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 [8]. Interestingly, the process of deregulation lead to an increasing interest in the field of electricity price forecasting (EF) within the academic community 2.2.
Moreover, the United Nations have identified the right to access affordable, reliable, sustainable and modern energy as one of their 17 SDGs [2]
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 [54]. 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.
Probabilistic forecasting may be useful to power producers, traders and consumers in order to improve their decision making process and managing risk(VaR). This hold in particular for traders, because probabilistic forecast enables them to simulate scenarios and carry out stress tests.
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

Electricity Forecasting 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.
As already pointed out, such variety of methods is characterized by the 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, it emerged the need of more homogeneity in the choice of the error valuation metrics (section 8), data quality and in the way of comparing model performances [61]; as a solution, [**?**] proposes a checklist to aid evaluating the meaningfulness of new research. Hence, throughout this thesis work, we will stick to the proposed principles and best practices peculiar of the EF field.

2.1 Electricity forecasting classification

Electricity forecasting is a vague term and it is used in the literature to refer to the whole field. Thus, in order to introduce some clarity it is useful to classify the range of EF articles in terms of their core attributes.

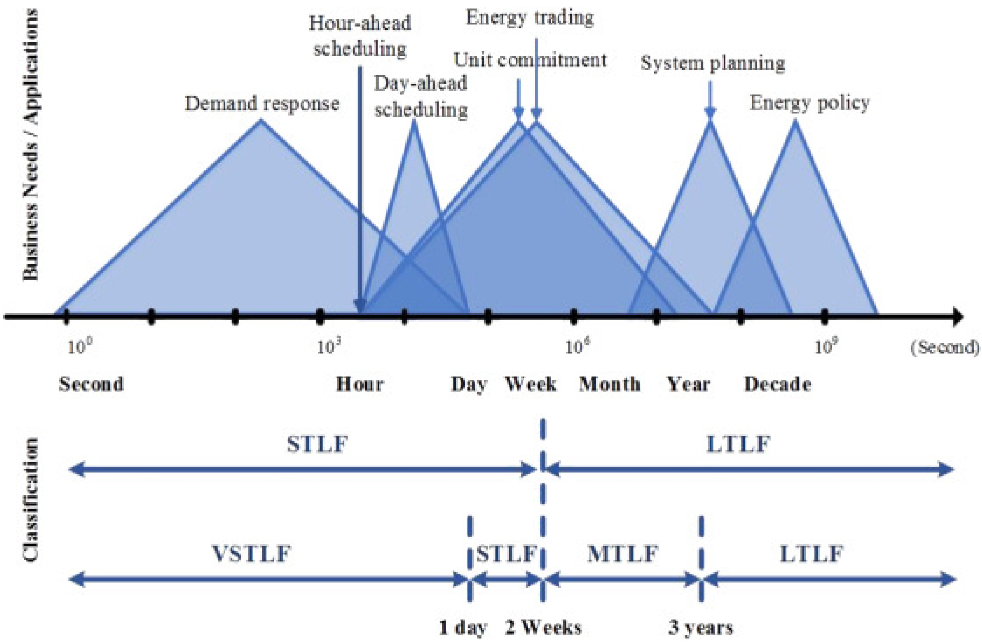


Figure 2.1: Time classsification [24]

2.1.1 Types

In the context of energy forecasting, the quantities of most interest are prices (EPF), loads (ELF) and renewables generation (mostly wind and solar).

2.1.2 Forecasting horizons

In terms of forecasting horizons, we can group EF into four major categories: very short term forcasting (VSTF), short term load forecasting (STF), medum term forecasting (MTF) and long term forecasting (LTF). Consensus in the literature is to use as cut off horizons one day, two weeks and three years respectilvely [23]; see 2.1 for a visualisation.

2.1.3 Size

Forecasts can either be for the whole target electricity network (system) or for a subset of it (zonal).

2.1.4 Point vs Probabilistic

Finally, EF literature distinguishes between point and probabilistic forecasts. Each of the two has its advantages and disadvantages; point forecasts are easier to generate and less computationally intensive while probabilistic forecast are more informative. Industry and research efforts have focused primarily on point forcasting. Nevertheless, interest in probabilistic forecasting

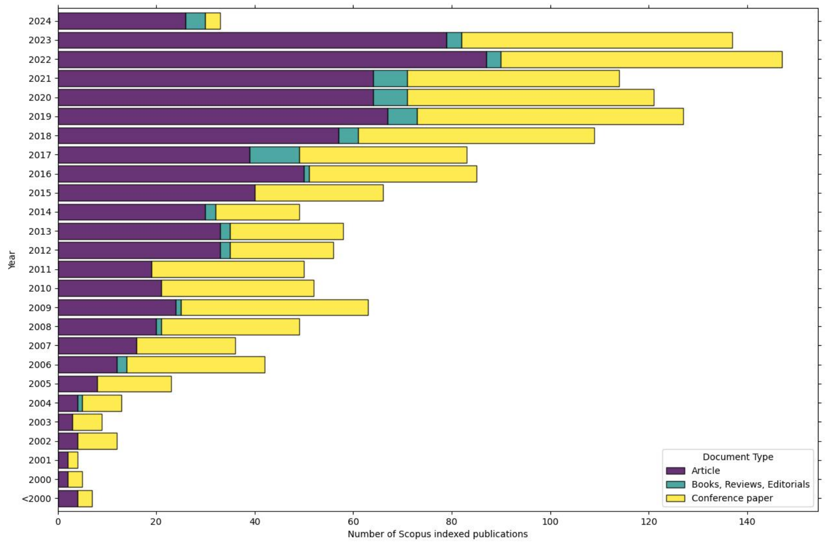


Figure 2.2: EF publications

has risen considerably over the last years due to renewable integration requirements, introduction of smart grids and increased market competitivness.

To conclude this section Ziel et Steinert carefully tables, are reported in the appendix for the interested reader.

2.2 Bibliographic analysis

This section presents the results of the bibliometric analysis we performed on 6th March 2024. This survey has been carried out by using the Scopus citation database. For details on the specific queries entered in Scopus, refer to A.2.
To get started, let us consider the evolution of the EF field over the years. This is visualized in figure 2.2, with results grouped by category; note, articles prior to the 2000 have been aggregated together due to their small number. Figure 2.2 shows the trend of an increasing interest in EF.
The next question was to compare the state of point versus probabilistic forecasting, this is visualized in figure 2.3. What it can be concluded is that probabilistic is less developed than point forecasting. To our mind this is due to the complexity of probabilistic forecast. Nevertheless, we can see a trend that suggests researchers are making an effort to fill this gap.
The EF literature is dominated by statistical and computational intelligence

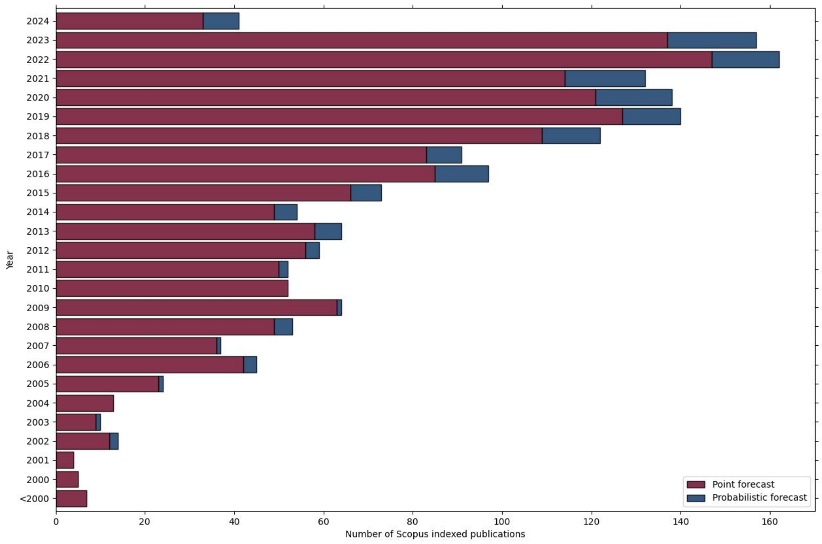


Figure 2.3: Point vs probabilistic pubblications

methods as can be seen from figure 2.4, with CI methods beign slightly preferred.
EF is an heterogenous field of research, its researchers come from a wide array of backgrounds, with electrical engineers and statisticians making up the top contributors; their different educational training may explain why the split between statistical and computational intelligence methods is so marked. Figure 2.5 depitcs the EF publications by subject area. What can be concluded, is that the bulk of pubblications come from engineering, computer science, mathematics and econometrics.
Finally, in order to refer to the most relevant source in the field, EF outlets have been ranked by popularity and plotted in 2.6.

2.3 EF literature review

To get started a few review articles were collected in order to understand conventions, best practicies and terminology of the EF community. Weron [61] reviews the state of the art for electricity price forecasting; beside analysing complexity of available solutions, strenghts and weaknesses it also stresses the need for objective comparative EPF studies. Specifically, it advocates for studies using similar datasets, using the same error evaluation metric and statistical testing model's outperformance. Hong et al. [24] discusses the state

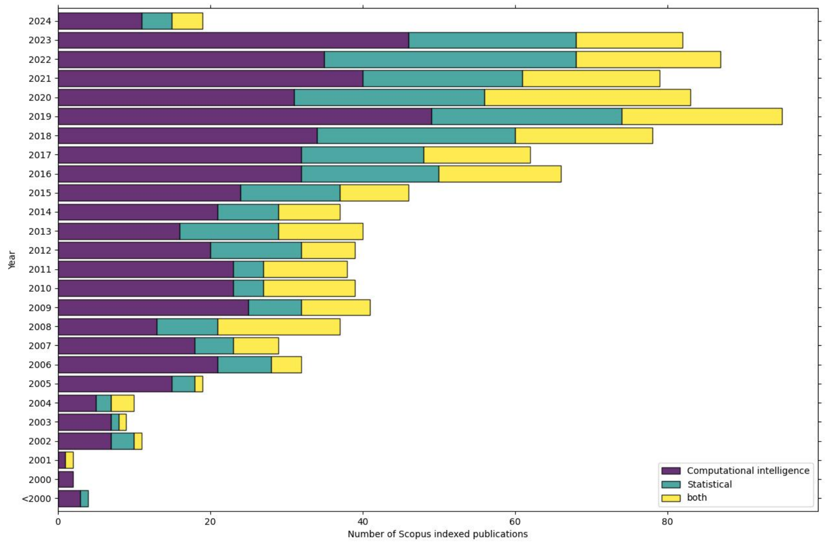


Figure 2.4: Pubblications by method

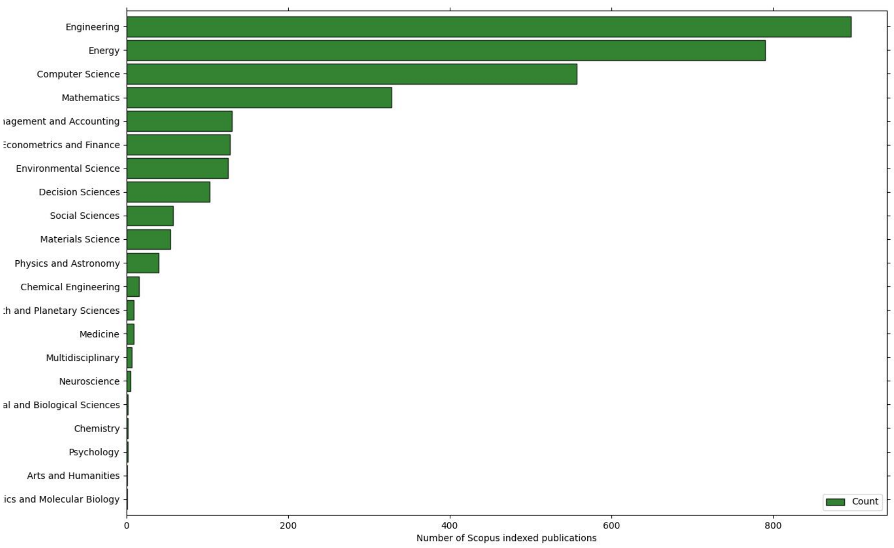


Figure 2.5: Pubblications by subject area

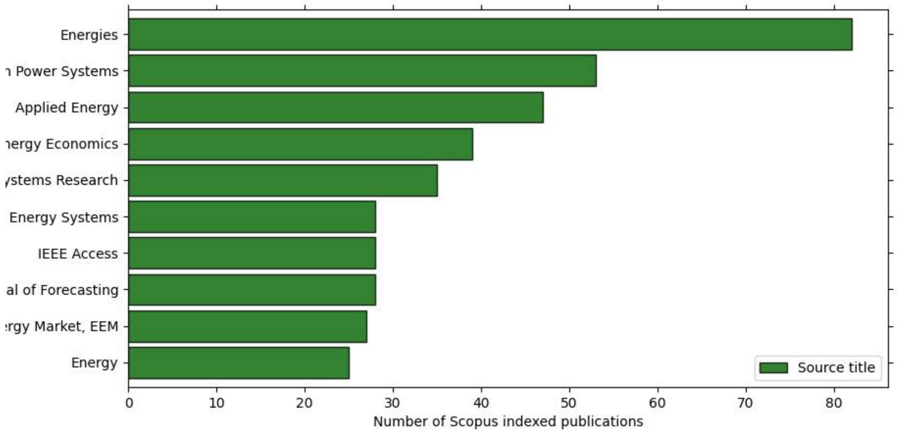


Figure 2.6: Most popolar sources/outlets

of the art in probabilistic electric load forecasting; it differentiates between techniques and methodologies. With techniques they refer to a family of models, like multiple linear regression or artificial neural networks. On the other hand, methodologies consist of general frameworks that can be incorporated into any method, for example variable selection mechanisms. Also this paper stresses the need for some guidelines to standardize research in the field. Nowotarski et al. [43] also carried out a thorough review of probabilistic forecasting. Weron et al. [35] offer a set of best practices when forecasting electricity prices in order to have a common framework to evaluate and compare future research. Zhang et al. [66] considers state of the art methods in wind power probabilistic forecasting and describes current challenges and possible future developments. Ziel et al. [68] provides detailed tables, grouping research papers by the time dimension 2.1.2 and objective 2.1.4 and reporting dataset, model and accuracy measures adopted. David et al. [12] adopt a combination of ARMA and GARCH in probabilistic forecasts of solar irradiance. Furthermore, they propose a recursive framework for parameter estimation. De Gooijer [13] reviews 25 years of time series forecasting for the period 1990-2005, highlighting the most influential works. He et al. [21] models multistep wind speed probabilistic forecasting by mixing CEEDMAN, LASSO and QRNN. In this work, CEEMDAN is used to decompose the wind speed time series; LASSO compresses high dimensional features; QR is used for obtaining quantile forecasts; finally KDE converts quantile forecasts into density estimates. Wan et al. [60] proposes a combination of QR and ELM to generate non parametric probabilistic forecast of wind generation. Van der Meer et al. [56] provides another thorough anlysis of the probabilistic forecasting realm by covering recent advances and identifying research gaps. The IEEE Power and Energy Society provides also insightful lecture notes

on probabilistic energy forecasting methodologies, implementations, and applications [1]. Marcjasz et al [37] uses distributional neural networks to crete probabilistic forecasts for the day ahead electricity prices in the german market. Nowotarski et al. [42] introduce a method for constructing PI and call it quantile regression averaging. Their idea is weighting a set of models' prediction such that the pinball loss of the weighted model is minimized. The observed results is that QRA performed better compared to twelve individual models. Arora et al. [6] focus on modelling electricity smart meter data by proposing a non parametric probabilistic technique based on kernel density estimation and conditional kernel density estimation [47] [29]. Their conclusion is that kernel density methods are competitive against exponential smoothing when forecasting residential data. Conversely, exponential smoothing has still an edge in predicting SME data. Zhang et al. [65] introduce a framework based on quantile regression and kernel density estimation in the context of short term wind forecasting. The proposed methods behave well compared to a ARMA(1,1). Haben et al. [19] analyses a variety of techniques in terms of both probabilistic and point forecasting, within this study they focus on load forecasting at the low voltage level. Zhang et al [64] proposes a two stage bootstrap sampling framework for probabilistic load forecasting. They test it for different regression models such as RF, GBRT, linear regression, and LSSVM regression. Jónsson et al. [31] introduces a density model for the day ahead market extending the adaptive QR framework of [39] by modelling the tails of the predicted density with an exponential distribution. Huurman et al. [28] surveys the predictive power of weather variables for electricity prices in danish market. Their empirical results suggest that weather is central for point forecasting day ahead prices. The opposite conclusion are drawn for density forecasting A major step forward in EF was the creation of the global energy forecasting competition (GEFCom) in 2012. Until then, no formal benchmarking process or data pool was established and new publications rarely reproduced the results from work done by others. Addressing these issues was the motivation behind the creation of GEFCom by the IEEE working group on energy forecasting. The EF field was positively affected by this competition; a number of ideas were tested on the same setting with only the best ones being published and it also contributed bridging the gap between industry practice and academic research. GEFCom 2012 had two tracks; the former about hierarchical load forecasting, the latter about wind power forecasting, see [25] for a comprehensive review.
The focus of GEFCom 2014 was on probabilistic forecasting, Hong et al. [26] discusses the problem tracks, the data and the winning methods. In this paragraph some of the winning entries of the 2014 GEFCom edition are discussed. Xie et al. [63] propose a two stage approach; in the first stage they use MLR to build a point forecast, then in the second stage they try different approaches for modelling the MLR residuals, among other they tried ESM,

ANN and ARIMA. Maciejowska et al. [36] proposes a new probabilistic model extending the idea of QRA [42]. Haben et al. [18] mixes CKD and QR in their competition entry. Gaillard et al. [14] [15] combines quantile regression with generalized additive models [20]. Ziel et al. [67] estimates an AR model through the LASSO [55] instead of the standard OLS.
The last GEFCom was held in 2017, its focus was providing probabilistic load forecasts, see [27]. The GEFCom competition has also inspired the organization of other competitions such as the RWEnpower competition in the UK, the RTE competition in France, the Tokyo electric power company competition in Japan and the BigDEAL forecasting competitions.

Considerations that can be drawn from the above literature review. 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 why sections after are destined to analyzing how these proposed methods so far work and their mathematical theory details
****
****
Lately, the idea of combining forecasts has gained popularity in the forecasting communit [45]; in the literature, combined forecasts are called ensemble [16]. 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.

2.4 Kernel methods 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. Throughout this thesis work, we will address the performance of kernel methods in the context of EF.
- Kernel history evolution. Their name comes from the german word kern, which translates to core in english. Such term was first used by David Hilbert in his paper on integral equations [22] where he defined a definite kernel. Following, Hilbert's and Schmidt's [48] work lead to the introduction of a new space, the Hilbert space. In 1909, James Mercer improved Hilbert's work by proposing his theorem [38]. This theorem underlies the power of

kernel methods, that is the kernel trick. In 1938, Schoenberg [49] developed the mathematical results that allow us to find the kernel associated to a specific feature space metric. In 1941, Kolmogorov [34] carried out stuides on representing kernel in linear spaces. In 1950, Aronszajn [5] published the first work on RKHS; developing the general method for representing kernels in linear spaces. In 1964, Aizerman [4] further improved the theory of RKHS. It was in the nineties that theory of kernel methods got popular, particularly in the field of machine learning. Kernels have been used in various different tasks such as SVM [58] [57], Gaussian process classifiers [62], spline methods [59], neural networks [46] and principal component analysis [50]. Nevertheless kernel methods received very little attention in the specific setting of EF literature.

The theory needed for this thesis work is covered in section 3. For an introduction to kernel methods, we referred to [52]. Kanagawa et Fukumizu introduces to the concept of kernel mean embedding [32]. Muandet et al. [40] surveys established results and new advances in the theory of Hilbert space distribution embeddings. It has to be said that computing and storing such embeddings becomes prohibitive for large scale settings. Rudi et al. [10] proposes an efficient approximation procedure based on the Nyström method [44], providing also an upper bound for the approximation error. Article [11] presents kernel herding; basically, Smola et al. used the kernel trick to extend the herding algorithm to continous spaces. The result is an infinite memory deterministic process that takes in a collection of samples and learns to approximate a pdf.

Chapter 3

Kernel Theory

3.1 Kernel Mean Embedding of Distributions: A Review and Beyond

From this first paper [40], 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 [32]. 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 [32] 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 [11]. 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:

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

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

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

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.

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

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

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

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

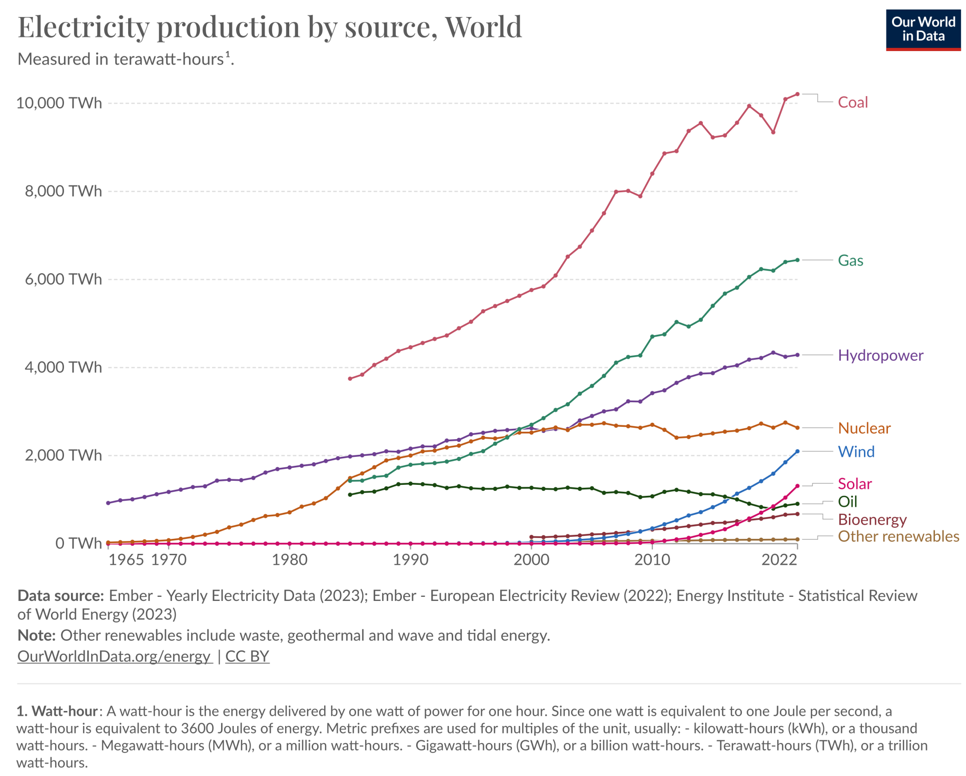
- simple example

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 [17], we can use lemma 2.2 of [53] or identity 17 of [7].
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

ANN	Artificial neural network
AR	Autoregressive model
ARMA	Autoregressive moving average model
ARIMA	Autoregressive integrated moving average model
ARX	Autoregressive exogenous model
CI	Computational intelligence
CKD	Conditional kernel density
CRPS	Continous ranked probability score
DDNN	Distributional neural network
EDA	Exploratory data analysis
EF	Electricity forecasting
ELF	Electricity load forecasting
EMD	Empirical mode decomposition
EPF	Electricity load forecasting
ESM	Exponential smoothing models
ETL	Extract transform load
EVs	Electric vehicles
GBRT	Gradient boosting regression tree
GEFCom	Global energy forecasting competition
HWT	Holt-Winters-Taylor exponential smoothing method
KDE	Kernel density estimation
LCTs	Low carbon technologies
LSSVR	Least squares support vector regression
LV	Low voltage
MLR	Multiple linear regression
OLS	Ordinary least squares
RF	Random forest

RKHS	Reproducing kernel hilbert space
PI	Prediction interval
PF	Price Forecasting
PPF	Probabilistic Price Forecasting
QR	Quantile regression
QRA	Quantile regression averaging
SME	Small and medium-sized enterprises
SNARX	Smoothed nonparametric ARX
SDGs	Sustainable development goals
SVM	Support vector machine
TOU	Time of use tariffs
TSO	Transmission system operator

Appendix A

Appendix

A.1 Feature Map Normalization

Proof

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

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

Bibliography

[1] GM22TutorialProbabilisticEnergyForecasting. https://resourcecenter.ieee-pes.org/education/tutorials/pes_ed_tut_gm22_0717_pefm_sld.

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

[3] 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.

[4] A Aizerman. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[5] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[6] Siddharth Arora and James W Taylor. Forecasting electricity smart meter data using conditional kernel density estimation. *Omega*, 59:47–59, 2016.

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

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

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

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

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

[12] Mathieu David, F Ramahatana, Pierre-Julien Trombe, and Philippe Lauret. Probabilistic forecasting of the solar irradiance with recursive arma and garch models. *Solar Energy*, 133:55–72, 2016.

[13] Jan G De Gooijer and Rob J Hyndman. 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473, 2006.

[14] Pierre Gaillard, Yannig Goude, and Raphaël Nedellec. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050, 2016.

[15] Pierre Gaillarda, Yannig Goudea, and Raphaël Nedelleca. Semi-parametric models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting.

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

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

[18] Stephen Haben and Georgios Giasemidis. A hybrid model of kernel density estimation and quantile regression for gefcom2014 probabilistic load forecasting. *International Journal of Forecasting*, 32(3):1017–1022, 2016.

[19] Stephen Haben, Georgios Giasemidis, Florian Ziel, and Siddharth Arora. Short term load forecasts of low voltage demand and the effects of weather. *arXiv preprint arXiv:1804.02955*, 2018.

[20] Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

[21] Yaoyao He, Yun Wang, Shuo Wang, and Xin Yao. A cooperative ensemble method for multistep wind speed probabilistic forecasting. *Chaos, Solitons & Fractals*, 162:112416, 2022.

[22] David Hilbert. Grundzuge einer allgemeinen theorie der linearen integralgleichungen. *Gott. Nachr.*, pages 307–338, 1904.

[23] Tao Hong. *Short term electric load forecasting*. North Carolina State University, 2010.

[24] Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[25] Tao Hong, Pierre Pinson, and Shu Fan. Global energy forecasting competition 2012, 2014.

[26] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

[27] Tao Hong, Jingrui Xie, and Jonathan Black. Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1389–1399, 2019.

[28] Christian Huurman, Francesco Ravazzolo, and Chen Zhou. The power of weather. *Computational Statistics & Data Analysis*, 56(11):3793–3807, 2012.

[29] Rob J Hyndman, David M Bashtannyk, and Gary K Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996.

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

[31] Tryggvi Jónsson, Pierre Pinson, Henrik Madsen, and Henrik Aalborg Nielsen. Predictive densities for day-ahead electricity prices using time-adaptive quantile regression. *Energies*, 7(9):5523–5547, 2014.

[32] 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.

[33] 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.

[34] Andrei Nikolaevitch Kolmogorov. Stationary sequences in hilbert space. *Bull. Math. Univ. Moscow*, 2(6):1–40, 1941.

[35] Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

[36] Katarzyna Maciejowska and Jakub Nowotarski. A hybrid model for gefcom2014 probabilistic electricity price forecasting. *International Journal of Forecasting*, 32(3):1051–1056, 2016.

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

[38] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

[39] Jan Kloppenborg Møller, Henrik Aalborg Nielsen, and Henrik Madsen. Time-adaptive quantile regression. *Computational Statistics & Data Analysis*, 52(3):1292–1303, 2008.

[40] 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.

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

[42] Jakub Nowotarski and Rafał Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30:791–803, 2015.

[43] Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018.

[44] Evert J Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. 1930.

[45] 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.

[46] Tomaso Poggio and Federico Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[47] Murray Rosenblatt. Conditional probability density and regression estimators. *Multivariate analysis II*, 25:31, 1969.

[48] Erhard Ing Grad Schmidt. Über die auflösung linearer gleichungen mit unendlich vielen unbekannten. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 25:53–77, 1908.

[49] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[50] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

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

[52] Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

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

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

[55] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[56] Dennis W Van der Meer, Joakim Widén, and Joakim Munkhammar. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, 81:1484–1512, 2018.

[57] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1995.

[58] Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

[59] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[60] Can Wan, Jin Lin, Jianhui Wang, Yonghua Song, and Zhao Yang Dong. Direct quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Transactions on Power Systems*, 32(4):2767–2778, 2016.

[61] 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.

[62] Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1342–1351, 1998.

[63] Jingrui Xie and Tao Hong. Gefcom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, 32(3):1012–1016, 2016.

[64] Jiawei Zhang, Yi Wang, Mingyang Sun, and Ning Zhang. Two-stage bootstrap sampling for probabilistic load forecasting. *IEEE Transactions on Engineering Management*, 69(3):720–728, 2020.

[65] Lei Zhang, Lun Xie, Qinkai Han, Zhiliang Wang, and Chen Huang. Probability density forecasting of wind speed based on quantile regression and kernel density estimation. *Energies*, 13(22):6125, 2020.

[66] Yao Zhang, Jianxue Wang, and Xifan Wang. Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, 32:255–270, 2014.

[67] Florian Ziel and Bidong Liu. Lasso estimation for gefcom2014 probabilistic electric load forecasting. *International Journal of Forecasting*, 32(3):1029–1037, 2016.

[68] Florian Ziel and Rick Steinert. Probabilistic mid-and long-term electricity price forecasting. *Renewable and Sustainable Energy Reviews*, 94:251–266, 2018.