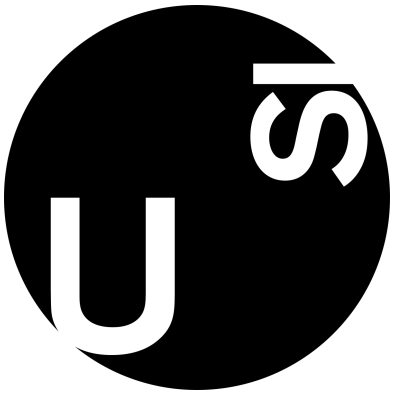

Energy Market Analysis Using Kernel Methods

Master Thesis Preparation

L. Pernigo

December 15, 2023

Advisor: Prof. Dr. M. Multerer

Co-Advisor: Dr. D. Baroli

Faculty of Informatics, USI Lugano

Abstract

The theory of kernel methods will be applied to the problem of probabilistic forecasting taking as input some data. Conceptually, the proposed analysis could be applied to any kind of data. In this study we considered the electricity market, because of the interesting implications on risk management tasks.

Contents

Chapter 1

Problem Description and Literature Review

Kernel methods are a class of algorithms for patter analysis. With kernel methods we are able to apply linear methods with predictors in a high dimensional space, without having to explicitly evaluate the involved dot products of the features. In this thesis work I will address the performance of kernel methods in the context of probabilistic forecasting; the area of application will be the electricity market. Probabilistic forecasting may be useful to power producers, traders and consumers in order to improve their decision making process and managing risk(VaR). This is because probabilistic forecast enables them to simulate scenarios and carry out stress tests.

This problem has been already addressed in [1] and [8]. These papers surveyed the performance of neural network architectures against simpler approaches like quantile regression and data fitting to Johnson distribution. Their conclusion is that distributional NN perform a little worse than quantile regression but the former has smaller computational cost; that is because the quantile regression is run for every quantile from 0.01 to 0.99. Nevertheless kernel methods received very little attention in this specific setting.
Kernel methods considered are kernel mean embedding [6], [9] and kernel herding [4]. Particularly extending the idea of [3], where the Nyström approximation is employed in computing the kernel mean embedding, experiments with the Pivoted Cholesky decomposition will be performed.
In section 2, three papers that lay the basis for this thesis' work are summarized.

Chapter 2

Paper review

2.1 Kernel Mean Embedding of Distributions: A Review and Beyond

From this first paper [9], the notation and terms used in the theory of Reproducing Kernel Hilbert Spaces are summarized.
Many algorithms use the inner product as similarity measure between data instances $x, x' \in \mathcal{X}$. However, this inner product spans only the class of linear similarity measures.
The idea behind kernel methods is to apply a non-linear transformation φ to the data x in order to get a more powerful non linear similarity measure.

$$\begin{aligned} \varphi(x) : \mathcal{X} &\longrightarrow \mathcal{F} \\ x &\mapsto \varphi(x) \end{aligned}$$

Then we take the inner product in the high dimensional space $\mathcal{F}$ mapped by $\varphi(x)$.

$$k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

$\varphi(x)$ is referred as feature map while k as kernel function.
Therefore, we can kernelize any algorithm involving a dot product by substituting $\langle x, x' \rangle_{\mathcal{X}}$ with $\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$
One would expect constructing the feature maps explicitly and then evaluate their inner product in $\mathcal{F}$ to be computationally expensive, and indeed it is. However, we do not have to explicitly perform such calculations. This is because of the existence of the kernel trick. To illustrate the idea behind the kernel trick consider the following example.
Suppose $x \in \mathbb{R}^2$ and assume to select $\varphi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$, then the

inner product in the feature space is $x_1^2 x_1'^2, x_2^2 x_2'^2 + 2x_1 x_2 x_1' x_2'$. Notice that this is the same of $\langle \varphi(x), \varphi(x') \rangle$; thus the kernel trick consists of just using $k(x, x') =: (x^T x')^2$.

2.1.1 RKHS

Following are the definitions that make up the basis for the theory of kernel methods.

Definition 2.1 *A sequence $\{v_n\}_{n=1}^{\infty}$ of elements of a normed space $\mathcal{V}$ is a Cauchy sequence if for every $\varepsilon > 0$, there exist $N = N(\varepsilon) \in \mathbf{N}$ such that $\|v_n - v_m\|_{\mathcal{V}} < \varepsilon \ \ \forall m, n \geq N$*

Definition 2.2 *A complete metric space is a metric space in which every Cauchy sequence is convergent.*

Definition 2.3 *A Hilbert space is a vector space $\mathcal{H}$ with an inner product $\langle f, g \rangle$ such that the norm defined by $\|f\| = \sqrt{\langle f, f \rangle}$ turns $\mathcal{H}$ into a complete metric space.*

Definition 2.4 *RKHS. A Reproducing Kernel Hilbert Space is an Hilbert space with the evaluation functionals $\mathcal{F}_x(f) := f(x)$ bounded, i.e. $\forall x \in \mathcal{X}$ there exists some $C > 0$ such that $\|\mathcal{F}_x(f)\| = \|f(x)\| \leq C\|f\|_{\mathcal{H}} \ \forall f \in \mathcal{H}$*

Theorem 2.5 *Riesz Representation. If $A : \mathcal{H} \to \mathbf{R}$ is a bounded linear operator in a Hilbert space $\mathcal{H}$, there exists some $g_A \in \mathcal{H}$ such that $A(f) = \langle f, g_A \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$.*

The Riesz representation theorem results in the following proposition for RKHS.

Proposition 2.6 *For each $x \in \mathcal{X}$ there exists a function $k_x \in \mathcal{H}$ such that $\mathcal{F}_x(f) = \langle k_x, f \rangle_{\mathcal{H}} = f(x)$*

The function k_x is the reproducing kernel for the point x. Furthermore, note that k_x is itself a function lying on $\mathcal{X}$

$$k_x(y) = \mathcal{F}_y(k_x) = \langle k_x, k_y \rangle_{\mathcal{H} = \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}}}$$

2.2 Recovering Distributions from Gaussian RKHS Embeddings

This paper covers the RKHS embedding approach to nonparametric statistical inference [6]. The idea here is computing an estimate of the kernel mean in order to obtain an approximation of the underlying distribution of the observed random variable. The kernel mean embedding $\mu_{\mathbb{P}}$ of a probability $\mathbb{P}$ corresponds to the feauture map $\varphi(x)$ integrated with respect to the $\mathbb{P}$

measure.
That is $\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(x, \cdot) d\mathbb{P}(x)$.
Kernel mean embedding serves as a unique representation of $\mathbb{P}$ in the RKHS $\mathcal{H}$. This holds provided that $\mathcal{H}$ is characteristic.

Definition 2.7 *The RKHS $\mathcal{H}$ and the associated kernel k are said characteristic, when the mapping $\mu : \mathbb{P} \to \mathcal{H}$ is injective.*

When the mapping is injective, we have that $\mu_{\mathbb{P}}$ is uniquely associated with $\mathbb{P}$; thus, $\mu_{\mathbb{P}}$ is a unique representation of $\mathbb{P}$ in $\mathcal{H}$.
Note that, by the reproducing property of RKHS $\langle f, k(x, \cdot)\rangle = f(x)$ we have:

$$E_{\mathbb{P}}[f(x)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}, \ \forall f \in \mathcal{H}$$

Proof

$$\begin{aligned} E_{\mathbb{P}}[f(x)] &= \int_{\mathcal{X}} f(x) d\mathbb{P}(x) \\ &= \int_{\mathcal{X}} \langle f, k(x, \cdot)\rangle_{\mathcal{H}} d\mathbb{P}(x) \\ &= \sum_{i=1}^{\infty} \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}} \mathbb{P}(\mathcal{X}_i) \\ &= \langle f, \sum_{i=1}^{\infty} k(x_i, \cdot)\mathbb{P}(\mathcal{X}_i)\rangle_{\mathcal{H}} \\ &= \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}} \end{aligned}$$

□

The kernel mean embedding can be employed to estimate the density p at any fixed point x_0. Letting δ_{x_0} to be the dirac delta function we have:

$$p(x_0) = \int \delta_{x_0} p(x) dx = E_{\mathbb{P}}[\delta_{x_0}]$$

Therefore, the idea is to define an estimator for the expectation of δ_{x_0} through $\mu_{\mathbb{P}}$; this would result in an estimator of $p(x_0)$.
A kernel $k(x_0, \cdot)$ is used to approximate the delta function, furthermore applying theorem 1 of [6] we have that a consistent estimator of $E_{\mathbb{P}}[k(x_0, \cdot)]$ is given by $\sum_{i=1}^{n} w_i k(x_0, x_i)$
When the weigths are all $1/n$ we end up with the standard kernel density estimation.
Alternatively, the optimal weights can be found by minimizing the following problem $\|\hat{\mu} - \Phi w\|^2$ where $\Phi : \mathbb{R}^n \to \mathcal{H}$.

2.3 Super-Samples from Kernel Herding

Kernel herding is a deterministic sampling algorithm designed to draw "Super Samples" from probability distributions [4]. The idea of herding, is to generate pseudo-samples that greedily minimize the error betweeen the mean operator and the empirical mean operator resulting from the selected herding points.

Letting $p(x)$ be a probability distribution, kernel herding is recursively defined as follows:

$$x_{t+1} = \arg\max_{x \forall \mathcal{X}} \langle w_t, \varphi(x) \rangle$$
$$w_{t+1} = w_t + E_{\mathbb{P}}[\varphi(x)] - \varphi(x_{t+1})$$

w denotes a weight vector that lies in $\mathcal{H}$.
Here, by assuming that the inner product between weights and the mean operator is equal to a general functional f evaluated at x, that is $\langle w, \varphi(x) \rangle_{\mathcal{H}} = f(x)$. We have:

$$\begin{aligned}
\langle w, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} &= \langle w, \int k(x, \cdot) d\mathbb{P}(x) \rangle_{\mathcal{H}} \\
&= \langle w, \sum_{i=1}^{\infty} k(x_i, \cdot) \mathbb{P}(\mathcal{X}_i) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{\infty} \langle w, k(x_i, \cdot) \rangle_{\mathcal{H}} \mathbb{P}(\mathcal{X}_i) \\
&= \sum_{i=1}^{\infty} f(x_i) \mathbb{P}(\mathcal{X}_i) \\
&= \int f(x) d\mathbb{P}(x) \\
&= \mathbb{E}_{\mathbb{P}}[f(x)]
\end{aligned}$$

Moreover, second assumption of the model is that $\|\varphi(x)\|_{\mathcal{H}} = R \quad \forall x \in X$. That is the Hilbert space norm of the feature vector is equal to a constant R for all states in the set $\mathcal{X}$.

This can be achieved by taking the new feature vector as $\varphi^{new}(x) = \frac{\varphi(x)}{\|\varphi(x)\|_{\mathcal{H}}}$. See A.1 for details.

By rewriting the formula for the weights we end up with

$$\begin{aligned} w_{t+1} &= w_t + E_{\mathbb{P}}[\varphi(x)] - \varphi(x_{t+1}) \\ &= w_{t-1} + E_{\mathbb{P}}[\varphi(x)] + E_{\mathbb{P}}[\varphi(x)] - \varphi(x_{t+1}) - \varphi(x_t) \\ &= w_{t-2} + E_{\mathbb{P}}[\varphi(x)] + E_{\mathbb{P}}[\varphi(x)] + E_{\mathbb{P}}[\varphi(x)] - \varphi(x_{t+1}) - \varphi(x_t) - \varphi(x_{t-1}) \end{aligned}$$

Considering w_T, we have

$$w_T = w_0 + TE_{\mathbb{P}}[\varphi(x)] - \sum_{t=1}^{T} \varphi(x_t)$$

Note that $E_{\mathbb{P}}[\varphi(x)] = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x)$ which corresponds to the definition of $\mu_{\mathbb{P}}$. That is μ is the mean operator associated with the distribution $\mathbb{P}$; it lies in $\mathcal{H}$.
Thus,

$$w_T = w_0 + T\mu_{\mathbb{P}} - \sum_{t=1}^{T} \varphi(x_t)$$

Notice we do not have to compute $\mu_{\mathbb{P}}$ explicitly, the terms involving $\mu_{\mathbb{P}}$ will be computed by applying the kernel trick.

Now we have everything we need in order to reformulate the original problem in a way such that it depends just on the states x. Plug the formula for the weights in the formula for the x_t and use the kernel trick; we end up with

$$\begin{aligned} x_{T+1} &= \arg\max_{x \forall \mathcal{X}} \langle w_0 + T\mu_{\mathbb{P}} - \sum_{t=1}^{T} \varphi(x_t), \varphi(x) \rangle_{\mathcal{H}} \\ &= \arg\max_{x \forall \mathcal{X}} \langle w_0, \varphi(x) \rangle_{\mathcal{H}} + \langle T\mu_{\mathbb{P}}, \varphi(x) \rangle_{\mathcal{H}} - \langle \sum_{t=1}^{T} \varphi(x_t), \varphi(x) \rangle_{\mathcal{H}} \\ &= \arg\max_{x \forall \mathcal{X}} \langle w_0, \varphi(x) \rangle_{\mathcal{H}} + T\langle \mu_{\mathbb{P}}, \varphi(x) \rangle_{\mathcal{H}} - \sum_{t=1}^{T} k(x_t, x) \end{aligned}$$

Notice $\langle \mu_{\mathbb{P}}, \varphi(x) \rangle_{\mathcal{H}}$ can be rewritten in the following way

$$
\begin{aligned}
\langle \mu_{\mathbb{P}}, \varphi(x) \rangle_{\mathcal{H}} &= \langle \int_{\mathcal{X}'} \varphi(x') d\mathbb{P}(x'), \varphi(x) \rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{\infty} \varphi(x'_i) \mathbb{P}(\mathcal{X}'_i), \varphi(x) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{\infty} \langle \varphi(x'_i), \varphi(x) \rangle_{\mathcal{H}} \mathbb{P}(\mathcal{X}'_i) \\
&= \sum_{i=1}^{\infty} k(x'_i, x) \mathbb{P}(\mathcal{X}'_i) \\
&= \int_{\mathcal{X}'} k(x', x) d\mathbb{P}(x') \\
&= E_{\mathbb{P}}[k(x', x)]
\end{aligned}
$$

Furthermore, by initializing $w_0 = \mu_{\mathbb{P}}$ we end up with the following function to be optimized, i.e.

$$
\begin{aligned}
x_{T+1} &= \arg\max_{x \forall \mathcal{X}} \langle w_0, \varphi(x) \rangle + T \langle \mu_{\mathbb{P}}, \varphi(x) \rangle - \sum_{t=1}^{T} k(x_t, x) \\
&= \arg\max_{x \forall \mathcal{X}} (T+1) E_{\mathbb{P}}[k(x', x)] - \sum_{t=1}^{T} k(x_t, x)
\end{aligned}
$$

Now consider the error term between the mean kernel operator and its estimation through herding samples

$$
\begin{aligned}
\varepsilon_{T+1} &= \| \mu_{\mathbb{P}} - \frac{1}{T+1} \sum_{i=1}^{T+1} \varphi(x_t) \|^2_{\mathcal{H}} \\
&= \mathbb{E}_{x,x' \sim \mathbb{P}}[k(x', x)] - \frac{2}{T+1} \sum_{t=1}^{T+1} \mathbb{E}_{x \sim \mathbb{P}}[k(x, x_t)] + \frac{1}{(T+1)^2} \sum_{t,t'=1}^{T+1} k(x_t, x_{t'}) \\
&= \mathbb{E}_{x,x' \sim \mathbb{P}}[k(x', x)] - \frac{2}{T+1} \sum_{t=1}^{T+1} \mathbb{E}_{x \sim \mathbb{P}}[k(x, x_t)] + \frac{1}{(T+1)^2} \sum_{\substack{t=1 \\ t=t'}}^{T+1} k(x_t, x_{t'}) + \\
&\quad + \frac{2}{(T+1)^2} \sum_{\substack{t=1 \\ t \neq t'}}^{T+1} k(x_t, x_{t'})
\end{aligned}
$$

So ε_{T+1} depends on x_{T+1} only through $-\frac{2}{T+1} \mathbb{E}_{x \sim \mathbb{P}}[k(x, x_{T+1})] + \frac{2}{(T+1)^2} \sum_{t=1}^{T} k(x_t, x_{T+1})$
The term $k(x_{T+1}, x_{T+1})$ is not included, because by assumption it is equal to

the constant R.
Recognize that this term is the negative of the objective function maximized with respect to x. So the sample x_{T+1} minimizes the error at time step $T+1$, i.e. ε_{T+1}
During the iterative step of kernel herding we maximize the negative of this quantity, thus we are minimizing the error greedily. In the sense that at each iteration we choose the x that minimizes our current error; however this does not guarantee that the samples states are jointly optimal.
Intuitively, at each iteration, herding searches for a new sample to add to the pool; it is attracted to the regions where p is high and pushed away from regions where samples have already been selected.

Chapter 3

Workplan

Workplan is to start by considering the performance of kernel herding in generating an empirical distribution of the electricity prices and compare the performance to the methods in [1]. To do so, with kernel herding we generate samples and as a consequence we also have an empirical distribution of electricity prices. Then we take the quantiles from 0.01 to 0.99 and finally we evaluate the distribution forecast through the Continous Ranked Probability Score CRPS.
In this study, the data from the EPX martket will be used; data is retrieved daily from the data provider through an automatic python script.
Depending on the results of kernel herding, we may also consider how kernelized quantile regression behaves in the same setting. Next, additional kernel methods applied to problems concerning energy prices and related subjects could be taken into account.

Appendix A

Appendix

A.1 Feature Map Normalization

Proof

$$
\begin{aligned}
\|\varphi^{new}(x)\|_{\mathcal{H}}^2 &= \|\frac{\varphi(x)}{\|\varphi(x)\|_{\mathcal{H}}}\|_{\mathcal{H}}^2 \\
&= \|\frac{\varphi(x)}{\sqrt{k(x,x)}}\|_{\mathcal{H}}^2 \\
&= \langle\frac{\varphi(x)}{\sqrt{k(x,x)}}, \frac{\varphi(x)}{\sqrt{k(x,x)}}\rangle_{\mathcal{H}} \\
&= \frac{1}{\sqrt{k(x,x)^2}}\langle\varphi(x), \varphi(x)\rangle_{\mathcal{H}} \\
&= 1
\end{aligned}
$$

□

Bibliography

[1] Ilyas Agakishiev, Wolfgang Karl Härdle, Karel Kozmik, Milos Kopa, and Alla Petukhina. Multivariate probabilistic forecasting of electricity prices with trading applications. *SSRN Electronic Journal*, 01 2023.

[2] Robert Bringhurst. *The Elements of Typographic Style*. Hartley & Marks, 1996.

[3] Antoine Chatalic, Nicolas Schreuder, Alessandro Rudi, and Lorenzo Rosasco. Nyström kernel mean embeddings, 2022.

[4] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. 2012.

[5] Tony Jebara and Risi Kondor. Bhattacharyya expected likelihood kernels. volume 2777, pages 57–71, 01 2003.

[6] Motonobu Kanagawa and Kenji Fukumizu. *Recovering Distributions from Gaussian RKHS Embeddings*, volume 33 of *Proceedings of Machine Learning Research*. PMLR, Reykjavik, Iceland, 22–25 Apr 2014.

[7] Yoshihito Kazashi and Fabio Nobile. Density estimation in RKHS with application to korobov spaces in high dimensions. *SIAM Journal on Numerical Analysis*, 61(2):1080–1102, apr 2023.

[8] Grzegorz Marcjasz, Michał Narajewski, Rafał Weron, and Florian Ziel. Distributional neural networks for electricity price forecasting. *Energy Economics*, 125:106843, September 2023.

[9] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[10] Michael Multerer, Paul Schneider, and Rohan Sen. Fast empirical scenarios, 2023.

[11] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2 edition, 2015.