Energy Market Analysis Using Kernel Methods

Master Thesis

L. Pernigo

June To be defined, 2024

Advisor: Prof. Dr. M. Multerer

Co-Advisor: Dr. D. Baroli

Faculty of Informatics, USI Lugano

PRELIMINARY ABSTRACT, ONCE I AM DONE WRITE THE FINAL ABSTRACT

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.

- Contribution: Kernel herding applied to the field of probabilistic forecasting

Contents

List of Figures

List of Tables

Chapter 1

Problem Description

Individuals and organizations constantly face situations of uncertainty. Thus, the need of robust forecasting methods. Such methods are crucial to the process of taking informed decision and to strategic planning.
The basic idea of forecasting is that we can extract knowledge from the past in order to make educated guess about the future. Consequently, the range of fields where forecasting can be applied is very wide. In this thesis, our focus lies on applying forecasting to the energy sector.

The reason of our decision to focus on the energy market is mainly motivated by the rapid changes it has experienced. Over the last decades, electricity markets have gone through an unprecedented transformation; this shift was driven by the liberalization of such markets, the development and integration of renewable energy sources, the increase of low carbon technologies and the adoption of smart meters. Events like the California electricity crisis help also motivating the choice of the electricity sector as subject of our studies, see [4]. Interestingly, the process of deregulation lead to an increasing interest in the field of electricity price forecasting (EF) within the academic community 2.1.
Moreover, the United Nations have identified the right to access affordable, reliable, sustainable and modern energy as one of their 17 SDGs [1]
Finally, the electricity market has a set of features that make it unique: electricity cannot be stored in an efficient way and supply and demand have to be matched instantly.

1.1 Motivation

The are mutliple reasons why the energy sector needs robust forecasting techniques. For power market companies, being able to predict prices with a

low MAPE 8.1 results in increased savings [19]. Furthermore,the adoption of smart meters provides power market companies with a ton of consumer data; this will enable them to better model consumer preferences.
Transmission system operators main goal is to match supply and deman, generally TSO do so by increasing or decreasing the generation supplied. Thus, from their view point forecasting is critical for balancing the electricity netwrok.
Other possible applications are: control of storage, demand side response, anomaly detection, network design and planning, simulating inputs and hangling missing data.

1.2 Point vs probabilistic forecast

A distinction has to be made between two types of forecasting approaches: point forecasts and probabilistic forecasts. Point prediction, also called deterministic forecasting in the literature, is all about predicting a particular value in time. On the other hand, with probabilistic forecasting we aim at predicting either a prediction interval, quantiles or a probability distribution for each point in time. For this reason, probabilistic forecasts are more informative than point forecasts; and this is why the interest of the research community is shifting towards them. Note that a probabilistic forecast can be turned into a point forecast by simply taking its expecatation. Alternatively, a probabilistic forecast can be derived from a point one by modeling the residuals of the point prediction.

1.3 Aims and objectives

- what is the question? objectives of the thesis the description of the problem tackled and the methodology used to solve it.

The scope of the thesis is analyzing state of the art forecasting methods in the energy market and to compare them with ideas coming from kernel theory. Applications of such methods to

1.4 Outline

We start with a literature review and bibliometric analysis in section 2. Then the theory underlying kernel methods is covered in section 3. Following, section **??** introduces the state of the art methods in the context of both point and probabilistic forecasting. Section 7 explains the core features and terminology of the energy market and of the electricity newtork. Evaluation metrics necessary to rank the forecasting techniques are presented in section

8. Section 9 goes on with the ETL pipeline and the exploratory analysis. Implementation details are included in 10. Finally section 11 presents the experiments, the results and discusses models' strenghts, weaknesses and possible improvements.

Chapter 2

Literature Review

Energy Forecasting

2.0.1 Energy forecasting classification

Types

In the context of energy forecasting, the quantities of most interest are price, load and renewables generation.

Forecasting horizons

Size

- Forecasting horizons

- Refer to section 8 for evaluating metrics

- plots come quelli delle review con i dati di Scopus se é una cosa fattibile

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.

Every paper uses different datasets (heterogeneous) So it is not possible to compare directly results from one paper to another without implementing the paper specific algorithms and the applying them to your dataset. This is

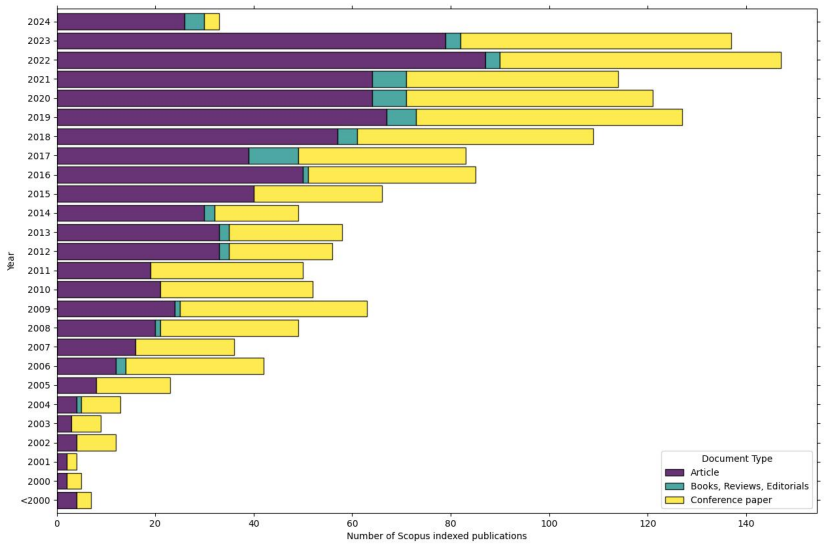


Figure 2.1: EPF publications [20]

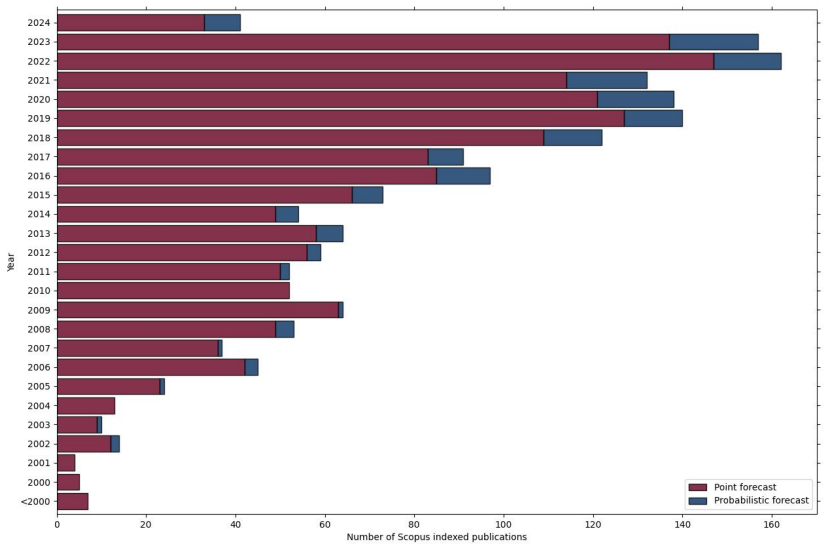


Figure 2.2: Point vs probabilistic pubblications

why sections after are destined to analyzing how these proposed methods so far work and their mathematical theory details

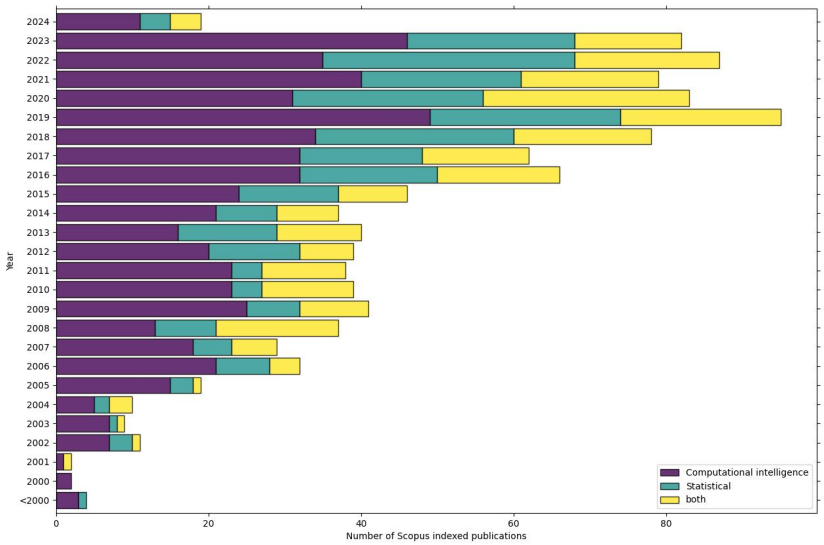


Figure 2.3: Pubblications by method

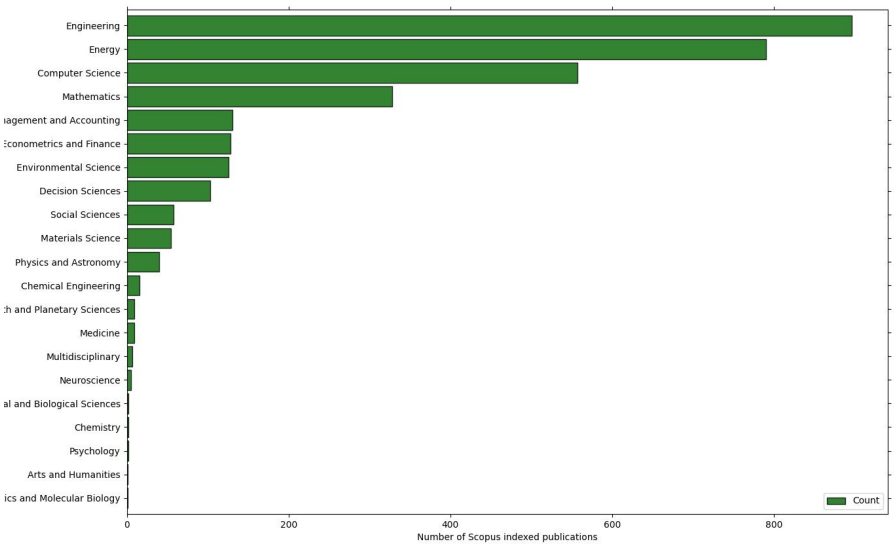


Figure 2.4: Pubblications by subject area

This problem has been already addressed in [2] and [13]. These papers surveyed the performance of neural network architectures against simpler

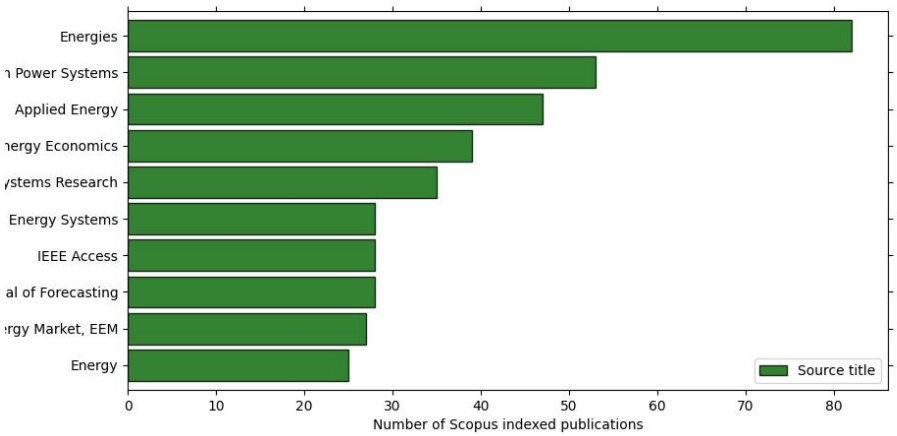


Figure 2.5: Most popolar sources/outlets

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 [11], [14] and kernel herding [7]. Particularly extending the idea of [6], 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.
****
During the past 25 years a wide range of new ideas have been proposed for point forecasting and for probabilistic forecasting.
The field benefitted greatly from the increse of computing power, the greater availability of dataset and the interest in data science. As a consequence, the forecaster's toolbox has grown in size and complexity.
Such variety of methods is characterized by heterogeneity in the fields from which they come from; methods come from statisitics, mathematics, econometrics, electrical engineering and the artificial intelligence communities.
Before delving into the literature review, it is important to make clear that at this point in time there is no superior method. Different solutions may outperform or underperform compared to other techniques depending on the problem settings. Thus, understanding the complexity, strenghts and weaknesses of each method is crucial for fitting the right model to the right setting.
Finally, within this research community, emerged the need of more ho-

mogeneity in the choice of the error valuation metrics and in the way of comparing model performances [20].
****
Lately, the idea of combining forecasts has gained popularity in the forecasting communit [16]; in the literature, combined forecasts are called ensemble [8]. Experimental results have shown ensemble methods to outperform their component forecasts.
Note that the more the errors of the combined models are not correlated the more we can benefit from ensembles.
It is also worth noting that older and simpler methods are still valuable(in combination with other models or on their own); these being less subject to overfitting than complex models.

Chapter 3

Kernel Theory

3.1 Kernel Mean Embedding of Distributions: A Review and Beyond

From this first paper [14], 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 $k(x, x') =: (x^T x')^2$.

3.1.1 RKHS

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

Definition 3.1 *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 3.2 *A complete metric space is a metric space in which every Cauchy sequence is convergent.*

Definition 3.3 *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 3.4 *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 3.5 *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 3.6 *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 $\mathcal{X}$

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

3.2 Recovering Distributions from Gaussian RKHS Embeddings

This paper covers the RKHS embedding approach to nonparametric statistical inference [11]. 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 3.7 *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 [11] 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 $\Phi : \mathbb{R}^n \to \mathcal{H}$.

3.3 Super-Samples from Kernel Herding

Kernel herding is a deterministic sampling algorithm designed to draw "Super Samples" from probability distributions [7]. 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.

- Explain kernel density estimation Explain under which conditions kernel mean embedding is equivalent to kernel density estimation. Kernel mean embedding generalization of kernel density estimation

- Other kernel theory concepts, I may need to restructure the structure of the kernel folder by putting in the right order the varies paper1,2,3,4

Chapter 4

Quantile Regression

- Explain the theory of quantile regression

Chapter 5

Kernel Density Estimation

- Explain the framework of kernel density estimation.

- Explain how it is applied in the literature.

- Extend to Conditional kernel density estimation and how it is applied in the literature

- Do a simple showcase with an example

Chapter 6

Ensemble Methods

- Explain the idea of ensemble methods.

- The most popular framework is based on autoregressive processes, so explain their Theory and how the procedure how they are used IMPORTANT: load series is not a stationary series so before applying AR we have to perform stationary tests or differencing steps

- simple example

Chapter 7

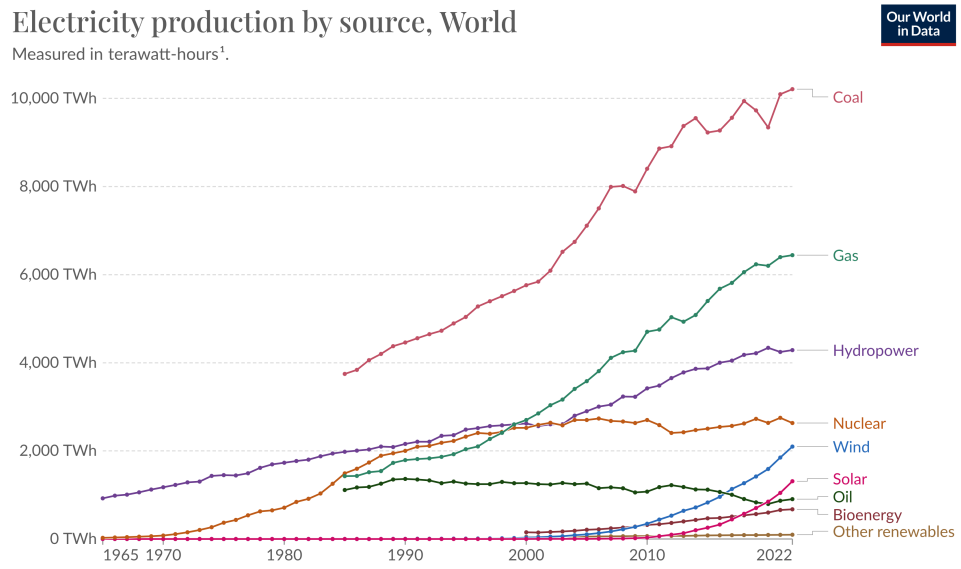
The Energy Market

Chapter explaining the energy market

- what are smart grids?

- introduction of smart grids

- renewable integration requirements



- How auctions work

- Difference intraday dayhead all other stuff
- EPEX Nordpool
- Prices can be negative

Chapter 8

Evaluation metrics

Proper evaluation methods guide researchers in choosing the model that best fits their needs; thus, this chapter is dedicated to the most common evaluation metrics adopted by academics in the field of EF.

8.1 MAPE

8.2 CRPS

Nevertheless, we can evaluate the integral in closed form. As pointed out in [9], we can use lemma 2.2 of [18] or identity 17 of [3].
The lemma states that ...
Note that in our case the distribution G of Y1 and Y2 is degenerate, with all probability mass on a single point (x notation will need to be cleaned here).
Since $G(t) =$:
It follows that the third addend in the summation is zero since the expectation of Y1 Y2 with distribution G means the difference of two equal constant numbers.
Additionally, since x is constant in the first term we have Y=x.

Chapter 9

Exploratory Analysis and Data ETL

- Explain how data has been retrieved.

- Data provider EPEX(entsoe retrieves its data)

- Explain the ETL(Extract Transform Load) pipeline I set up.

- Explore data in order to get useful insights for how to tune the models to get the most of them.

- correlation between temperatures and load

- Correlation and auto correlation plots

- Split train and test dataset Explain carefully why it is important to carry out an out of sample test and not in sample. In sample test involves look ahead bias because we are fitting the model on the data we want to predict, thus it overfits on the data considered but it does not generalize well.

Chapter 10

Implementation

Section documenting code

- indicate computer specifics

- see whether something can be parallelized

- Explain how quantile regression, random forest, gradient boosting have been used that is explain how their implementation has been adapted to my specific setting.

- Explain in detail how to my src code has been implemented its rationale and how to use it.

- As I explain code scripts go over the test, to explain better my ideas.

Chapter 11

Experiments Analysis

Analysis of experiments and results

- Comments
- Comparison
- Table of loss scores

Plots:

- Plots for visualizing timeseries with quantiles bounds
- Other plots that will come up to mind

List of Symbols

a	acceleration
CRPS	Continous ranked probability score
EDA	Exploratory data analysis
EF	Electric forecasting
EVs	Electric vehicles
ETL	Extract transform load
LCTs	Low carbon technologies
LV	Low voltage
PI	Prediction interval
PF	Price Forecasting
PPF	Probabilistic Price Forecasting
SDGs	Sustainable development goals
TSO	Transmission system operator

Appendix A

Appendix

A.1 Feature Map Normalization

Proof

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

□

A.2 Src code

The whole code for the project is hosted on https://github.com/luca-pernigo/ThesisKernelMethods.

- query: folder containing Scopus data and scripts to generate bibliometric survey plots in section 2.0.1

Bibliography

[1] THE 17 GOALS Sustainable Development - the United Nations. https://sdgs.un.org/goals.

[2] 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.

[3] Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.

[4] Ashutosh Bhagwat. Institutions and long term planning: Lessons from the california electricity crisis. *Administrative Law Review*, 55(1):95–125, 2003.

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

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

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

[8] Tilmann Gneiting and Adrian E Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[9] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

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

[11] 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.

[12] 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.

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

[14] 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.

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

[16] Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K Barrow, Souhaib Ben Taieb, Christoph Bergmeir, Ricardo J Bessa, Jakub Bijak, John E Boylan, et al. Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871, 2022.

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

[18] Gábor J Székely and Maria L Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

[19] Hong Tao. Crystal ball lessons in predictive analytics. *EnergyBiz*, pages 35–37, 2015.

[20] Rafał Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30, 10 2014.