Multi-Target Forecasting using Conditional Mean Embedding

FRE Lab Seminar (2024-08-01)

Giyeong Lee

Table of Contents

Motivation

- **Measures in Finance**
  - Statistics for asset returns
  - Performance measures
  - Risk measures
  - …

- **Q. Is there no unified framework to forecast various financial measures?**
  - Lots of financial measures admit the following form:

$$\mathbb{E}(f(Y)|X = x)$$

    where x is a past observation of covariates X and Y is a target variable
  - Might be possible if we know the distribution $Y|X = x$ and a function f is given as an input

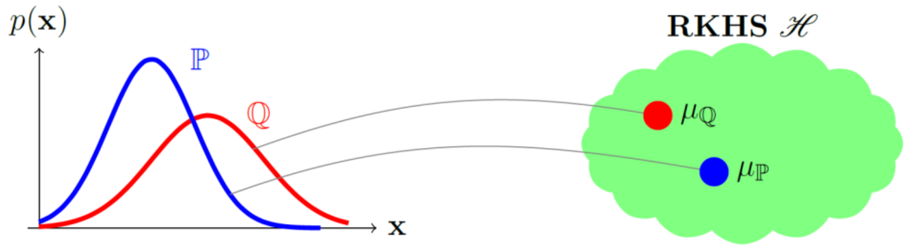
Kernel Mean Embedding

- **Maximum Mean Discrepancy (MMD)**
  - Measures difference between two distributions P, Q using a class of some functions $\mathcal{F}$

$$\mathrm{MMD}(P, Q; \mathcal{F}) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y) \right)$$

- **Kernel Mean Embedding (KME)**
  - Embeds probability distributions into reproducing kernel Hilbert spaces
  - Requires more relaxed conditions than likelihood-based models
  - Gretton, et al. (2012) proved MMD can be evaluated efficiently using KME
    - For some RKHS $\mathcal{H}$ (which depends on $\mathcal{F}$),

$$\mathrm{MMD}(P, Q; \mathcal{F})^2 = \left\| \mu_P - \mu_Q \right\|_{\mathcal{H}}^2$$

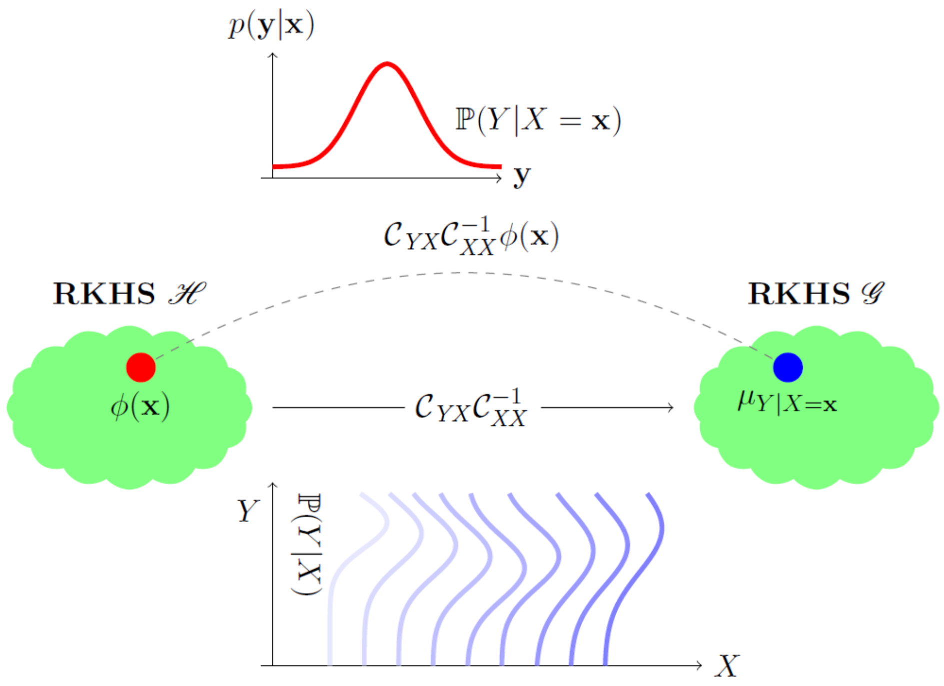


Muandet, et al. (2017)

Conditional Mean Embedding

- **Conditional Mean Embedding (CME)**
  - Embeds **conditional distributions** into RKHSs
  - Enables us to evaluate conditional expectations via the inner product

- **Definition**
  - CME $\mu_{Y|X=x}$ of $P(Y|X=x)$ is the element of an RKHS such that
    1. $\mu_{Y|X=x} = \mathbb{E}(\psi(Y)|X=x)$ $(\psi$: canonical feature map of $\mathcal{G})$
    2. $\mathbb{E}(f(Y)|X=x) = \langle f, \mu_{Y|X=x} \rangle_{\mathcal{G}}$ for all $f \in \mathcal{G}$
  - Denote as $\mu_{Y|x} = \mu_{Y|X=x}$ briefly



Muandet, et al. (2017)

Reproducing Kernel Hilbert Space

- **Hilbert Space**
  - An inner product space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ such that $\mathcal{H}$ is complete with respect to the norm $\|f\| = \langle f, f\rangle$

- **Reproducing Kernel Hilbert Space (RKHS)**
  - A Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ of functions on $\mathcal{X}$ such that
    1. $k(x,\cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$
    2. $f(x) = \langle f, k(x,\cdot)\rangle$ for all $x \in \mathcal{X}, f \in \mathcal{H}$
  - The second condition is called the <u>**reproducing property**</u>
  - The map $\phi: x \in \mathcal{X} \mapsto k(x,\cdot) \in \mathcal{H}$ is called the canonical feature map of $\mathcal{H}$

Empirical Estimator of CMEs

- **Song, Fukumizu, Gretton. (2013)**

$$\hat{\mu}_{Y|x} = \sum_{\ell=1}^{N} w_\ell(x)\psi(y_\ell) \quad \text{where} \quad \begin{pmatrix} w_1(x) \\ \vdots \\ w_N(x) \end{pmatrix} = \left(\left(k(x_i, x_j)\right)_{1\le i,j\le N} + \lambda I\right)^{-1} \begin{pmatrix} k(x, x_1) \\ \vdots \\ k(x, x_N) \end{pmatrix}$$

- Evaluation of conditional expectations
  - Reproducing property gives

$$\widehat{\mathbb{E}}(f(Y)|X = x) = \langle f, \hat{\mu}_{Y|x} \rangle_{\mathcal{G}} = \sum_{\ell=1}^{N} w_\ell(x) f(y_\ell)$$

- Requires a matrix inversion
  - Not scalable to large datasets

Estimator of CMEs – Neural Network Approach

- **Grünewälder, et al. (2012)**
  - Provides an interpretation in the view of function-valued regression

$$\underset{C:\mathcal{H}\to\mathcal{G}}{\operatorname{argmin}} \frac{1}{N}\sum_{\ell=1}^{N} \|\psi(y_i) - C\phi(x_i)\|_{\mathcal{G}}^2 + \lambda\|C\|_{\mathrm{HS}}^2$$

- **Simizu, Fukumizu, Sejdinovic. (2024)**
  - Suggests the NN-based estimator of CMEs based on the above interpretation
  - $w_a(\cdot;\theta)$: weight map implemented by neural networks
  - η_a: fixed or learnable location parameter

$$\hat{\mu}_{Y|x} = \sum_{a=1}^{M} w_a(x;\theta)\psi(\eta_a)$$

Simple Experiment & Comparison

- **Data: 10 Currencies (EUR, GBP, AUD, NZD, CAD, CHF, SGD, KRW, JPY, CNY)**
  - Train: 2014~2020, Validation: 2021~2022
  - Input: log return of adj. close (daily, 60 days)
  - Horizon T: 20 days

- **Metrics**
  - $\mathrm{MSE} = \frac{1}{DT}\sum_{i,t}(\hat{r}_{it} - r_{it})^2$
  - $\mathrm{RV} = \frac{1}{D}\sum_i \sqrt{r_{i1}^2 + \cdots + r_{iT}^2}$
  - …

		GRU + Point Prediction (MSE)	**GRU + CME**
#parameters		10,648,264	100,689
MSE	$(r_1, \cdots, r_T)$	6.22×10^{-3}	6.15×10^{-3}
MAE	$\mathrm{std}(r_1, \cdots, r_T)$	4.43×10^{-3} (1.35×10^{-3})	1.03×10^{-3}
	$\max(r_1, \cdots, r_T)$	8.80×10^{-3} (3.56×10^{-3})	3.01×10^{-3}
	$\min(r_1, \cdots, r_T)$	8.36×10^{-3} (3.31×10^{-3})	2.72×10^{-3}
	$\mathrm{MDD}(r_1, \cdots, r_T)$	1.54×10^{-2} (1.09×10^{-2})	1.01×10^{-2}
	$\mathrm{RV}(r_1, \cdots, r_T)$	1.90×10^{-2} (6.04×10^{-3})	4.53×10^{-3}

Future Work

- **Choice of kernel**
  - What kernel is most suitable for time-series data?
    - Signature kernel
  - What class of functions can be represented using CMEs with this kernel?

- **Error correction for each f**

- **Computation cost during training**
  - M locational parameters → Computation cost increases quadratically
  - Might be solved with gradient accumulation (guess)

- **More benchmarks**
  - SOTA for each metric

References

[1] Gretton, Arthur, et al. "A kernel two-sample test." *The Journal of Machine Learning Research* 13.1 (2012): 723-773.

[2] Grünewälder, Steffen, et al. "Conditional mean embeddings as regressors-supplementary." *arXiv preprint arXiv:*1205.4656 (2012).

[3] Muandet, Krikamol, et al. "Kernel mean embedding of distributions: A review and beyond." *Foundations and Trends® in Machine Learning* 10.1-2 (2017): 1-141.

[4] Shimizu, Eiki, Kenji Fukumizu, and Dino Sejdinovic. "Neural-Kernel Conditional Mean Embeddings." *arXiv preprint arXiv:*2403.10859 (2024).

[5] Song, Le, Kenji Fukumizu, and Arthur Gretton. "Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models." *IEEE Signal Processing Magazine* 30.4 (2013): 98-111.