Kernel Conditional Factor Models

Pierre Collin-Dufresne[1,2] Damir Filipović[1,2] **Urban Ulrych**[1,2]

[1]École Polytechnique Fédérale de Lausanne

[2]Swiss Finance Institute

EEA-ESEM 2024

Erasmus School of Economics, Rotterdam

Version: August 2024

1. Introduction

- *Factor models* are one of the fundamental tools in finance, used to uncover the relationship between *asset returns* and *underlying factors*:

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t} f_{t+1} + \epsilon_{i,t+1}.$$

- Factor models traditionally perform *factor* (f_{t+1}) and *beta* ($\beta_{i,t}$) *learning* in a linear manner, which does *not* capture the *non-linear dynamics* of financial markets.

- Our paper, Collin-Dufresne, Filipović, and Ulrych (2024), proposes an innovative approach to enhance factor models by *incorporating non-linearity* through *(low-rank) kernel functions*.

- Utilize a reproducing kernel Hilbert space (RKHS) with the associated reproducing kernel as *hypothesis space* for modeling *factor loadings*, and employ *cross-sectional ridge regression* to directly learn *factors* (i.e., factor portfolios).

- The proposed framework *improves* the *accuracy* and *predictive power* of factor models in asset pricing.

2. Kernel Factors

- Let $x_{t,i} \in \mathcal{X}$ denote the *characteristics* of assets $i \in \mathcal{I}_t$, where $\mathcal{X}$ denotes the set of observable asset characteristics.

- Let $m \geq 1$ represent the number of factors.

- *Excess returns* of assets $i \in \mathcal{I}_t$ are, in vector notation, given as

$$\boldsymbol{r}_{t+1} = g(\boldsymbol{x}_t)F_{t+1} + \boldsymbol{\epsilon}_{t+1}.$$

- Objective: *Learn* the *factor loading* function $g = (g_1, \ldots, g_m) : \mathcal{X} \to \mathbb{R}^{1\times m} \cong \mathbb{R}^m$, and *factors* $F_{t+1} \in \mathbb{R}^m$ by a cross-sectional ridge regression, similar as in Kelly et al. (2019).

- As hypothesis space for g we select $\mathbb{R}^m$-valued reproducing kernel Hilbert space (RKHS) $\mathcal{G} = \mathcal{H} \otimes \mathbb{R}^m$ with operator-valued reproducing kernel $K(x, x') = k(x, x')I_m$ on $\mathcal{X}$.

2. Alternating Kernel Ridge Regression

- We then solve the *regularized optimization problem*

$$\min_{g \in \mathcal{G},\, F \in \mathbb{R}^{m \times T}} \left\{ \mathcal{E}(g, F) + \lambda_1 \|g\|_{\mathcal{G}}^2 + \lambda_2 \|F\|_2^2 \right\}, \tag{1}$$

for an error function $\mathcal{E} : \mathcal{G} \times \mathbb{R}^{m \times T} \to \mathbb{R}_{>0}$ and some penalty parameters $\lambda_1, \lambda_2 > 0$.

- Using the (weighted) mean squared error function, (1) can be equivalently written as

$$\min_{g \in \mathcal{G}} \left\{ \sum_{t=0}^{T-1} \min_{F_{t+1} \in \mathbb{R}^m} \left\{ \sum_{i \in \mathcal{I}_t} \omega_{t,i} \big(r_{t+1,i} - g(x_{t,i}) F_{t+1} \big)^2 + \lambda_2 \|F_{t+1}\|_2^2 \right\} + \lambda_1 \|g\|_{\mathcal{G}}^2 \right\}, \tag{2}$$

which reflects the *alternating kernel ridge regression algorithm* for solving it.

- This is a *generalization* of the linear alternating least-squares approach of Kelly et al. (2019) towards *non-linear factor loadings*.

2. Cross-Sectional Ridge Regression

- For a given function $g \in \mathcal{G}$ and for each $t = 0, \ldots, T-1$, we solve the *cross-sectional ridge regression*:

$$\min_{F_{t+1} \in \mathbb{R}^m} \left\{ \left(\boldsymbol{r}_{t+1} - g(\boldsymbol{x}_t) F_{t+1} \right)^\top \boldsymbol{\Omega}_t \left(\boldsymbol{r}_{t+1} - g(\boldsymbol{x}_t) F_{t+1} \right) + \lambda_2 \| F_{t+1} \|_2^2 \right\}. \tag{3}$$

- The solution to (3) is *unique* and, for each t, *explicitly* given by

$$\hat{F}_{t+1} = \left(g(\boldsymbol{x}_t)^\top \boldsymbol{\Omega}_t g(\boldsymbol{x}_t) + \lambda_2 \boldsymbol{I}_m \right)^{-1} g(\boldsymbol{x}_t)^\top \boldsymbol{\Omega}_t \boldsymbol{r}_{t+1},$$

reflecting that $\hat{F}_{t+1}$ is a *factor portfolio*.

2. Time-Series Kernel Ridge Regression

- Conversely, for given factors $F_{t+1} \in \mathbb{R}^m$, solving the outer optimization of (2) for g amounts to *time-series kernel ridge regression*:

$$\min_{g \in \mathcal{G}} \left\{ \sum_{t=0}^{T-1} \left(\boldsymbol{r}_{t+1} - g(\boldsymbol{x}_t) F_{t+1} \right)^\top \boldsymbol{\Omega}_t \left(\boldsymbol{r}_{t+1} - g(\boldsymbol{x}_t) F_{t+1} \right) + \lambda_1 \|g\|_{\mathcal{G}}^2 \right\}. \tag{4}$$

- By the *representer theorem*, this would lead to a ridge regression of dimension $M = \sum_{t=0}^{T-1} M_t$, which is too *computationally costly*.

- Following Filipović et al. (2023), we compute a *low-rank approximation* of the *kernel function*

$$k(x, x') \approx \sum_{j=1}^{d} \phi_j(x) \phi_j(x'),$$

for the orthonormal functions $\phi(\cdot) = (\phi_1(\cdot), \ldots, \phi_d(\cdot))$ in $\mathcal{H}$.

2. Low-Rank Approximation

- Accordingly, we replace the full hypothesis space $\mathcal{G}$ by the subspace $\mathcal{G}_\phi$ spanned by $\phi_j(\cdot)v_j$, for $v_j \in \mathbb{R}^m$ and $j = 1, \ldots, d$. Any g in this subspace is of the form

$$g(\cdot) = \sum_{j=1}^{d} \phi_j(\cdot) v_j, \quad v_j \in \mathbb{R}^m, \tag{5}$$

  and problem (4) becomes *quadratic* in $\boldsymbol{v} = [v_1; \ldots; v_d] \in \mathbb{R}^{dm}$.

- Differentiating the objective function (4) in $\boldsymbol{v}$ yields the FOC with the *unique solution*:

$$\hat{\boldsymbol{v}} = \left(\left(\sum_{t=0}^{T-1} \left(\phi(\boldsymbol{x}_t)^\top \Omega_t \phi(\boldsymbol{x}_t) \right) \otimes \left(F_{t+1} F_{t+1}^\top \right) \right) + \lambda_1 I_{dm} \right)^{-1} \left(\sum_{t=0}^{T-1} \left(\left(\phi(\boldsymbol{x}_t)^\top \Omega_t \right) \otimes F_{t+1} \right) \boldsymbol{r}_{t+1} \right).$$

3. Kernel Selection

- Analyzed kernels:
  - Linear: $k(\boldsymbol{x}, \boldsymbol{y}) = 1 + \frac{\boldsymbol{x}^T \boldsymbol{y}}{l^2}$
  - Quadratic: $k(\boldsymbol{x}, \boldsymbol{y}) = (1 + \frac{\boldsymbol{x}^T \boldsymbol{y}}{l^2})^2$
  - Gaussian: $k(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2l^2}}$

- Extend the approach by *including industry classification* as an additional characteristic, encoded by $a \in \{1, \ldots, A\}$.

- At any t and for any asset i, the original characteristics $x_{t,i}$ are thus extended to $(x_{t,i}, a_{t,i})$.

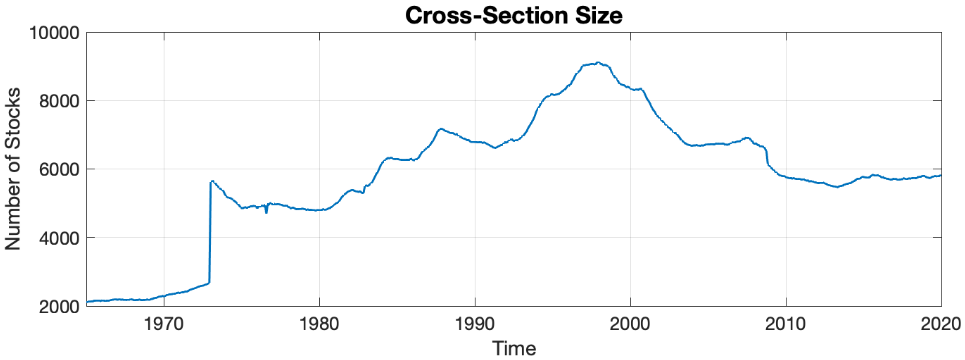
- A simple separable approach for *incorporating industries*:
  - $\bar{k}((\boldsymbol{x}, \boldsymbol{a}), (\boldsymbol{y}, \boldsymbol{b})) = k(\boldsymbol{x}, \boldsymbol{y}) k_a(\boldsymbol{a}, \boldsymbol{b})$, with $k_a(\boldsymbol{a}, \boldsymbol{b}) = \begin{cases} 1, & \boldsymbol{a} = \boldsymbol{b} \\ \rho, & \text{else} \end{cases}$

    for $\rho \in [0, 1]$.

3. Empirical Analysis

- Conducted an *Out-of-Sample* (OOS) *analysis* of the proposed smart kernel factor model.
- Analyzed *US stocks* from January 1965 to December 2019 with *monthly cross-sections* and 94 characteristics per stock.
- Rank-normalized all characteristics into the interval $(-1, 1)$ for each month t.
- Incorporated *industry classifications* $(A = 11)$ based on Standard Industrial Classification (SIC) two-digit codes.
- Utilized a *training period* of 10 years, followed by *hyperparameter validation* on the subsequent 5 years.
- Performed a *rolling window out-of-sample backtest* with 10 years of training and 1 year of testing.

3. Empirical Analysis

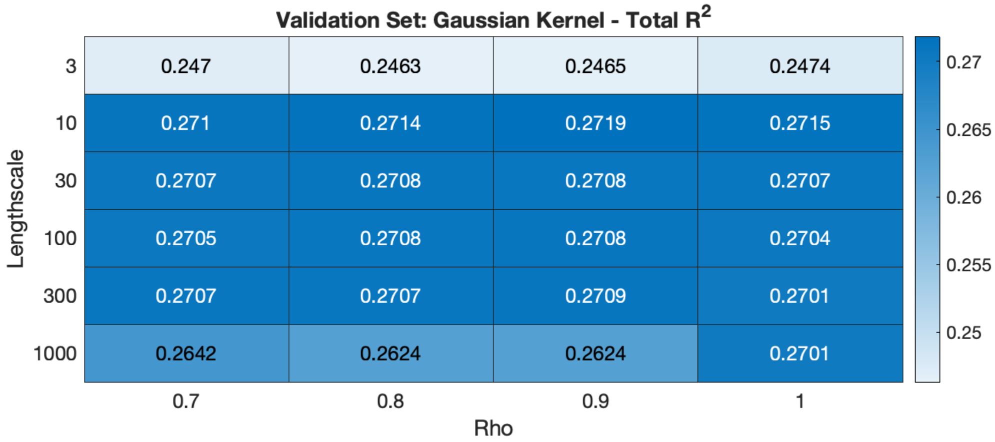


3. Empirical Performance

- $R^2_{Total} = 1 - \frac{\sum_{(i,t)\in OOS}(r_{i,t} - \hat{g}(x_{i,t-1})\hat{F}_t)^2}{\sum_{(i,t)\in OOS} r^2_{i,t}}$, $R^2_{Pred} = 1 - \frac{\sum_{(i,t)\in OOS}(r_{i,t} - \hat{g}(x_{i,t-1})\hat{\lambda}_{t-1})^2}{\sum_{(i,t)\in OOS} r^2_{i,t}}$, where $\hat{\lambda}_{t-1}$ is the prevailing sample average of $\hat{F}$ up to month $t-1$.
- Comparison of OOS R^2_{Total} and R^2_{Pred} for different methods and $m = 5$ factors:

Method ($m = 5$)	R^2_{Total} **(%)**	R^2_{Pred} **(%)**
$1/N$ Portfolio	8.80	0.02
Kelly et al.	14.11	< 0
Linear Kernel ($d = 94$)	14.13	0.65
Quadratic Kernel ($d = 200$)	14.42	0.22
Gaussian Kernel ($d = 200$)	14.38	0.36
Linear with Industry ($d = 95$)	14.20	0.74
Quadratic with Industry ($d = 200$)	14.43	0.11
Gaussian with Industry ($d = 200$)	14.40	0.06

3. Gaussian Kernel: Validation & Effect of Industry



4. Conclusion

- This paper presents a novel approach for learning *factors* and *factor loadings* in a *non-linear* manner using *kernel-based methodology*.

- We *extend* existing linear-learning-based approaches by introducing:
  1. *Non-linear dependence* on characteristics, allowing for greater flexibility in modeling factor relationships. Notably, our linear kernel specification *nests* Kelly et al. (2019).
  2. *Regularization*, allowing for more factors and *improving OOS performance* in terms of explained variation (R^2).
  3. Additional characteristics, such as *industries*, potentially enhancing model accuracy.

- A preliminary *empirical analysis* demonstrates that our kernel-based extension *outperforms* current linear-learning-based models in terms of OOS R^2, demonstrating the *effectiveness* of *non-linear learning* and *regularization*.

- An *interpretable* approach with *fast* computation speed $\Longrightarrow$ *practical* applicability!

Thank you for your attention!

References I

P. Collin-Dufresne, D. Filipović, and U. Ulrych. Smart kernel factors. *Work in Progress*, 2024.

D. Filipović, M. Multerer, and P. Schneider. Kernel conditional distribution machines. *Working Paper*, 2023.

B. T. Kelly, S. Pruitt, and Y. Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019.

Appendix: Low-Rank Approximation

- We assume that the kernel matrix $\boldsymbol{K} := k(\boldsymbol{X}, \boldsymbol{X}^\top)$ admits a *low-rank approximation* $\boldsymbol{K} \approx \boldsymbol{L}\boldsymbol{L}^\top$ for some $M \times d$-matrix $\boldsymbol{L}$ and such that there exists a bi-orthogonal $M \times d$ matrix $\boldsymbol{B}$ with
  - $\boldsymbol{K}\boldsymbol{B} = \boldsymbol{L}$,
  - $\boldsymbol{B}^\top \boldsymbol{L} = \boldsymbol{I}_d$, and
  - $\operatorname{Im} \boldsymbol{B} = \operatorname{span}\{e_{\pi_1}, \dots, e_{\pi_d}\}$,

  for some pivot indices $\{e_{\pi_1}, \dots, e_{\pi_d}\} \subseteq \{1, \dots, M\}$. That is, only d rows of $\boldsymbol{B}$ are different from zero.

- Matrices $\boldsymbol{B}$ and $\boldsymbol{L}$ can be computed recursively, see Filipović et al. (2023).

- This yields the *low-rank approximation* of the *kernel function*

$$k(x, x') \approx \sum_{j=1}^{d} \phi_j(x)\phi_j(x'),$$

for the orthonormal functions $\phi(\cdot) = (\phi_1(\cdot), \dots, \phi_d(\cdot))$ in $\mathcal{H}$ given by $\phi(\cdot) := \boldsymbol{B}^\top k(\cdot, \boldsymbol{X})$.