
Forschungsgemeinschaft through the SFB 649 "Economic Risk".

<http://sfb649.wiwi.hu-berlin.de>
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin



Smooth Principal Component Analysis for High Dimensional Data

Yingxing Li^{*}, Wolfgang K. Härdle[†] and Chen Huang[‡]

Abstract

This paper considers smooth principle component analysis for high dimensional data with very large dimensional observations p and moderate number of individuals N . Our setting is similar to traditional PCA, but we assume the factors are smooth and design a new approach to estimate them. By connecting with Singular Value Decomposition subjected to penalized smoothing, our algorithm is linear in the dimensionality of the data, and it also favors block calculations and sequential access to memory. Different from most existing methods, we avoid extracting eigenfunctions via smoothing a huge dimensional covariance operator. Under regularity assumptions, the results indicate that we may enjoy faster convergence rate by employing smoothness assumption. We also extend our methods when each subject is given multiple tasks by adopting the two way ANOVA approach to further demonstrate the advantages of our approach.

Keywords: Principal Component Analysis; Penalized Smoothing; Asymptotics; Multilevel; fMRI;

^{*}The Wang Yanan Institute for Studies in Economics, Xiamen University, China.

[†]Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Germany.

[‡]Center for Applied Statistics and Economics, School of Business and Economics, Humboldt-Universität zu Berlin, Germany.

1 Introduction

Massive data attract lots of research interest these days, as the volume and dimensionality of these data leads to intensive analysis and calls for new statistical tools. One particular example is to study functional Magnetic Resonance Images (fMRI), which aims at identifying the active reaction regions in brain and better understanding how brain makes investment decisions (see Rangel, Camerer and Montague, 2008; Mohr et al. 2010, Majer et al. 2014). In these neuroeconomics studies, subjects were given risk related tasks and needed to make decisions. Simultaneously, the fMRI scanner records the changes in each subject's brain blood flow at volume and oxygen level during neural activity. Such blood-oxygen-level-dependent (BOLD) signals are captured on 3 dimensional spatial maps of brain voxels over time. If vectorization is conducted, the dimensionality of data for one person in one task could be roughly 10^6 at each second.

Principle component analysis (PCA) has been a very popular tool in multivariate data analysis. However, it faces challenges when applied to very high dimensional data. For example, the huge dimensionality of the BOLD signals feasibility of the conventional eigen-decomposition and reduce the estimation accuracy when extracting the fundamental factors. Although Singular Value Decomposition (SVD) could be adopted to partly reduce the dimension of covariance (see Golub and Reinsch 1970), PCA could be further improved if one utilizes some mild structure assumptions on the factors. For example, Heller, Stanley, Yekutieli, Rubin and Benjamini (2006) points out that the active risk related regions should be contiguous region in brain rather than discrete voxels.

This motivates the analysis functional principal component analysis (fPCA), see Rao (1958), Ramsay and Silverman (1997). In fPCA, data are represented with continuous curves and decomposed to factors and factor loadings. When applied to fMRI analysis, the factors could be viewed as a nonparametric ap-

proach for spatial connection and it could represent the risk related region. The factor loadings are then used to capture the temporal and subject behaviors, and they may be associated with risk altitude of individuals (van Bömmel et al 2013). Viviani, Gron and Spitzer (2005) implemented FPCA to the smoothed or functional fMRI data of 3 subjects, where single subject was considered separately. Zipunnikov et al. (2011) reached higher computational efficiency by exploiting SVD. Chen et al. (2015) propose a 3D image fPCA approach, which avoids vectorizing the 3D location coordinates to 1D and improves quality of spatial representation.

In this paper, we consider smooth principal component analysis for massive data. In our settings, we assume a balanced design. The number of individuals might be moderate, but the observations (subjected to non-negligible measurement errors) for each individual are high-dimensional. Different from most existing methods, we avoid direct estimating the smooth covariance operator and extract its eigenfunctions, thus achieve much more computation advantages. The use of SVD makes our algorithm to be linear in the dimensionality of the data, and it also favors block calculations and sequential access to memory. We shall provide details about our approach and its theoretical property in Section 2, and discuss more of its extensions in Section 3.

2 Methodology

Let X_1, \dots, X_n denote independent identically distributed (i.i.d.) random functions on a compact interval \mathcal{F} , satisfying $\int_{\mathcal{F}} E(X_i^2) < \infty$. For example, if X_i denotes the images of individual i , then it could be a d -dimensional array structure of dimension $p = \prod_{k=1}^d p_k$. Denote the mean function as μ . Then we may represent each function X_i as

$$X_i(s) = \mu(s) + \sum_{j=1}^{\infty} \bar{\psi}_{ij} \phi_j(s), \quad (1)$$

where ϕ_j and $\bar{\psi}_{ij}$ are the j th factor and its loading respectively, and $\bar{\psi}_{ij}$ are i.i.d. random variables with mean 0 and variance σ_j^2 . In practice, we may obtain the observations $Y_i(s) = X_i(s) + \epsilon_{is}$ instead. The error terms ϵ_{is} satisfy $E(\epsilon_{is}|X_i) = 0$ and $var(\epsilon_{is}|X_i)$ is finite.

First, we rewrite the representation in equation (1) by projecting μ on ϕ_j 's. Then we have

$$X_i(s) = \sum_{j=1}^{\infty} \psi_{ij} \phi_j(s), \quad (2)$$

where ϕ_j are still the factors and $\psi_{ij} = \bar{\psi}_{ij} + \langle \mu, \phi_j \rangle$. The inner product between μ and ϕ_j might lead to nonzero mean ψ_{ij} , but it helps us to avoid estimating μ , whose accuracy might be low when the number of individuals is fixed or moderately small. Utilizing representation (2), we estimate ψ_{ij} and $\phi_j(s)$ by minimizing a constraint minimization criterion:

$$\sum_i \sum_s (Y_i(s) - \sum_j \psi_{ij} \phi_j(s))^2, s.t. \int_{\mathcal{F}} (\phi_j^{(m)}(s))^2 < C_j,$$

where $\phi_j^{(m)}(s)$ is the m th derivative of $\phi_j(s)$ and the choice of m reflects one's assumption on the degree of smoothness over $\phi_j(s)$.

Note that the above minimization criterion is equivalent to a penalized least square criterion

$$\sum_i \sum_s (Y_i(s) - \sum_j \psi_{ij} \phi_j(s))^2 + \sum_j \lambda_j \int_{\mathcal{F}} (\phi_j^{(m)}(s))^2,$$

where the penalty term tries to control the roughness of the estimated $\phi_j(s)$. Let $\tilde{\psi}_{ij}$ and $\tilde{\phi}_j(s)$ be the minimizer. For fixed $\tilde{\phi}_j$'s, we could easily obtain $\tilde{\psi}_{ij}$ by regressing $Y_i(s)$ on $\phi_j(s)$. For fixed ψ_{ij} 's, we could estimate $\phi_j(s)$ as if adopting penalized nonparametric smoothing on varying coefficient model (Chen, Fang and Li 2015). Hence one could use an iterative algorithm to obtain all estimates.

As we mentioned in the Introduction session, the dimension of the observations might be huge. For example, in the study of van Bömmel et al 2013,

the dimension per scan per person could be $91 \times 109 \times 91$, roughly 10^6 voxels. Therefore, we further consider a two stage procedure, which could achieve the same estimation efficiency with much less computation burden. Specifically, our algorithm is designed as below:

Stage 1: Use a common smoother and simultaneously smooth all individual curves. Apply SVD on these smooth curves and extract $\hat{\psi}_{ij}$.

Stage 2: Conduct penalized regression between $Y_i(s)$ and $\hat{\psi}_{ij}$ to obtain the smooth estimate $\hat{\phi}_j(s)$. Regress $Y_i(s)$ on $\hat{\phi}_j(s)$ to obtain one step update of $\hat{\psi}_{ij}$.

We shall discuss the asymptotic properties of our approach very soon. Here are more explanations about this two stage procedure. Stage 1 is used to obtain a consistent estimate of $\hat{\phi}_{ij}$ and the smoothing parameter may not be tuned very delicately. In contrast, all refinement is done in Stage 2 and the estimates are then improved as if we have chosen the optimal smoother in Stage 1. To achieve computation efficiency, we propose to use a common smoother in stage 1 for all individual curves. On the other hand, this could be shown as equivalent as adopting a fast bivariate smoothing approach on the covariance function (Xiao, Li and Ruppert 2013). However, we do not rely on extracting the factors from the covariance function. Though this could be achieved similarly by regressing the smooth curves on $\hat{\psi}_{ij}$ to obtain $\hat{\phi}_j(s)$, we prefer not to do so as such an approach is very sensitive to the choice of the smoothing parameter in Stage 1.

To study the asymptotics, we consider the following conditions:

- C1 Suppose $E(Y_i^{2+\delta}(s)) < \infty$ and $Y_i(s) = X_i(s) + \epsilon_{is}$ instead, where the error terms ϵ_{is} satisfy $E(\epsilon_{is}|X_i) = 0$ and $var(\epsilon_{is}|X_i)$ is finite. The random function $X_i(s)$ admits representation (2) and the factors $\phi_j(s)$ has $2m$ th continuous derivative.
- C2 We use the p th degree B-splines to model $\phi_j(s)$. The knots are equally spaced on \mathcal{F} and the number of knots K satisfies $K \sim Cp^\gamma$ for some fixed

constant γ^1 .

C3 We adopt the m th order penalty and denote λ_c as the common penalty parameter used in Stage 1. Denote λ_j as the penalty parameter for $\phi_j(s)$ used in Stage 2. For any penalty parameter λ , define the equivalent bandwidth as

$$h = K^{-1}(\lambda K p^{-1})^{1/(2m)}. \quad (3)$$

Denote h_c and h_j based on λ_c and λ_j accordingly. Moreover, assume $h_c \rightarrow 0$ and $ph_c \rightarrow \infty$.

Theorem 2.1. *Assume Conditions C1-C3, then we have*

$$\hat{\psi}_{ij} \rightarrow \psi_{ij} \text{ in prob}$$

Remark 1: Condition 1 impose regularity conditions on the observations. As we adopt the penalized approach for estimation, condition 2 and 3 discuss settings on the choice of penalty and the splines. As pointed out by Li and Ruppert (2008), the placement of the knots are not crucial as long as the number of knots exceed some threshold. Under this large penalty setting, the key smoothing parameter is the penalty term and the penalized smoothing approach using penalty λ is equivalent to the kernel smoothing approach using an equivalent bandwidth h defined as equation (3). The last assumption on h_c is very mild and it yields consistent estimate of the smooth function $X_i(s)$ as well as the loadings ψ_{ij} .

Theorem 2 below shows that we could achieve higher estimation efficiency if we carefully tune the smoothing parameter in Stage 2.

Theorem 2.2. *Assume Conditions C1-C3. Suppose s is univariate. Let $h_j \sim p^{-1/5}$. Then we have $p^{2/5}(\hat{\phi}_j(s) - \psi_{ij}) \Rightarrow N(B(s), V(s))$, where $B(s)$ and $V(s)$ are the asymptotic bias and variance.*

¹In the univariate case, the minimal choice of γ is $1/5$ if $p \geq 1$ and $2/5$ if $p = 0$.

Remark 2: We obtain the optimal nonparametric rate for our estimate $\hat{\phi}_j(s)$. Note that this rate is faster than the rate of $\min(\sqrt{n}, \sqrt{p})$ in traditional PCA analysis (Bai and Li 2012, Bai and Ng 2013) or \sqrt{n} in dense functional PCA analysis (Hall and Hosseini-Nasab 2006, Li and Hsing 2010). We are able to achieve a faster rate as we adopt the representation (2) which avoids estimating the mean function μ .

3 Extensions

Now we extend our procedures by allowing multiple tasks for each individual. We still consider a balanced design such that each subject has the same number of visits. The number of tasks J could be much large than the number of individuals n , but $n * J$ is still far smaller than p . To represent the data, we consider the functional two-way ANOVA approach (Di et al. 2009, Staicu et al. 2014),

$$X_{ik}(s) = \mu_i(s) + Z_k(s) + W_{ik}(s),$$

where $\mu_i(s)$ are the individual-specific curves, $Z_k(s)$ are the task-specific curves and $W_{ik}(s)$ are the subject- and task-specific deviation.

By the KL decomposition, we could write X_{ik} as

$$X_{ik}(s) = \mu_i(s) + \sum_{ij} \psi_{ij,1} \phi_{j,1}(s) + \sum_{ij} \psi_{ij,2} \phi_{j,2}(s),$$

where $\phi_{j,1}$ and $\phi_{j,2}$ denote the level 1 and level 2 factors and $\psi_{ij,1}$ and $\psi_{ij,2}$ are their corresponding loadings. Note that $\phi_{j,1}$ for $j = 1, 2, \dots$ and $\phi_{j,2}$ for $j = 1, 2, \dots$ are assumed to be orthonormal bases, but they are not required to be mutually orthogonal. When measurement errors are presented, then $Y_{ik}(s) = X_{ik}(s) + e_{iks}$ are the observations we collect.

We could extend our algorithm for multilevel PCA analysis as below:

1. Average $Y_{ik}(s)$ across k to obtain $\bar{Y}_i(s)$.

2. Perform two stage analysis on $Y_{ik}(s) - \bar{Y}_i(s)$ to estimate $\psi_{ij,2}$ and $\phi_{j,2}$.
3. Define $\hat{W}_{ik}(s) = \sum_{ij} \hat{\psi}_{ijk,2} \hat{\phi}_{j,2}(s)$.
4. Perform two stage analysis on $Y_{ik}(s) - \hat{W}_{ik}(s)$ to estimate $\psi_{ij,1}$ and $\phi_{j,1}$.

Under the two way functional ANOVA setting, we are able to capture the task specific variations as well as the subject- and task-specific deviation. We expect a further analysis on the loadings $\psi_{ij,1}$ may help us to better distinguish the importance of our tasks in studying risk perception, while the loadings $\psi_{ij,2}$ may reveal more information about personal risk preference. Further studies along this area may open more new research direction in neuroeconomic analysis.

4 Reference

- Bai, J. and Li, K. (2012). Statistical Analysis of Factor Models of High Dimension. *Annals of Statistics*, 40, 1, 436-465.
- Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors, *Journal of Econometrics*, 176, 1, 18-29.
- Chen, H.Q., Fang, Y. and Li, Y.X., 2015, Estimation and Inference for Varying-coefficient Models with Nonstationary Regressors using Penalized Splines. *Econometric Theory*, 31(4), 753-777.
- Chen, Y., Hardle, W. K., He, Q. and Majer, P. (2015), Risk Related Brain Regions Detected with 3D Image FPCA, working paper.
- Di, C.Z., Crainiceanu, C.M, Caffo, B.S and Punjabi, N.M. (2009). Multi-level Functional Principal Component Analysis. *Annals of Applied Statistics* 3(1):458-488.
- Hall, P. and Hosseini-Nasab, M. (2006). On Properties of Functional Principal Components Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 68, No. 1, pp. 109-126

- Heller, R., Stanley, D., Yekutieli, D., Rubin, N. and Benjamini, Y. (2006). Clusterbased analysis of FMRI data, *NeuroImage* 33(2): 599-608.
- Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics* 38, 6, 3321-3351.
- Li, Y. and Ruppert, D. (2008). On the Asymptotics of Penalized Splines. *Biometrika*, 95, 415-436.
- Majer, P., Mohr, P. N., Heekeren, H. R. and Hardle, W. K. (2015). Portfolio decisions and brain reactions via the CEAD method, *Psychometrika*. DOI: 10.1007/s11336-015-9441-5.
- Mohr, P., Biele, G. and Heekeren, H. (2010). Neural Processing of Risk, *Journal of Neuroscience* 30(19): 6613-6619.
- Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, Springer, New York.
- Rangel, A., Camerer, C. and Montague, P. (2008). A framework for studying the neurobiology of value-based decision making, *Nat Rev Neurosci* 9: 545-556.
- Rao, C. R. (1958). Some statistical methods for comparison of growth curves, *Biometrics* 14(1): 1-17.
- Staicu, A. M., Li, Y., Crainiceanu, C. M., and Ruppert, D. (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, 41(4), 932-949. DOI: 10.1111/sjos.12075
- Xiao, L., Li, Y. and Ruppert, D. (2013). Fast Bivariate Penalized Splines, *Journal of the Royal Statistical Society Series B*.
- van Bommel, A., Song, S., Majer, P., Mohr, P. N. C., Heekeren, H. R. and Hardle, W. K. (2013). Risk Patterns and Correlated Brain Activities.

Multidimensional Statistical Analysis of fMRI Data in Economic Decision Making Study, *Psychometrika* . DOI: 10.1007/s11336-013-9352-2.

Viviani, R., Gron, G. and Spitzer, M. (2005). Functional principal component analysis of fMRI data, *Human Brain Mapping* 24: 109-129.

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S. and Crainiceanu, C. (2011). Functional principal component model for highdimensional brain imaging, *NeuroImage* 58(3): 772-784.

SFB 649 Discussion Paper Series 2017

For a complete list of Discussion Papers published by the SFB 649, please visit <http://sfb649.wiwi.hu-berlin.de>.

- 001 "Fake Alpha" by Marcel Müller, Tobias Rosenberger and Marliese Uhrig-Homburg, January 2017.
- 002 "Estimating location values of agricultural land" by Georg Helbing, Zhiwei Shen, Martin Odening and Matthias Ritter, January 2017.
- 003 "FRM: a Financial Risk Meter based on penalizing tail events occurrence" by Lining Yu, Wolfgang Karl Härdle, Lukas Borke and Thijs Benschop, January 2017.
- 004 "Tail event driven networks of SIFIs" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle and Yarema Okhrin, January 2017.
- 005 "Dynamic Valuation of Weather Derivatives under Default Risk" by Wolfgang Karl Härdle and Maria Osipenko, February 2017.
- 006 "RiskAnalytics: an R package for real time processing of Nasdaq and Yahoo finance data and parallelized quantile lasso regression methods" by Lukas Borke, February 2017.
- 007 "Testing Missing at Random using Instrumental Variables" by Christoph Breunig, February 2017.
- 008 "GitHub API based QuantNet Mining infrastructure in R" by Lukas Borke and Wolfgang K. Härdle, February 2017.
- 009 "The Economics of German Unification after Twenty-five Years: Lessons for Korea" by Michael C. Burda and Mark Weder, April 2017.
- 010 "Data Science & Digital Society" by Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, May 2017.
- 011 "The impact of news on US household inflation expectations" by Shih-Kang Chao, Wolfgang Karl Härdle, Jeffrey Sheen, Stefan Trück and Ben Zhe Wang, May 2017.
- 012 "Industry Interdependency Dynamics in a Network Context" by Ya Qian, Wolfgang Karl Härdle and Cathy Yi-Hsuan Chen, May 2017.
- 013 "Adaptive weights clustering of research papers" by Larisa Adamyan, Kirill Efimov, Cathy Yi-Hsuan Chen, Wolfgang K. Härdle, July 2017.
- 014 "Investing with cryptocurrencies - A liquidity constrained investment approach" by Simon Trimborn, Mingyang Li and Wolfgang Karl Härdle, July 2017.
- 015 "(Un)expected Monetary Policy Shocks and Term Premia" by Martin Kliem and Alexander Meyer-Gohde, July 2017.
- 016 "Conditional moment restrictions and the role of density information in estimated structural models" by Andreas Tryphonides, July 2017.
- 017 "Generalized Entropy and Model Uncertainty" by Alexander Meyer-Gohde, August 2017.
- 018 "Social Security Contributions and the Business Cycle" by Anna Almosova, Michael C. Burda and Simon Voigts, August 2017.
- 019 "Racial/Ethnic Differences In Non-Work At Work" by Daniel S. Hamermesh, Katie R. Genadek and Michael C. Burda, August 2017.
- 020 "Pricing Green Financial Products" by Awdesch Melzer, Wolfgang K. Härdle and Brenda López Cabrera, August 2017.
- 021 "The systemic risk of central SIFIs" by Cathy Yi-Hsuan Chen and Sergey Nasekin, August 2017.
- 022 "Das deutsche Arbeitsmarktwunder: Eine Bilanz" by Michael C. Burda and Stefanie Seele, August 2017.

SFB 649, Spandauer Straße 1, D-10178 Berlin
<http://sfb649.wiwi.hu-berlin.de>

This research was supported by the Deutsche
Forschungsgemeinschaft through the SFB 649 "Economic Risk".



SFB 649 Discussion Paper Series 2017

For a complete list of Discussion Papers published by the SFB 649, please visit <http://sfb649.wiwi.hu-berlin.de>.

- 023 "Penalized Adaptive Method in Forecasting with Large Information Set and Structure Change" by Xinjue Li, Lenka Zbonakova and Wolfgang Karl Härdle, September 2017.
- 024 "Smooth Principal Component Analysis for High Dimensional Data" by Yingxing Li, Wolfgang K. Härdle and Chen Huang, September 2017.

SFB 649, Spandauer Straße 1, D-10178 Berlin
<http://sfb649.wiwi.hu-berlin.de>

This research was supported by the Deutsche
Forschungsgemeinschaft through the SFB 649 "Economic Risk".

