Singular value decomposition for time series analysis with applications to smart energy systems

Khoshrou, A.

DOI
10.4233/uuid:9bf59202-4b7a-4313-b972-c12b7d272c06

Publication date
2022

Citation (APA)
Khoshrou, A. (2022). *Singular value decomposition for time series analysis with applications to smart energy systems*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:9bf59202-4b7a-4313-b972-c12b7d272c06

Singular Value Decomposition for Time Series Analysis with Applications to Smart Energy Systems

SINGULAR VALUE DECOMPOSITION FOR TIME SERIES ANALYSIS WITH APPLICATIONS TO SMART ENERGY SYSTEMS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 19th December 2022 at 15:00 o'clock

by

Abdolrahman KHOSHROU

Master of Science in Information Engineering,
University of Porto, Portugal
born in Ghaemshahr, Iran

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. J.A. La Poutré	Delft University of Technology and Centrum Wiskunde & Informatica, promotor
Dr. E.J.E.M. Pauwels	Centrum Wiskunde & Informatica, copromotor

Independent members:

Dr. J. Kazempour	Technical University of Denmark, Denmark
Prof. dr. A.E. Çetin	University of Illinois Chicago, USA
Prof. dr. A.A. Salah	Utrecht University, The Netherlands
Prof. dr. W.G.J.H.M. van Sark	Utrecht University, The Netherlands
Prof. dr. ir. C. Vuik	Delft University of Technology
Prof. dr. P. Palensky	Delft University of Technology, reserve member

ISBN 978-94-6384-396-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/

SUMMARY

In a world replete with observations (physical as well as virtual), many data sets are represented by time series. In its simplest form, a time series is a set of data collected sequentially, *usually* at fixed intervals of time. In a number of applications, the mean and the variance of the time series is time-invariant and there is no seasonality in the data (such time series is called stationary). However, in many more applications, e.g., time series that are related to smart energy systems, the data have non-stationary characteristics.

This thesis focuses primarily on matrices as an alternative representation of the latter type of time series, in order to take advantage of matrix decomposition methods. The rationale is straightforward: numerically stable matrix decomposition techniques enable us to extract underlying patterns in the data and use them to construct approximations of the corresponding time series. In particular, we will focus on singular value decomposition (SVD) as a powerful and numerically stable matrix factorization technique. Therefore, as the first step in this thesis, the SVD and its geometrical interpretation are extensively studied, in order to acquire a firm understanding of how it performs. That in turn enables us to look at different problems in time series analysis from a fresh perspective.

For most of the applications of SVD in various fields, it is important to understand the properties of the SVD of a matrix whose entries show some degree of random fluctuations. Therefore, to determine how the noise level affects the singular value spectrum, it is essential to study the singular value decomposition of random matrices. As we will explain in the introductory chapter, one of the early applications of the SVD in time series analysis is in periodicity detection of the time series data. Therefore, we explore how the geometry of a matrix (the position of the data points with respect to the origin) and the aspect ratio of the matrix (the ratio between the number of columns and the number of rows) can affect its SVD results.

Matrix factorisation techniques such as principal component analysis (PCA) and singular value decomposition (SVD) are both conceptually simple and effective. However, it is well-known that they are sensitive to the presence of noise and outliers in input data. One way to mitigate this sensitivity is to introduce *regularisation*. To this aim, we hark back to the interpretation of SVD and PCA in terms of low-rank approximations, which involve the minimisation of specific functionals. We then derive algorithms for the minimisation of the regularised version of such functionals.

After the above-described theoretical investigations of SVD, we considered novel applications of SVD to various problems. The first one concerned challenges related to the integration of renewable energy sources (RES). With increasing RES-integration such as wind and solar energy to the power grid, balancing the grid has become more challenging. This is mostly due to the inherently intermittent nature of RES, on the one hand, and shortcomings in bulk energy storage systems, on the other. Therefore, studies on *scenario-based* probabilistic energy production and demand forecasts have gained mo-

mentum, as they are highly valuable from both a technical and an economic point of view. A particular application of such models in the energy sector is where having the distribution of energy consumption for the coming days is desired. Furthermore, as extensively argued in the literature, a decisive variable in predicting energy demand is temperature data. There are mainly three practical and popular methods for generating temperature scenarios, namely fixed-date, shifted-date, and bootstrap approaches. Nevertheless, these methods have mostly been used on an ad-hoc basis without being formally compared or quantitatively evaluated. Moreover, as we discuss, the performance of such models depends to a large extent to the quality of input data. Therefore, we propose a generic, data-driven and computationally efficient SVD-based approach to simulate temperature scenarios. The strength of our proposed method lies in its simplicity and robustness, in terms of the training window size, with no need for subsetting or thresholding in order to generate temperature scenarios. The empirical case studies performed on the data from the load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L) show that the proposed method outperforms the top two scenario-based models with a similar set-up.

Another topic of considerable interest is to investigate what effect the transition of energy to RES can have on the overall trend and volatility of electricity prices. This impact could be complex because there are two contradicting forces at play. The marginal cost of RES is relatively low and even negative (especially if subsidized), therefore, increased penetration of wind and solar would result in a downward trend in electricity prices. Opposing this is the associated uncertainty regarding the availability of wind and solar energy, which causes spikes in the market. In other words, the integration of RES provokes assertions that the stability of the power grid can (surely) be compromised due to the inherent intermittency of such sources. Therefore, the increased price volatility will cause additional market risks for suppliers and consumers in the market.

In the literature, numerous methods have been introduced to determine the volatility of the time series data. However, as we exemplify, the emergence of non-positive price values in the energy transition era has introduced new challenges in the electricity market volatility analysis. This new aspect of the market renders many traditional volatility indices ineffective. More precisely, the standard approach to switch to logarithmic measures can be done only after shifting up all values above zero by a certain threshold. On the other hand, price volatility has a dependence on the price level, which is even more pronounced when the spot prices are low. Therefore, the generalizability of conventional approaches is questioned, as the volatility measures can vary drastically, with respect to the magnitude of the aforementioned thresholds. We tackle this problem by introducing a new notion of volatility which is obtained by reconstructing the time series using the SVD technique. In other words, we detect and remove the *deterministic* part of the price data using the SVD and consider the *stochastic* part (residuals) as a notion of volatility. Using the matrix representation of the data, we then highlight the evidence of the effect of renewables on daily price profiles in the German day-ahead market, i.e., the emergence of non-positive prices and also shifts of peak price values to hours where solar is less available.

Overall, in this thesis, we study in-depth the SVD technique and propose novel applications of it in time series analysis. Our findings can be used as innovative components

of future smart grid systems, which are characterized by the increasing uncertainty on both the supply and demand parts.

SAMENVATTING

In een wereld vol fysieke en virtuele waarnemingen zijn veel datasets opgebouwd in de vorm van tijdreeksen. Een tijdreeks, in zijn eenvoudigste vorm, is een reeks gegevens die opeenvolgend worden verzameld, meestal met vaste tussenpozen. In sommige toepassingen zijn het gemiddelde en de variantie van de reeks tijdsinvariant en vertonen de gegevens geen seizoenseffecten; een dergelijke tijdreeks wordt stationair genoemd. Voor de meeste toepassingen, bijvoorbeeld tijdreeksen die betrekking hebben op slimme energiesystemen, vertonen de gegevens echter niet-stationaire kenmerken.

Dit proefschrift richt zich voornamelijk op alternatieve representaties van de laatste soort tijdreeksen in de vorm van matrices, zodat geprofiteerd kan worden van matrixontledingsmethoden. De reden is eenvoudig: numeriek stabiele matrixontledingstechnieken stellen ons in staat om onderliggende patronen in de gegevens te extraheren, en hiermee benaderingen van de bijbehorende tijdreeks te construeren. In het bijzonder zullen we ons richten op singuliere waardenontbinding (SWO); een krachtige en numeriek stabiele matrixfactorisatie techniek. De eerste stap in dit proefschrift is daarom het uitgebreid bestuderen van de SWO en zijn geometrische interpretatie, om zo een goed begrip van de prestaties te verkrijgen. Dit begrip stelt ons in staat om verschillende problemen in tijdreeksanalyses vanuit een nieuw perspectief te bekijken.

Voor de meeste toepassingen van SWO is het cruciaal om inzicht te verkrijgen in de eigenschappen van de SWO van een matrix, waarvan de gegevens enige mate van willekeurige fluctuaties vertonen. Om te bepalen hoe ruis het singuliere waardenspectrum beïnvloedt, is het essentieel om de ontleding van singuliere waarden van willekeurige matrices te bestuderen. Zoals we toelichten in het inleidende hoofdstuk, is periodiciteitsdetectie van tijdreeksgegevens één van de eerste toepassingen van de SWO in tijdreeksanalyse. Om deze reden onderzoeken we hoe de geometrie van een matrix (de positie van de datapunten ten opzichte van de oorsprong) en de aspectverhouding van de matrix (de verhouding tussen het aantal kolommen en het aantal rijen) de SWO-resultaten kunnen beïnvloeden.

Matrixfactorisatie technieken zoals principale-componentenanalyse (PCA) en singuliere waardeontbinding (SWO) zijn zowel effectief als conceptueel eenvoudig. Het is echter algemeen bekend dat ze gevoelig zijn voor de aanwezigheid van ruis en uitschieters in invoergegevens. Het toepassen van regularisatie is een manier om deze gevoeligheid te verminderen. Om dit doel te bereiken, grijpen we terug naar interpretaties van SWO en PCA in termen van lage-rangapproximaties, die betrekking hebben op het minimaliseren van specifieke functionalen. Vervolgens leiden we algoritmen af voor de minimalisatie van de geregulariseerde versie van dergelijke functionalen.

Na het hierboven beschreven theoretische onderzoek van SWO hebben we nieuwe toepassingen van SWO voor verschillende problemen bestudeerd. De eerste heeft betrekking op uitdagingen in verband met de integratie van hernieuwbare energiebronnen. Met toenemende integraties van bronnen zoals wind- en zonne-energie binnen

het elektriciteitsnet, is het balanceren van het net een grotere uitdaging geworden. Dit is voornamelijk te wijten aan het inherent intermitterende karakter van hernieuwbare energiebronnen enerzijds, en tekortkomingen in bulkopslag van energie anderzijds. Om die reden hebben studies naar op scenario's gebaseerde probabilistische prognoses voor energieproductie en -vraag aan momentum gewonnen, aangezien ze zeer waardevol zijn vanuit zowel technisch als economisch oogpunt. Een specifieke toepassing van dergelijke modellen in de energiesector is het bepalen van de gewenste verdeling van energieverbruik voor de komende dagen. Zoals uitvoerig besproken in de literatuur, zijn temperatuurgegevens een beslissende variabele bij het voorspellen van de energievraag. Er zijn drie praktische en populaire methoden voor het genereren van temperatuurscenario's, namelijk benaderingen met vaste datum, verschoven datum en bootstrapping. Deze methoden worden echter meestal op ad-hocbasis gebruikt, zonder formele vergelijking of kwantitatieve beoordeling. Bovendien zijn de prestaties van dergelijke modellen in grote mate afhankelijk van de kwaliteit van de invoergegevens. Om die reden stellen we simulatie van temperatuurscenario's voor, als een generieke, datagedreven en rekenkundig efficiënte op SWO gebaseerde benadering. De kracht van onze voorgestelde methode ligt in zijn eenvoud en robuustheid, in termen van de grootte van het trainingsvenster, zonder noodzaak tot het creëren van deelverzamelingen of drempelwaarden om temperatuurscenario's te genereren. Uit de empirische casusstudies, uitgevoerd op de gegevens van de load forecasting-sessie van de Global Energy Forecasting-competitie 2014 (GEFCom2014-L), blijkt dat de voorgestelde methode beter presteert dan de twee beste scenario-gebaseerde modellen met een vergelijkbare opzet.

Een ander belangrijk onderzoeksonderwerp is het effect dat de overgang naar hernieuwbare energiebronnen kan hebben op de algemene trend en volatiliteit van de elektriciteitsprijzen. Deze impact kan complex zijn, omdat er twee tegenstrijdige krachten spelen. De marginale kosten van hernieuwbare energiebronnen zijn relatief laag of zelfs negatief (vooral indien gesubsidieerd), daarom zou verhoogd gebruik van wind- en zonne-energie resulteren in een neerwaartse trend in elektriciteitsprijzen. Daartegenover staat de onzekerheid omtrent de beschikbaarheid van wind- en zonne-energie, die pieken in de marktprijs veroorzaakt. Met andere woorden, door de inherente fluctuaties van hernieuwbare energiebronnen kan de integratie van dergelijke bronnen de stabiliteit van het elektriciteitsnet in het gedrang brengen. De verhoogde prijsvolatiliteit zal daardoor leiden tot extra marktrisico's voor leveranciers en consumenten.

In de literatuur zijn talloze methoden geïntroduceerd om de volatiliteit van tijdreeksgegevens te bepalen. We illustreren echter dat de opkomst van niet-positieve prijzen in het tijdperk van energietransitie heeft geleid tot nieuwe uitdagingen in de volatiliteitsanalyse van de elektriciteitsmarkt. Dit nieuwe aspect van de markt zorgt ervoor dat veel traditionele volatiliteitindices niet effectief zijn. Preciezer geformuleerd: de standaardaanpak om over te schakelen naar logaritmische maten kan alleen worden uitgevoerd nadat alle waarden boven nul met een bepaalde drempel zijn opgeschoven. Aan de andere kant is prijsvolatiliteit afhankelijk van het prijsniveau, een effect dat zelfs meer uitgesproken is wanneer de spotprijzen laag zijn. Daarom wordt de generaliseerbaarheid van conventionele benaderingen in twijfel getrokken, aangezien de volatiliteitsmaatregelen drastisch kunnen variëren met betrekking tot de hoogte van de bovengenoemde drempels. We pakken dit probleem aan door een nieuwe notie van volatiliteit

te introduceren, die wordt verkregen door de tijdreeks te reconstrueren met behulp van de SWO-techniek. Met andere woorden: we detecteren en verwijderen het deterministische deel van de prijsgegevens met behulp van de SWO en beschouwen het stochastische deel (residuen) als een notie van volatiliteit. Met behulp van de matrixweergave van de gegevens belichten we vervolgens het bewijs van het effect van hernieuwbare energiebronnen op dagprijsprofielen in de Duitse day-aheadmarkt, d.w.z. de opkomst van niet-positieve prijzen en ook verschuivingen van piekprijswaarden naar uren waarop zonne-energie is minder beschikbaar.

Samenvattend bestuderen we in dit proefschrift op diepgaande wijze de SWO-techniek en stellen we nieuwe toepassingen voor op het gebied van tijdreeksanalyse. Onze bevindingen kunnen worden gebruikt als innovatieve componenten van toekomstige slimme elektriciteitsnetwerken, die worden gekenmerkt door toenemende onzekerheid omtrent zowel het vraag- als het aanbodgedeelte.

ACKNOWLEDGEMENTS

This thesis concludes a big chapter of my life. In the last couple of years, I have grown both as a humble researcher and a person. I have learned to always keep an open mind and embrace all the tough, challenging, rewarding and fulfilling moments that a PhD can offer. I also have honed my hands-on mentality and drive to get things done.

I am indebted to many excellent collaborators and colleagues for their invaluable time, feedback, and suggestions.

Foremost, I express my greatest gratitude to my promotor Prof. dr. ir. J.A. La Poutré, and my co-promotor, Dr E.J.E.M. Pauwels for their constant help and guidance during my PhD journey. Throughout the writing of this thesis, I have always received great feedback and comments from them.

Next, I would like to thank all the members of the Intelligent and Autonomous Systems (IAS) group at CWI for making it fun and exciting to come to work every day. My special appreciation goes to Dr Brinn Hekkelman, Roland Saur and Xinyu Hu for their friendship. I have also always benefited from extensive interactions and conversations with Dr Tim Baarslag, Dr Hoang Luong and Dr Swasti R. Khuntia. I am also grateful to my dear friend, Dr Wouter van Heeswijk for translating this thesis summary into Dutch.

And, last but not least, I would like to thank my family for their love and support. You always have encouraged me to chase my dreams without fear. I am so grateful to have you all in my life.

CONTENTS

Abdolrahman KHOSHROU

Dead yesterdays and unborn tomorrows,
why fret about it, if today be sweet.

Omar Khayyam

1

INTRODUCTION

1.1. CONTEXT AND MOTIVATION

In a world replete with observations (physical as well as virtual), many data sets come in the shape of time series. In their simplest incarnation, time series represent an ordered sequence of values of a variable at equally spaced time intervals. If the variable of interest is basically stationary (e.g., the output of a stable production process), the global characteristics of the time series do not change much. In such cases, one can describe the time series processes in terms of a random variable with constant statistical moments.

However, time series that are related to human activities — such as data streams produced by **smart infrastructures** — often have a non-stationary structure. For instance, whereas the electrical consumption of households will be similar throughout the week, it will be markedly different from consumption on the weekend. Similarly, significant gradual shifts will be noticeable over the course of a year.

This thesis focuses primarily on an **alternative representation of the time series** data that offers some advantages when it comes to the analysis of these types of data. Specifically, we will focus on representing time series as matrices in order to take advantage of **matrix decomposition methods.** The rationale is straightforward: since matrix decomposition provides principled methods to find **low-rank approximations** of a matrix, they will also give rise to an **approximation of the corresponding time series**. However, by their very nature, these methodologies are conceptually different from the standard statistical techniques for averaging or summarizing time series.

In particular, in this thesis, we will focus on the **singular value decomposition (SVD)** as a powerful, numerically stable matrix factorization technique which is then applied to time series analysis. Therefore, the SVD is herein extensively studied to acquire a firm understanding of how it performs. That in turn enables us to look at different applications in time series analysis from a fresh perspective.

For now, it suffices to announce the gist of SVD, but we will provide more details in Chapter 2. Given any matrix A, SVD allows us to represent it as the product of three matrices (also see Figure 1.1) :

$$A = USV^T \tag{1.1}$$

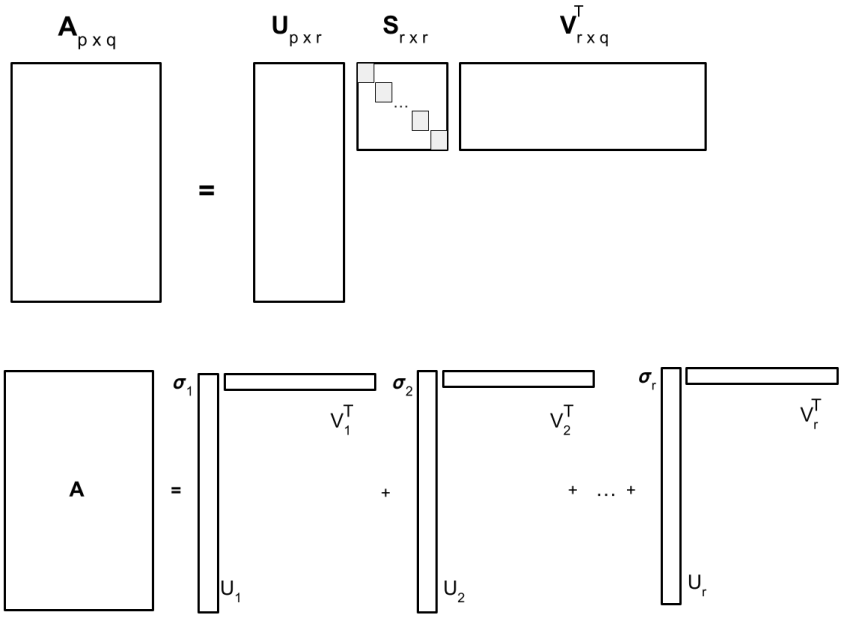


Figure 1.1: An overview of matrix decomposition using SVD, in matrix form corresponding to Eq. 1.1 (**top**) and dyadic form corresponding to Eq. 1.2 (**bottom**).

where U and V have orthonormal columns, and S is essentially diagonal. Because the diagonal matrix is located between the orthogonal matrices, Eq. (1.1) can be expanded as the following simple sum:

$$A = \sum_{i=1}^{r} \sigma_i U_i V_i^T, \tag{1.2}$$

where U_i, V_i are the i-th columns of U and V, respectively. Chapter 2 explains the SVD method in details.

1.2. RECASTING TIME SERIES AS MATRICES

To see how matrix decomposition can be applied to time series analysis, let us start with the simplest conceivable example. Suppose we have a perfectly periodic signal (e.g., a pure sine wave) which is sampled at some (large) multiple of the wavelength. Suppose in addition that we observe q (identical) cycles, with each cycle having p sample points. Let us register the observations during one cycle in the p-dim vector $\mathbf{p}$. For reasons that will become clear shortly, let us re-arrange the times series as a matrix, with each cycle in a separate column (from now on we assume that vectors are columns, see Figure 1.2). In that case, we can express the data concisely as data matrix A (denoting the column vector $\mathbb{1}_q = (1,1,\ldots,1)^T$):

$$A = \underbrace{[\mathbf{p},\mathbf{p},\ldots,\mathbf{p}]}_{q \text{ cycles}} = \mathbf{p}\mathbb{1}_q^T \tag{1.3}$$

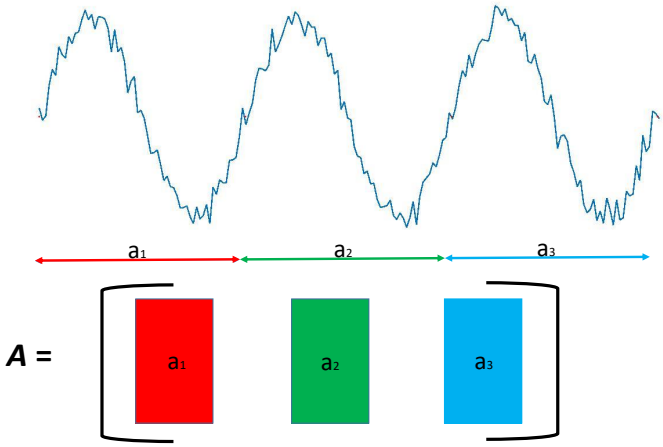


Figure 1.2: Turning a periodic signal into a matrix. The data of each successive period are stacked next to each other as columns of the matrix.

Now suppose that we are dealing with a time series that still has a clear and fixed period, by which we mean that the length of each cycle is constant throughout the observed time window (e.g., due to diurnal activities). The amplitude, however, can vary significantly and erratically from cycle to cycle. This type of time series is often encountered in smart infrastructures. As a case in point, consider traffic data [1]:

- These data show fixed periods of 24 hours (due to human activity);
- Excluding weekends, the shape of the observed patterns is similar, reflecting patterns in human activity (e.g., schools and businesses start at a certain time every morning);
- However, the amplitude could differ from day to day: e.g., depending on the weather — which to a first approximation is stochastic;

Pursuing the stance taken by Eq. (1.3) we represent such a variable-amplitude as:

$$A = \mathbf{p}\mathbf{q}^T \tag{1.4}$$

where the q-dimensional vector $\mathbf{q}$ now summarizes the amplitude information.

In fact, it is customary to recast the $\mathbf{p}$ and $\mathbf{q}$ vectors as unit vectors and collect the factored-out amplitudes in a single coefficient (customarily denoted as σ):

$$A = \mathbf{p}\mathbf{q}^T = \sigma\mathbf{u}\mathbf{v}^T \quad \text{where} \quad \mathbf{u} = \mathbf{p}/\|\mathbf{p}\|, \quad \mathbf{v} = \mathbf{q}/\|\mathbf{q}\| \quad \text{and} \quad \sigma = \|\mathbf{p}\| \cdot \|\mathbf{q}\|. \tag{1.5}$$

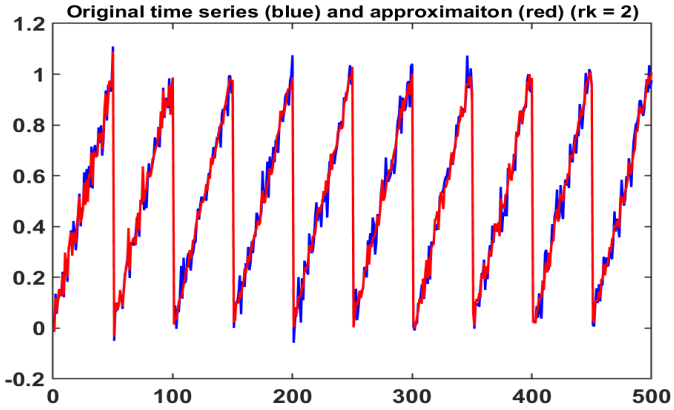


Figure 1.3: A toy example of a noisy time series (blue) that has 10 cycles with length 50. The low-rank approximation (see main text for more details) is drawn in red.

For an example of such a time series, see Figure 1.3. The noisy time series (blue) has cycles of length p =50, but the amplitude of each cycle varies erratically.

Finally, real data are frequently noisy, so a more appropriate representation of the data matrix A would be:

$$A = \sigma \mathbf{u}\mathbf{v}^T + \varepsilon Z \tag{1.6}$$

where Z has the same size as A and represents independent unit-variance (Gaussian) noise, while ε determines the noise amplitude. Notice that in general, adding a noise matrix will inflate the rank rk(A) from rank 1 to full rank. However, as long as the noise is small with respect to the signal, the matrix is essentially — in a sense that will be made precise later on in Section 1.3 — still a rank-1 matrix. Therefore, we have constructed a data matrix A that is essentially rank-1 as the superposition of a pure rank-1 matrix and a (small amount of) random noise. In what follows we will take the complementary view: take a given data matrix, and decompose it into a rank-1 (or more generally, low rank) approximation and some remaining noise.

Up to this point, the examples are artificial but throughout most of this thesis, we will take the perspective of **smart energy systems** applying data-driven approaches to analyse time series data generated by this type of smart infrastructures that often contain underlying patterns that facilitate low-rank representations.

Figure 1.4 illustrates an example of one year worth of price and load data (Top), alongside their alternative representation (Bottom). The latter plots were obtained by recasting each time series into a matrix of size 24×365. In Chapter 4, we will elaborate on this and put these visual impressions on a more sound, mathematical footing.

1.3. APPLYING SVD TO TIME SERIES

In the previous section, we have indicated how we can recast a time series as a matrix. The reason for this alternative representation is to take advantage of matrix decomposition techniques to extract useful information from data. In this thesis, we will mainly

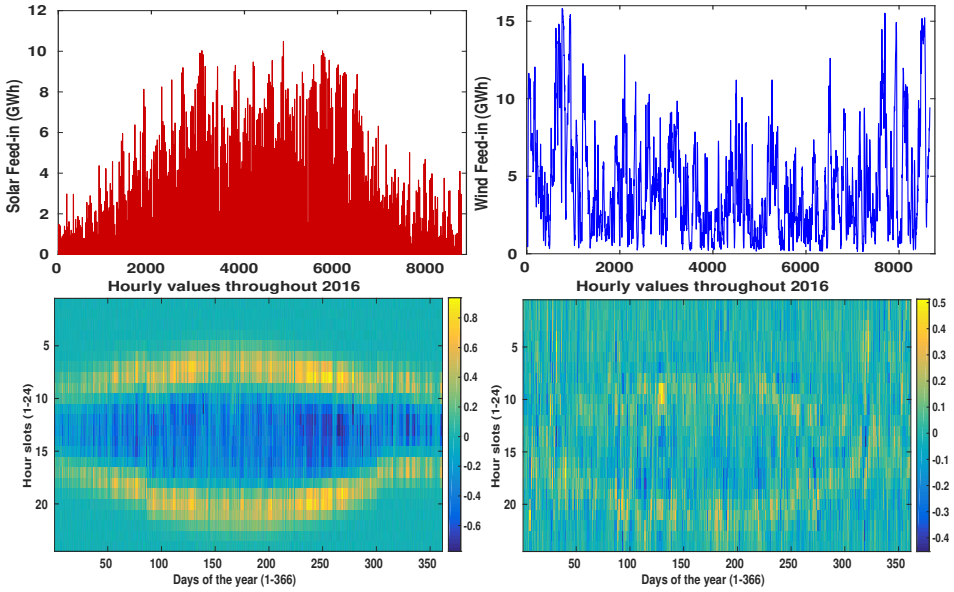


Figure 1.4: **Top:** time series of hourly solar (left) and wind (right) energy feed-in. **Bottom:** Matrix representation of the derivatives of the solar (left) and wind (right) data. For full description of the data and the methodology see Chapter 4.

focus on the following:

- Period extraction
- Approximation
- Visualisation
- Pattern extraction

In what follows we will look at each of these applications in slightly more details.

1.3.1. PERIOD EXTRACTION

To handle period estimation for such aforementioned time series, the authors in [2] reasoned as follows: If we have a noiseless time series with periods that are identical in shape but possibly vary in amplitude, it could be recast as the rank-1 matrix in Eq. (1.4). Hence we could determine the period by reshaping the time series as a matrix with different dimensions (i.e., number of rows and columns) until we hit upon a matrix that has rank 1. The corresponding number of rows would correspond to the sought-after period.

In practical applications, however, it is rare to come across time series that yields a data matrix of this exact simple form. Therefore, it is more realistic to think of Eq. (1.5) as the first term in an expansion:

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots \tag{1.7}$$

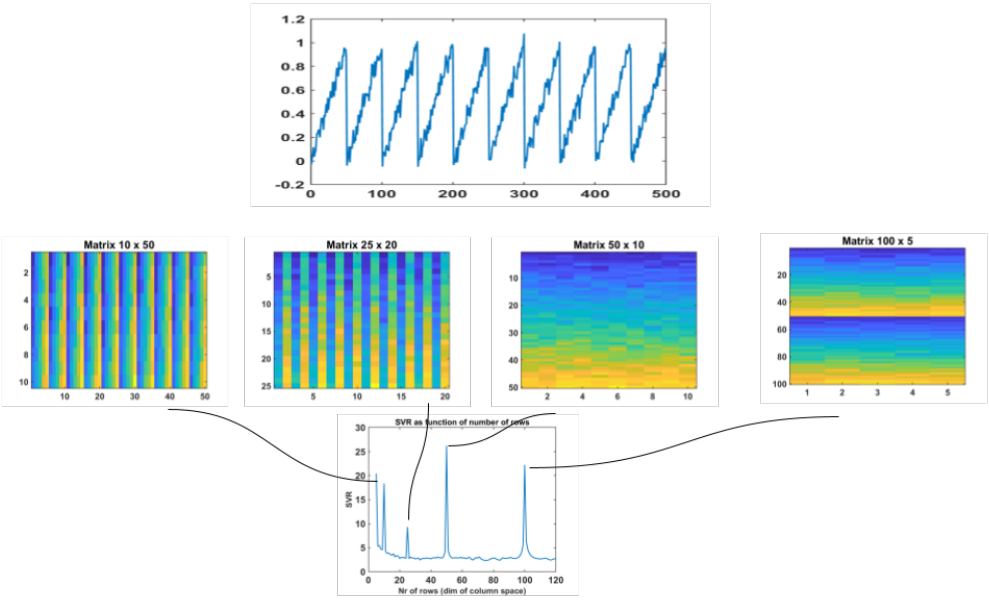


Figure 1.5: Determining time series period based on SVR. The input time series (**top** panel) is reshaped in matrices of different column-length (**middle panels**). For each of these matrices we compute the SVR = σ_1/σ_2 and plot it against the column size (p). The actual time series period produces a (dominant) peak in the SVR-curve (**bottom** panel).

Worth to be noted that the smaller the amount of noise is, the closer Eq. (1.7) is to Eq. (1.5). The factorization in Eq. (1.7) can be efficiently computed using the singular value decomposition (SVD – for a recap of the important results, see Chapter 2). For such an approximation to make sense, we assume that the signal (captured by the first term) dominates the noise (captured by the subsequent terms). Mathematically, this amounts to the assumption that $\sigma_1 \gg \sigma_2$, or equivalently $\sigma_1/\sigma_2 \gg 1$. This latter expression is called the **singular value ratio** (SVR). This observation was the starting point for the authors in [2, 3] who proposed the following simple procedure for period extraction (also see Figure 1.5):

- For a given time series of length n, rearrange the data as successive columns in a matrix. For each length p of the column this produces a differently shaped matrix of size $p \times q$ where $q = \lfloor n/p \rfloor$[1].
- For each different value of p, compute the SVD expansion in Eq. (1.7) and the corresponding $SVR = \sigma_1/\sigma_2$, for the corresponding matrix.
- When the dimension of the column corresponds with the actual underlying periodicity, there will be a peak in the SVR. Hence, by identifying the peaks in the $p - SVR$ plot, one can determine the periodicity.

[1] where [.] is the Greatest Integer Function.

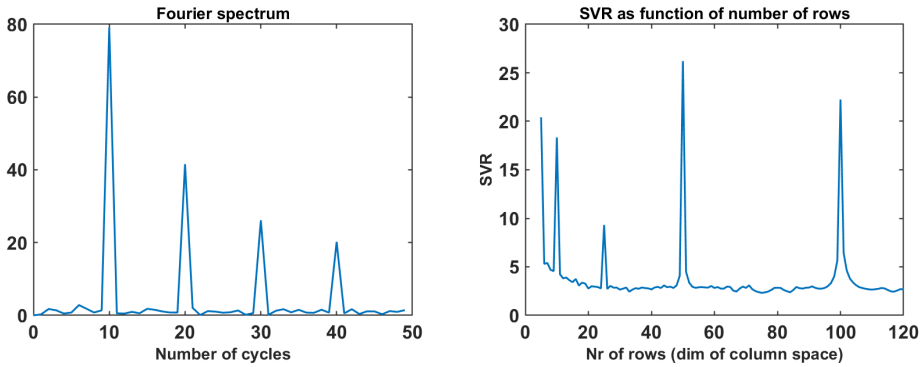


Figure 1.6: Illustrating alternative ways to quantify the periodicity of the signal in Figure 1.3. **Left:** FFT spectrum for time series in Figure 1.3. The highest spectrum indicates the number of cycles in the time series. As can be seen in the aforementioned figure, it is 10. **Right:** Period extraction based on singular value ratio (SVR) plotted as a function of the cycle length. The dominant SVR peak indicates the correct cycle length (50).

Comparing SVR and FFT for period extraction When introducing a new method for period estimation it is only natural to ask oneself how it compares to standard methods such as Fast Fourier Transform (FFT). Figure 1.6 illustrates the above methodology along with the outcome of the FFT algorithm for the time series of Figure 1.3. It is evident that SVR finds the correct period of the signal (p), whereas FFT finds the correct number of cycles (q) which in turn leads to the *correct* p value.

Fourier transform, however, falls short when it comes to more complex time series, especially with erratic behaviour. Figure 1.7 illustrates an example of a periodic signal that is amenable to period extraction by SVR but not by FFT. In other words, the reconstructed approximation need not be continuous. This produces high frequency noise due to successive under and overshoots. The reason for that is the sign changes in the amplitude that wreak havoc on the FFT spectrum, as can be seen in the left panel of Figure 1.8, (it shows the FFT-spectrum for the time series in Figure 1.7). In fact, it turns out that the FFT-power at 10 cycles (which is the correct number of cycles) is a relative minimum rather than a maximum. However, $p - SVR$ plot correctly shows a spike at $p = 50$.

Another difference between approximations based on SVD and Fourier is the choice of basis function used in the expansion. This will be addressed in the next section.

1.3.2. TIME SERIES APPROXIMATION

Fourier analysis decomposes a given signal into a fixed set of sine-waves with steadily increasing frequencies. This works well for signals that are relatively slowly varying. However, if the signal includes jumps or swings, FFT needs to aggregate a large number of terms in order to reach an acceptable accuracy. In contrast, the basis functions used in the SVD approach are obtained from the data at hand, and are therefore be determined in a **data-driven** fashion. The following example illustrates this observation.

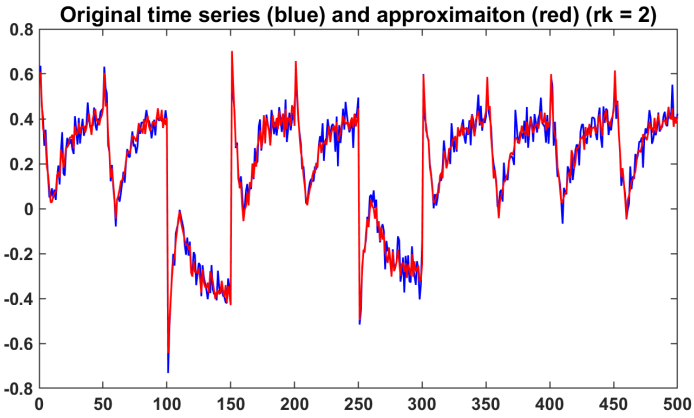


Figure 1.7: An example of a noisy more complex signal that has 10 cycles with length 50), but the amplitude of each cycle varies erratically. The low-rank approximation (see main text for more details) is drawn in red.

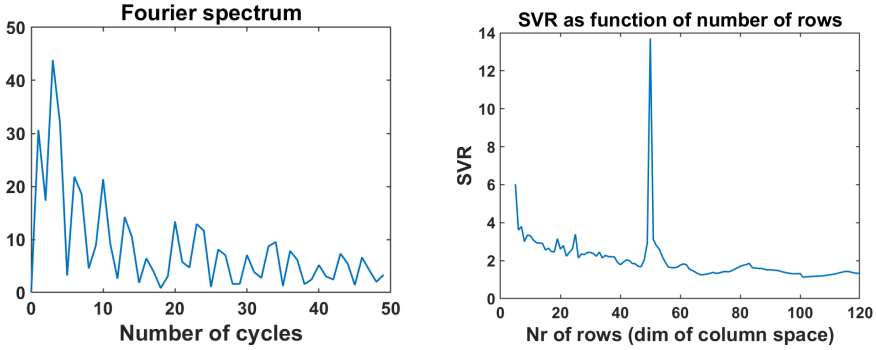


Figure 1.8: **Left:** FFT spectrum for time series in Figure 1.7. The spectrum fails to identify the correct number of cycles (which is 10). **Right:** Period extraction based on singular value ratio (SVR). There is a clear peak that the correct cycle length (which is 50).

EXAMPLE: BURSTS WITH EXPONENTIAL DECAY

The left panel of Figure 1.9 provides an example of a noisy periodic time series with sharp up-swings followed by exponential decays. In this example the underlying signal has a single and constant wavelength λ with exponentially decaying bursts that are repeated at the beginning of every wavelength:

$$x_0(t) = \exp(-\alpha t/\lambda) \qquad 0 \le t \le \lambda$$

To this uncorrupted signal we add various amounts of Gaussian noise:

$$x(t) = x_o(t) + \sigma w(t) \quad , \quad w(t) \sim N(0,\sigma^2)\, i.i.d$$

Because of the burst-like nature of the signal, Fourier analysis is not well-suited for signal reconstruction in this case. The right panel of Figure 1.9 illustrates the power spec-

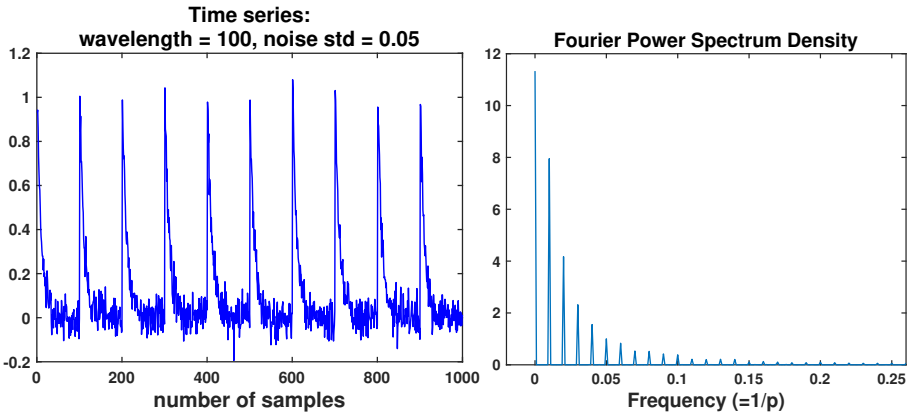


Figure 1.9: **Left:** Burst-like signal with exponential decay, wavelength $p = 100$, amplitude = 1, Gaussian noise std = 0.05. **Right:** The corresponding Fourier power spectrum. The slow decay of the power spectrum is an early indication that approximation might be problematic

trum density (PSD) of the aforementioned time series. The slow decay of the power spectrum implies that FFT is not well suited to recover the underlying patterns of such time series. That is the prime motivation in many applications to search for more data-driven methods.

Figure 1.10 provides examples of the reconstruction of the aforementioned time series using the Fast Fourier Transform (FFT) approach with a different number of components. It is evident that FFT, in low-order approximations, is incapable of fully reconstructing the signal due to the burst-like shape of the signal. The left panel of Figure 1.11 illustrates the residuals of the FFT based approach with 9 components. The spikes in this figure indicates that 9 components were not enough for the FFT to be able to fully capture the high-frequency burst-like shape of the time series. A comparison of the aforementioned residuals with a normal distribution is provided on the right panel.

Figure 1.12 illustrates the first two left and right singular vectors (obtained by recasting the original time series as a matrix and then applying the SVD). We will discuss the SVD approach in more detail in Chapter 2. The rank-2 SVD-based reconstruction of the time series along with its residuals is illustrated in Figure 1.13. The corresponding residuals for this rank-2 reconstruction are presented in Figure 1.14. As can be seen on the right panel the residuals nicely adhere to the normal distribution.

Whereas Fourier decomposes all signals into sine waves, this is not the case for SVD. Complicated waves will give rise to complicated initial profiles. Put differently, whereas low-order approximations based on Fourier will be smooth and clearly sinusoidal, low-rank SVD can be arbitrarily complicated. This is already borne out in the example in Figures 1.10 and 1.12.

1.3.3. VISUALIZATION

In the previous section, we explained how reshaping a time series as a matrix suggests an alternative way to determine the periodicity. Furthermore, there are additional advan-

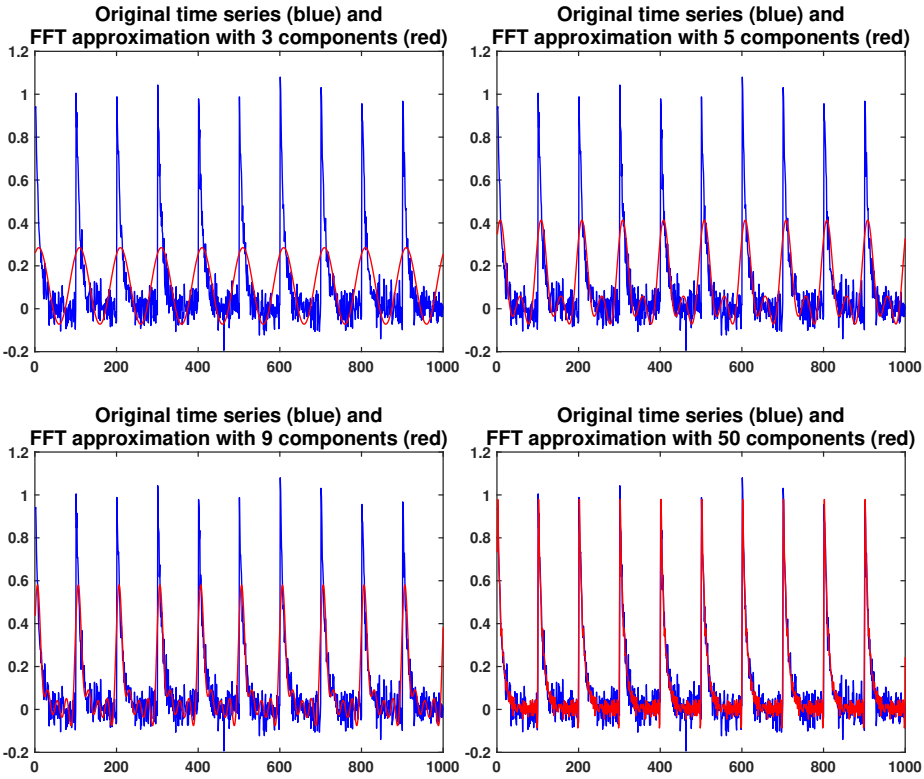


Figure 1.10: Signal reconstruction for the time series on the left panel of Figure 1.9. Restricting the number of Fourier components to 3, 5, 9 and 50 fails to capture the discontinuous nature of the bursts.

tages of such alternative representation. This change of viewpoint has two important applications:

- Representing time series as images allow one to visually integrate patterns across longer time spans, and hence improves the discriminatory power.
- Recasting such time series as matrices also suggest drawing on matrix decomposition theorems to elucidate the underlying structure by constructing approximations which are more tightly linked to the structure of the time series.

The SVD factorization suggests a straightforward method to smooth time series in such a way that the overall structure is preserved. Figure 1.15 provides another example of how SVD can be used to elucidate the underlying patterns in the data. Chapter 4 provides a detailed description of such a technique. For a given noisy time series, one constructs the corresponding data matrix A and then applies SVD to construct a low-rank approximation A_r which is then re-expanded as time series. This is illustrated in Figure 1.7 where a smoothed (red) based on a rank-2 approximation is constructed for the noisy signal (blue).

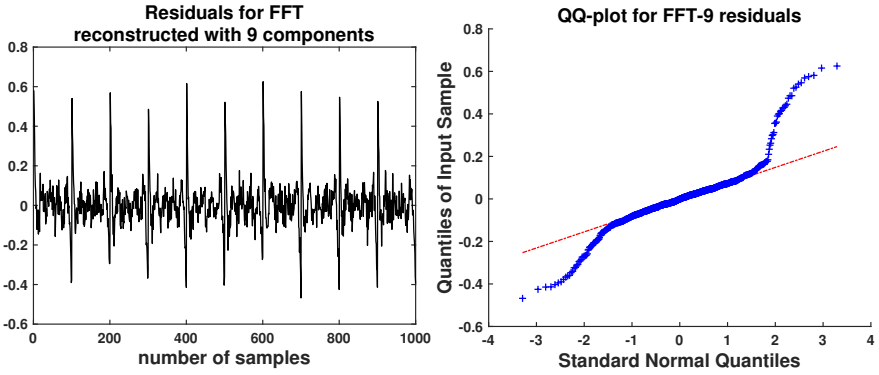


Figure 1.11: **Left:** The residuals and its distribution compared to a normal distribution for a FFT-based reconstruction with 9 components As expected due to the burst-like shape of the signal FFT could not fully capture the pattern of the original time series. **Right:** QQ-plot of the residuals. Due to the spikes, the residuals are not fully normally distributed.

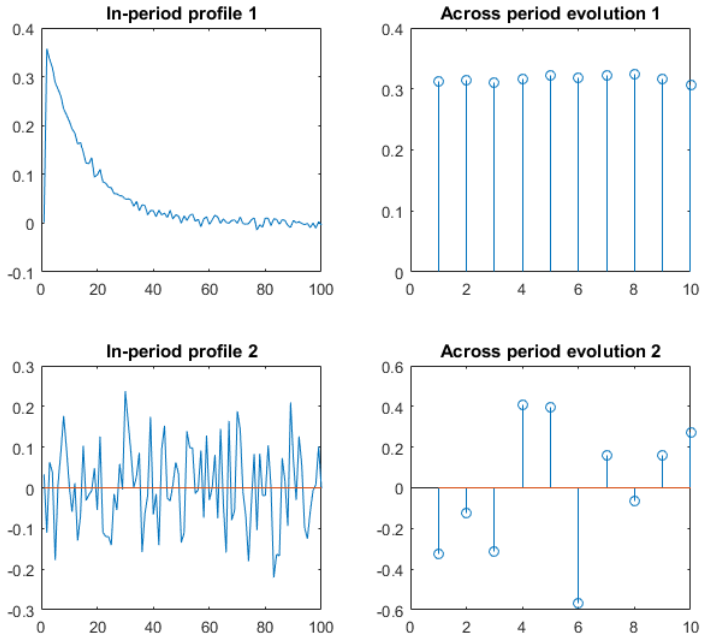


Figure 1.12: Burst-like signal with exponential decay, wavelength = 100, amplitude = 1, Gaussian noise std = 0.05. The extracted profile (top left) clearly shows the exponentially burst profile, while the corresponding amplitudes (top right) indicate that the profile is repeated with constant strength. The second profile (bottom left) looks like pure noise, an impression further corroborated by the randomly distributed amplitudes (bottom right).

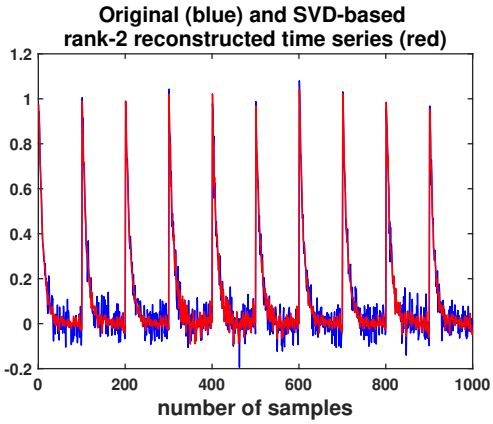


Figure 1.13: SVD-based rank-2 reconstruction of the time series. The details of the SVD is discussed in Chapter 2.

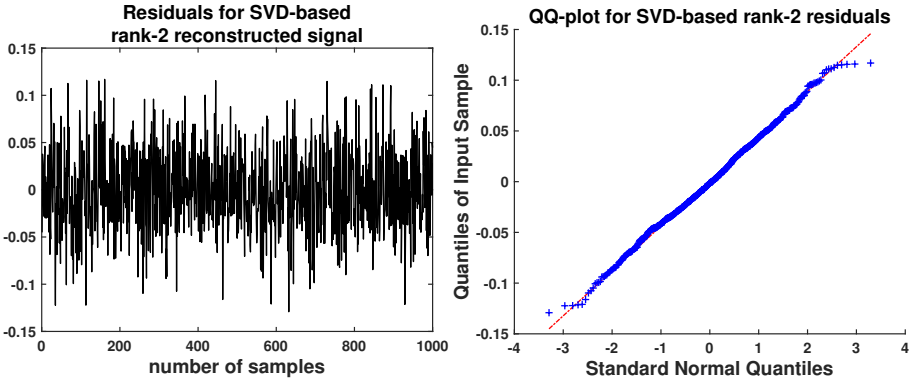


Figure 1.14: **Left:** Residuals for the SVD-based rank-2 reconstruction of the time series. In contrary to the FFT where even by considering 9 components the reconstruction was not satisfactory (see Figure 1.11), here a rank-2 reconstructed time series accurately captures the underlying pattern of the time series. **Right:** The residuals are almost fully normally distributed.

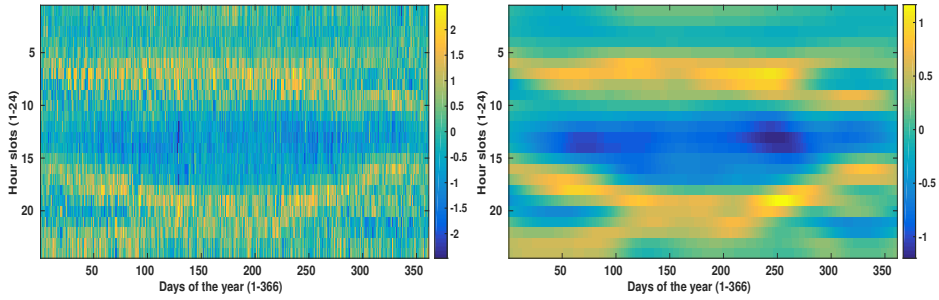


Figure 1.15: An example of enhancing the underlying patterns in the data using lower-rank approximation (the original data on the left and the enhanced version on the right)

1.3.4. PATTERN EXTRACTION

Figure 1.16 provides an illustrative example of the SVD decomposition. The top left panel shows a noisy zero-mean block signal of length $n = 1000$ with a pronounced period $p = 100$ and $q = 10$ full cycles. In addition to the noise, there are three irregularly occurring spikes. After recasting this time series as a 100×10 matrix A, we then apply the SVD algorithm to obtain $A = USV^T$ where S is a 100×10 "rectangular diagonal" matrix with the 10 singular values on its main diagonal. The top right panel shows those ten singular values, clearly illustrating that all except the first two are negligible, which means that the matrix (and therefore the time series) can be accurately represented by truncating the expansion in Eq. (2.5) after the first two terms, i.e., rank-2 approximation (see Figure 1.17). Finally, the bottom panel of Figure 1.16 displays the first three columns of U (left) and V

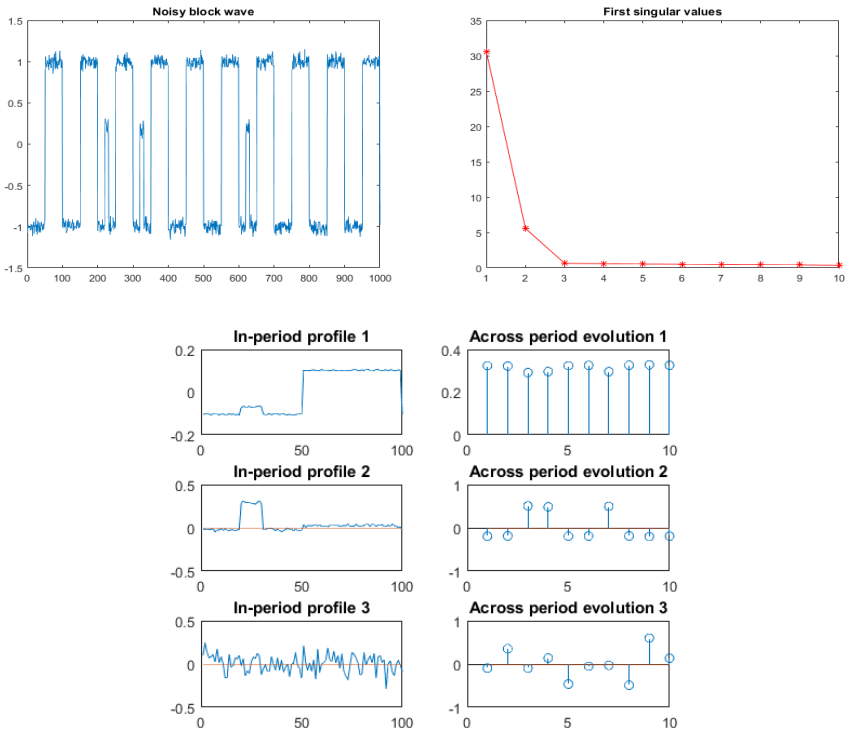


Figure 1.16: SVD application to pattern-extraction in noisy block signal. **Top left:** Original data of noisy block signal with period 100. In addition to the noise there are three irregularly occurring spikes. **Top right:** The 10 singular values for SVD with period $p = 100$. Clearly, only the two first are significant and $\sigma_1 \gg \sigma_2$ confirming that $p = 100$ corresponds to a valid periodicity. **Bottom:** The first three columns of U (left) and V (right).

(right), respectively. As they correspond to the most significant singular values, they are most important for the reconstruction of the signal. The U-columns cover one cycle and can be interpreted as successive profiles needed to reconstruct a generic cycle. In that sense, they are analogous to various trigonometric basis functions in Fourier analysis. The V-columns, on the other hand, specify the amplitudes with which these basis func-

tions need to be combined in order to reproduce the individual cycles observed in the data. Not surprisingly, the main profile (U_1 top left) reflects the step-like behaviour seen during each cycle. As the amplitude of each of these steps is essentially constant, the 10 V_1-entries displayed in the top-right panel show little variation. The U_2 profile (middle, left) captures the shape of the additional spikes that occur at irregular intervals. The positive values in the corresponding V_2-coefficients (middle, right) clearly indicate in which intervals these spikes occur. Finally, the erratic appearance of both U_3 and V_3 is a further indication (in line with $\sigma_3 \approx 0$) that all structural information has been extracted from the signal.

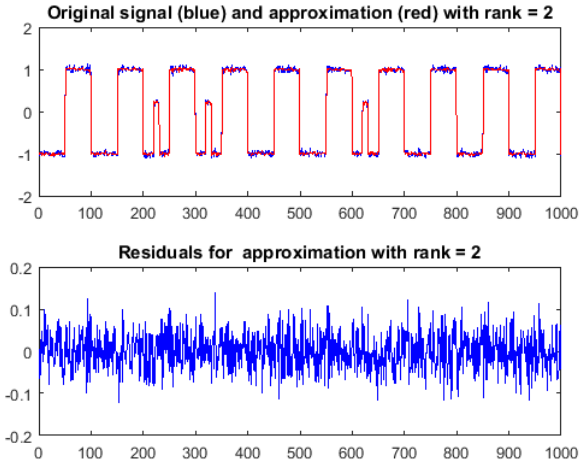


Figure 1.17: **Top:** Original (blue) and rank-2 approximation (red) of the block-signal. **Bottom:** Residuals with respect to the approximation.

1.4. CHALLENGES AND RESEARCH QUESTIONS

So far in this chapter, we have outlined the essential theoretical concepts of the SVD that are used throughout this thesis. We also have introduced the research domain and main motivations of this thesis. In this section, we delve deeper into specific issues that arise in time series analysis, and the SVD in its traditional form. We further distinguish various problems in energy market analysis. In particular, we outline the following research questions:

Research Question 1 As explained in section 1.3.1 the authors in [2] introduced the idea of using the ratio of the two largest singular values (i.e., the so-called singular value ratio SVR) to determine the period of a time series (signal). However, they did not realise that the SVR depends on the signal mean. As a consequence, thresholding the SVR value is meaningless unless we take this effect into account. This is the topic of our first research question:

> **Q1: How does the mean of a signal affect the singular value ratio SVR?**

Chapter 2 provides a thorough understanding of the SVD result and its geometrical interpretation. This chapter also contains a number of examples of how the position and orientation of the same cloud of points from the origin have led to different SVD results. After providing a background on random matrices in Chapter 3, we have delved deeper into this question and have measured the effect of distance from the origin on the singular values, in an example.

Research Question 2 In the above setup, the number of columns corresponds to the number of observed cycles and will therefore increase as more data are acquired. Indeed, as each column represents the data from one period, observing more periods gives rise to additional columns, making the matrix "fatter" and thus increasing the aspect ratio (i.e. number of columns over the number of rows). Therefore, it is natural to ask how changes in the aspect ratio will affect the results. Many math papers focus on square matrices, but for applications to time series, it is important to understand the more general case of rectangular matrices. This is addressed by the next research question:

> **Q2: How does the aspect ratio of a matrix affect the singular value ratio (SVR)?**

In Chapter 3, we have extended the work in earlier papers by initiating a more systematic analysis of these effects.

Research Question 3 Matrix factorisation techniques such as principal component analysis (PCA) and singular value decomposition (SVD) are both conceptually simple and effective. However, it is well-known that they are sensitive to noise and outliers in input data. One way to mitigate this sensitivity is to introduce *regularisation* terms. In order to do this, in Chapter 4 we hark back to the interpretation of SVD and PCA in terms of low-rank approximation. Adding regularisation terms to these functionals gives rise to new but related minimisation problems.

- **Q3 a: How can we develop a regularized version of the PCA problem?**

  This problem is addressed in Section 4.2.

- **Q3 b: How can we develop a regularized version of the SVD problem?**

  This problem is addressed in Section 4.3.

Research Question 4 Scenario-based probabilistic forecasting models have been extensively explored in the literature in recent years. A particular application of such models is in the energy sector, where e.g., having the distribution of the energy consumption for the coming days is desired. A decisive variable in predicting the energy demand (target variable) is the temperature data (an attribute or a predictor) [4]. There are mainly three practical and popular methods for generating temperature scenarios, namely fixed-date, shifted-date, and bootstrap approaches [5]. Nevertheless, these methods have mostly been used on an ad-hoc basis without being formally compared or quantitatively evaluated.

> **Q4: In scenario-based probabilistic forecasting problems, how can we simulate realistic profiles for an independent variable?**

In Chapter 5 we propose a generic, data-driven and computationally efficient SVD-based approach to simulate different scenarios.

Research Question 5 Volatility principally refers to random fluctuations of a time series about its mean or expected value. Generally speaking, in financial time series data analytics, volatility is measured by the standard deviation of the logarithmic return or a derivation of that [6]. In the literature, numerous methods have been introduced to determine the volatility of the time series data. However, the emergence of non-positive price values in the *energy transition* era has introduced new challenges in the market volatility analysis.

> **Q5 a: How can we quantify the volatility of the electricity price time series data with zero or negative prices?**

In the first part of Chapter 6, we will address this problem and propose a notion of volatility that can handle negative prices. Furthermore, along with the increase in the utilization of intermittent renewable sources, short-term electricity market studies are becoming increasingly popular. Therefore, in the final part of this chapter we address the following question:

> **Q5 b: Has the stimulation of renewable energy sources led to a noticeable changes in the (day-ahead) electricity market?**

In the final part of Chapter 6, we will analyse the evolution of the day-ahead electricity market in Germany in 2006-2016.

1.5. THESIS OUTLINE

The rest of this thesis is structured as follows. The SVD technique is the cornerstone of this thesis. Therefore, we first in Chapter 2 review the relevant theorems underpinning this method. We then delve deeper into the geometrical interpretation of the SVD. This chapter also provides a number of examples of how the position with respect to the origin and the alignment of data points affects the singular vectors and the singular values. That in turn enables us to have an intuitive answer for *Research Question 1*. However,

more in-depth discussions over the first two research questions are provided in the next chapter.

For most applications of the SVD in various fields, it is vital to understand the properties of SVD of a matrix whose entries show some degree of random fluctuations. Therefore, in order to determine how the noise level affects the singular value spectrum, it is essential to study the singular value decomposition of random matrices. Having provided a background in random matrices, Chapter 3 addressed *Research Question 1* and *Research Question 2* in full detail.

The SVD and PCA techniques are both conceptually simple and effective. However, it is well-known that they are sensitive to the high level of presence of noise and outliers in input data. In the literature, some modifications of the original algorithms of SVD and PCA have been proposed to alleviate the effect of these disturbances. In particular, one way to mitigate this sensitivity is to introduce regularisation terms. To this aim, in Chapter 4 we first hark back to the interpretation of PCA and SVD in terms of low-rank approximations. We then have offered solutions to *Research Question 3a* and *Research Question 3b* in Section 4.2 and Section 4.3, respectively.

With the growing integration of renewable energy sources (RES) such as wind and solar energy into the power grid, balancing the grid has become more challenging. It is mostly due to the inherently intermittent nature of RES, on the one hand, and shortcomings in bulk energy storage systems, on the other. Therefore, studies on scenario-based probabilistic energy production and demand forecasts have gained momentum, as they are highly valuable from both a technical and an economic point of view [7]. Chapter 5 proposes a generic framework for probabilistic load forecasting using an ensemble of regression trees. This chapter proposes a solution to *Research Question 4* by generating various examples of a predictor (temperature in this case) using the SVD results. The generated samples are then used in an ensemble of regression trees to obtain the distribution of the target variable (load profile) in future times.

Chapter 6 is dedicated to *Research Question 5*, i.e., what effect the transition of energy to RES can have on the overall trend and also the volatility of the electricity prices. In this chapter, we exemplify how the emergence of zero or even negative price values in the day-ahead market in Germany in recent years has introduced new challenges in the electricity market volatility analysis. More precisely, in this new market, the traditional approaches to switch to logarithmic measures can only be done after shifting up all values above zero by a certain threshold. However, price volatility has a dependence on the price level, which is even more pronounced when the spot prices are low. Therefore, the aforementioned *pre-processing* step can affect the final outcome and its generalizability. The first part of this chapter offers a solution to *Research Question 5a* by introducing a new notion of volatility which was obtained by reconstructing the time series using the SVD. Our observations indicate price volatility reduction, in the day-ahead market, in the years 2006-2016. The second part of this chapter addressed *Research Question 5b* and provided pieces of evidence of the effect of renewables on daily price profiles – the emergence of non-positive prices and also shifts of peak price values to hours where solar is less available. A summary of this thesis along with some future research directions is provided in Chapter 7.

1.6. LIST OF PUBLICATIONS

In this section, we present an overview of the research publications comprising this thesis.

Journals

1. Abdolrahman Khoshrou, Eric J Pauwels. **Regularisation for PCA-and SVD-type matrix factorisations**. preprint 2021. Springer - Advances in Computational Intelligence[2].

2. Abdolrahman Khoshrou, and E.J. Pauwels. **Short-term scenario-based probabilistic load forecasting: A data-driven approach**. 2019. Elsevier - Applied Energy.

3. Abdolrahman Khoshrou, André Dorsman, Eric J. Pauwels. **The evolution of electricity price on the German day-ahead market before and after the energy switch**. 2019. Elsevier - Renewable Energy.

Conferences

1. Abdolrahman Khoshrou, Eric J. Pauwels. **Data-driven pattern identification and outlier detection in time series**. 2018. Springer, Cham - Science and Information Conference.

2. Abdolrahman Khoshrou, Eric J Pauwels. **Quantifying volatility reduction in German day-ahead spot market in the period 2006 through 2016**. 2018. IEEE - Power & Energy Society General Meeting (PESGM).

3. Abdolrahman Khoshrou, André Dorsman, Eric J. Pauwels. **SVD-based Visualisation and Approximation for Time Series Data in Smart Energy Systems**. 2017. IEEE - Innovative Smart Grid Technologies Conference Europe (ISGT-Europe).

4. Abdolrahman Khoshrou, Eric J Pauwels. **Propagating uncertainty in tree-based load forecasts**. 2017. IEEE - Electrical and Electronics Engineering (ELECO), 2017 10th International Conference.

5. André Dorsman, Abdolrahman Khoshrou, Eric J. Pauwels. **The influence of the switch from fossil fuels to solar and wind energy on the electricity prices in Germany**. 2016. ISINI conference in Groningen.

[2] Part of this work was published at Belgian-Netherlands Artificial Intelligence Conference (BNAIC) 2021.

REFERENCES

[1] V. Verendel and S. Yeh, *Measuring traffic in cities through a large-scale online platform,* Journal of Big Data Analytics in Transportation **1**, 161 (2019).

[2] P. P. Kanjilal and S. Palit, *On multiple pattern extraction using singular value decomposition,* IEEE transactions on signal processing **43**, 1536 (1995).

[3] P. P. Kanjilal and S. Palit, *The singular value decomposition—applied in the modelling and prediction of quasi-periodic processes,* Signal processing **35**, 257 (1994).

[4] A. Khoshrou and E. J. Pauwels, *Short-term scenario-based probabilistic load forecasting: A data-driven approach,* Applied Energy **238**, 1258 (2019).

[5] T. Hong *et al.*, *Energy forecasting: Past, present, and future,* Foresight: The International Journal of Applied Forecasting , 43 (2014).

[6] *Financial chaos theory,* http://quantonline.co.za/Articles/article_volatility.htm.

[7] T. Hong and S. Fan, *Probabilistic electric load forecasting: A tutorial review,* International Journal of Forecasting **32**, 914 (2016).

2

Singular Value Decomposition: A Recap

2.1. Introduction

The singular value decomposition (SVD) technique is the cornerstone of this thesis. We hence start this chapter by reviewing the relevant theorems underpinning the SVD method.

Before proceeding with that, however, let us first recall the definition of a (multiplicative) group of orthogonal matrices of dimension n:

$$\mathcal{O}(n) := \left\{U \in R^{n\times n} \mid UU^T = I_n = U^T U\right\} \tag{2.1}$$

Notice that from the definition it immediately follows that

$$\det(U) = \pm 1 \tag{2.2}$$

since

$$\det(UU^T) = \det(U)\det(U^T) = (\det(U))^2 = \det(I_n) = 1.$$

This observation motivates the introduction of a subgroup of special orthogonal matrices of unit determinant:

$$SO(n) := \left\{U \in R^{n\times n} \mid UU^T = I_n = U^T U \text{ and } \det(U) = 1\right\} \tag{2.3}$$

2.2. Singular Value Decomposition

As we exemplified earlier in Chapter 1, the FFT works in an idealized setting. Furthermore, the SVD, in a sense, generalizes the concept of FFT. In other words, SVD allows us to transform or tailor a coordinate system, based on the data itself (data-driven). It is a widely adaptive method and is mostly based on simple and interpretable linear algebra. As a result of that, every time we have a matrix of data, we can compute the SVD and address different problems based on that [2]. The following sections provide a detailed description of the SVD and its geometrical intuitions.

Parts of this chapter have been published in [1].

2.2.1. SVD: MAIN RESULT

We herein recall the well-known SVD technique and develop an intuition for how to apply it in the following chapters. For more details, we refer to standard textbooks such as [3, 4].

Theorem 1 (**Singular Value Decomposition**). *Any real-valued $p \times q$ matrix A can be factorized into the product of three matrices:*

$$A_{p\times q} = U_{p\times p} S_{p\times q} V^T_{q\times q} \tag{2.4}$$

where $U \in \mathcal{O}(p)$ and $V \in \mathcal{O}(q)$ are orthogonal, and S is a $p \times q$ diagonal matrix where the elements on the main "diagonal" (so-called singular values) are non-negative (i.e., $\sigma_i := S_{ii} \geq 0$ for $1 \leq i \leq \min(p,q)$).
Assuming that the rank $rk(A) = r \leq \min(p,q)$, we can sort the singular values such that

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 = \sigma_{r+1} = \ldots = \sigma_{\min(p,q)}$$

and recast Eq. (2.4) as (see Figure 2.1, bottom panel, for an "economic" expansion of SVD.)

$$A = \sum_{i=1}^{r} \sigma_i U_i V_i^T \quad \text{where } U_i, V_i \text{ are the } i\text{-th columns of } U \text{ and } V, \text{ respectively.} \tag{2.5}$$

For the singular values sorted as above, we then introduce the short-hand notation $U_{(1:k)}$ and $V_{(1:k)}$ to denote the matrix comprising the first k columns of U and V, respectively:

$$U_{(1:k)} := [U_1, U_2, \ldots, U_k] \qquad \text{and} \qquad V_{(1:k)} := [V_1, V_2, \ldots, V_k]$$

In this notation, Eq. (2.5) can be expressed concisely as (see Figure 2.1, top panel):

$$A = U_{(1:r)}\, diag(\sigma_1, \ldots, \sigma_r)\, V^T_{(1:r)} \tag{2.6}$$

□

To appreciate the significance of Theorem 1, it is helpful to highlight its geometric interpretation. Recall that any $p \times q$ matrix A gives rise to a corresponding linear transformation $A : R^q \longrightarrow R^p$ that maps the standard basis in R^q into the columns of A:

$$A\mathbf{e}_k = A_k \quad \text{where} \quad \mathbf{e}_k = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^T$$

Roughly speaking, the SVD theorem, therefore, tells us that it is always possible to select an *orthonormal* basis in R^q (columns of V) that is mapped (up to non-negative scaling factors, i.e., the singular values) into an *orthonormal* basis in $\mathbb{R}^p$ (columns of U). This is immediately obvious from Eq. (2.5):

$$AV_\ell = \sum_{k=1}^{r} \sigma_k U_k V_k^T V_\ell = \sum_{k=1}^{r} \sigma_k U_k \delta_{k\ell} = \sigma_\ell U_\ell$$

where $\delta_{k\ell}$ is a Kronecker delta function. It is worth noting that insisting on the orthogonality of V ($V^T V = I_q$) is not restrictive. Indeed, a linear transformation is completely

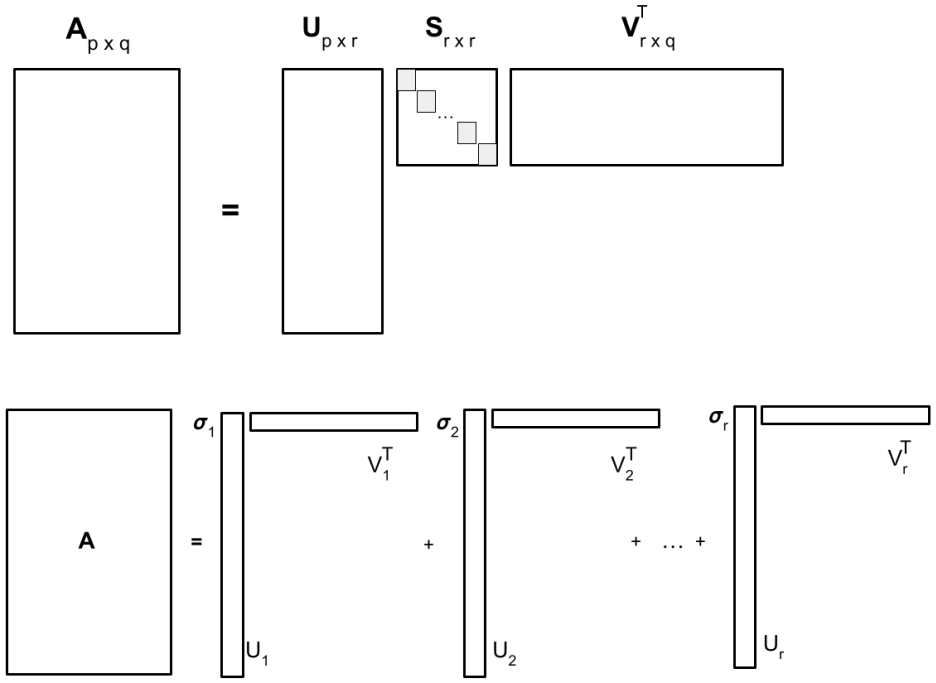


Figure 2.1: An overview of matrix decomposition using SVD, in matrix form (top) and dyadic form (bottom).

and uniquely determined by specifying its effect on any basis; hence, there is no loss of generality by insisting on the orthonormality of this basis.

However, the non-trivial message of this theorem is that orthonormal bases (V) can be chosen in such a way that their image (transformation) U under A is also orthonormal (up to non-negative scaling). Furthermore, in a generic case, where all the singular values are different, the SVD is unique up to an arbitrary relabeling of the *basis-vectors* and a simultaneous sign-flip of the corresponding columns in U and V, i.e., $(U_\ell, V_\ell) \to (-U_\ell, -V_\ell)$ for any number of columns.

2.2.2. SINGULAR VALUES ARE EIGENVALUES OF SQUARED MATRICES

From $A = USV^T$ it follows that $A^T = VSU^T$ and consequently[1]:

$$AA^T = USS^TU^T \qquad \text{and therefore} \qquad (AA^T)U = U(SS^T) \tag{2.7}$$

This means that the columns of U are eigenvectors of the positive definite, symmetric matrix AA^T, with positive eigenvalues:

$$\lambda_i(AA^T) = \sigma_i^2$$

[1] S is diagonal, hence $S = S^T$.

A similar observation can be made for V which turn out the be the eigenvectors of A^TA and again:

$$\lambda_i(A^TA) = \sigma_i^2$$

Because the matrices AA^T and A^TA are symmetric and semi-positive it follows that all the eigenvalues are real and non-negative.

In summary, we derive the following useful relationship between the singular values of a matrix $A \in \mathbb{R}^{p\times q}$ and the eigenvalues of the related matrices AA^T and A^TA:

$$\sigma_i^2(A) = \lambda_i(AA^T) = \lambda_i(A^TA) \tag{2.8}$$

or equivalently:

$$\sigma_i(A) = \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^TA)} \qquad \text{for } i = 1:\min(p,q) \tag{2.9}$$

where $i = 1:\min(p,q)$.

For a given matrix A we use the notation $\sigma_i(A)$ or $\lambda_i(A)$ to denote the i-th (in descending order) singular or eigenvalue, respectively. If there is no danger of confusion, the explicit reference to the matrix will be suppressed.

2.2.3. SVD SOLVES A MINIMISATION PROBLEM

The importance of the SVD result is the following well-known theorem (more details can be found in [5, 6]).

Theorem 2 (**Eckart-Young-Mirsky Theorem**[2])**.** *Let us consider a $p \times q$ matrix A with rank $rk(A) = r \le \min(p,q)$. For $k < r$, finding the rank-k matrix A_k that is closest to A in (Frobenius) norm gives rise to the following constrained minimisation problem:*

$$\min_{A_k} \|A - A_k\|^2 \qquad \textit{subject to} \quad rk(A_k) \le k$$

The solution to this problem is obtained by truncating the SVD expansion Eq. (2.5) after the k-th largest singular value:

$$A_k = \sum_{i=1}^{k} \sigma_i U_i V_i^T = U_{(1:k)}\, diag(\sigma_1, \ldots, \sigma_k)\, V_{(1:k)}^T \tag{2.10}$$

□

Recall that a rank-k matrix of size $p \times q$ can always be written as a product $A_k = PQ^T$ where $P \in R^{p\times k}$ and $Q \in R^{q\times k}$ are matrices of full rank k. In this factorisation, there is no loss of generality in requiring $Q^TQ = I_k$. In fact, it is necessary to remove indeterminacy due to arbitrary but trivial rescalings such as $P \longmapsto rP$ while $Q \longmapsto (1/r)Q$ (with $r \neq 0$), and the like. We will discuss this alternative formulation of Theorem 2 as the factorisation result in Theorem 3.

[2] Also referred to as the optimal low rank approximation theorem.

2.3. GEOMETRIC INTERPRETATION

The following section describes important mathematical properties of SVD including geometric interpretations of the unitary matrices U and V.

2.3.1. INCREMENTAL SVD: COMPUTING BEST RANK-1 APPROXIMATION

The general SVD result can be obtained incrementally, by constructing the best rank-1 approximation, then subtracting this approximation and repeating this procedure. This follows directly from Eq. (2.5) which can be recast as:

$$A - \sigma_1 U_1 V_1^T = \sum_{i=2}^{r} \sigma_i U_i V_i^T \tag{2.11}$$

which shows that the next term in the expansion can be obtained by computing the SVD of the residual $A - \sigma_1 U_1 V_1^T$. Using this insight, it follows that we can focus on computing the first singular value and vector. This is helpful as it turns out that the first singular value and vectors have a straightforward interpretation which we will explain next (also see Figure 2.2). To extract the first singular value and vectors of A we proceed as follows:

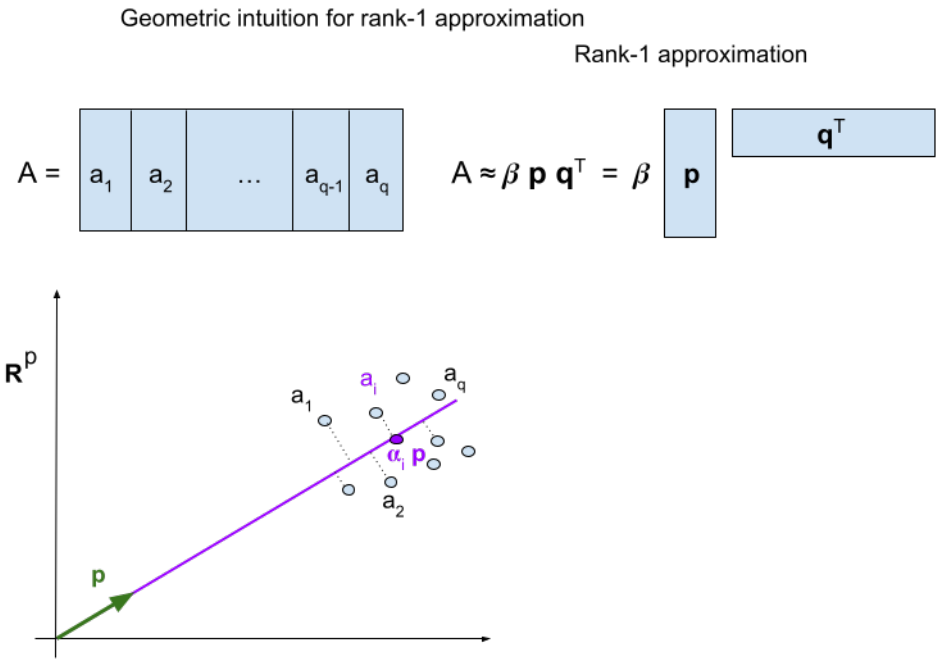


Figure 2.2: A geometric interpretation on how SVD recursively factorise a given matrix A.

- Assuming A is a $p \times q$ matrix, the optimal rank-1 approximation A_1 is of the form:

$$A_1 = \beta \mathbf{p}\mathbf{q}^T$$

where $\mathbf{p}$ and $\mathbf{q}$ are unit column vectors of size p and q respectively.

- By virtue of Theorem 2, the p-dimensional vector $\mathbf{p}$ can be determined uniquely (up to a sign change) by constructing the regression line L passing from the origin towards the p-dimensional points $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$. Denote the orthogonal projection of $\mathbf{a}_i$ onto the regression line by $\alpha_i \mathbf{p}$.
- To determine the corresponding $\mathbf{q}$ we use the fact that the rank-1 matrix A_1 maps the standard unit vectors $\mathbf{e}_i$ to $\alpha_i \mathbf{p}$ (for $i = 1, 2, \ldots, q$). Hence,

$$A_1 \mathbf{e}_i = \beta \mathbf{p} \mathbf{q}^T \mathbf{e}_i = \beta q_i \mathbf{p}$$

  From this it follows that

$$\forall i = 1, 2, \ldots, q: \quad \alpha_i = \beta q_i$$

  and consequently:

$$\beta^2 = \sum_{i=1}^{q} \alpha_i^2 \qquad \text{since} \quad \sum_{i=1}^{q} q_i^2 = 1$$

- From this we can now explicitly determine $\mathbf{q}$:

$$q_i = \frac{\alpha_i}{\beta} = \frac{\alpha_i}{\sqrt{\sum_{i=1}^{q} \alpha_i^2}}$$

Take-home message From the above derivation we conclude that the regression line (through the origin) L essentially determines the rank-1 approximation $A_1 = \beta \mathbf{p} \mathbf{q}^T$:

1. $\mathbf{p}$ is the unit vector along the regression line L; this fixes the coefficients α_i.
2. Singular value: $\beta = \sqrt{\sum \alpha_i^2}$.
3. $\mathbf{q}$ is the unit vector proportional to $(\alpha_1, \alpha_2, \ldots, \alpha_q)$.

This observation has a number of important consequences:

1. The SVD does not solely depend on the shape of the point cloud, but also on its position with respect to the origin. More precisely: if the position of the cloud is large compared to its size (is far from the origin), the first singular vectors and the corresponding singular value (and hence the rank-1 approximation) are determined by its position. However, when we shift it closer to the origin, the singular vector switches adapting to the shape of the cloud. Figure 2.3 provides an illustrative example of how moving further from the origin can drastically affect the first singular vector.
2. The above reasoning would also suggest that movement on the line along with the first singular vector away from the origin will not affect the 2nd singular value/vectors (see Figure 2.4).
3. Figure 2.4 exemplifies why the ratio σ_1/σ_2 does not necessarily tell us something about the shape (and therefore periodicity) of the point cloud (time series).

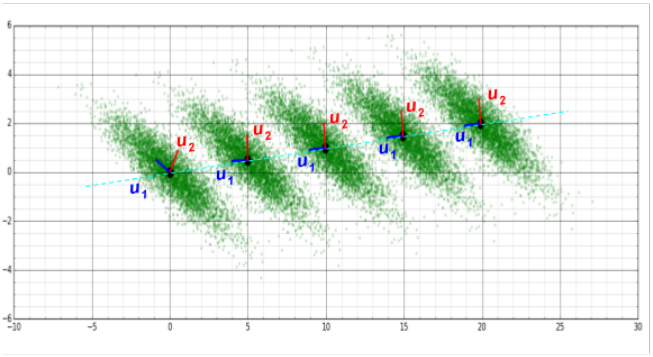


Figure 2.3: A plain example of how the direction of the first (and higher order) singular vectors can change with respect to the distance of the cloud of points from the origin.

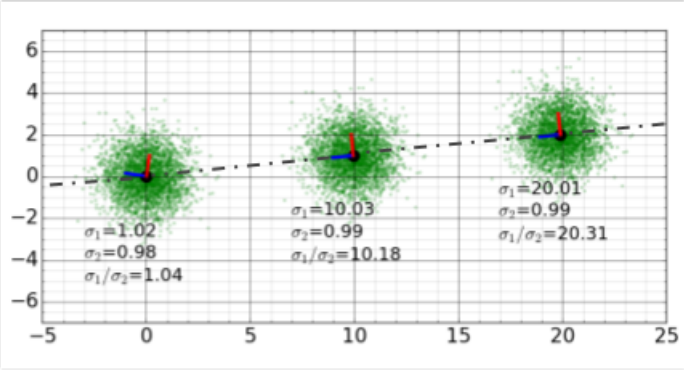


Figure 2.4: Example where the position rather than the shape of the point cloud determines the ratio σ_1/σ_2.

2.3.2. INTUITIVE CONTINUITY-BASED ARGUMENT FOR SVD

When restricting our attention to the 2-dimensional case, it is easy to get a feel as to why SVD holds. It also hints at why the result is less surprising than might seem at first blush. The following is not meant as proof, but simply as an aid to intuition. Consider Figure 2.5, we can make the following observations:

- **Subfig 0:** Consider an ortho-frame, i.e., orthonormal frame (red and green vector) that is mapped under the linear transformation A to the non-orthonormal basis on the RHS of the figure. Let us assume that the angle between the red and green vectors is acute.

- **Subfig 1:** Now rotate the ortho-frame counter-clockwise over 90^o to the new position. Notice that the red vector is now in the same position as the green vector previously. This will also rotate the image of the ortho-frame to a new position.

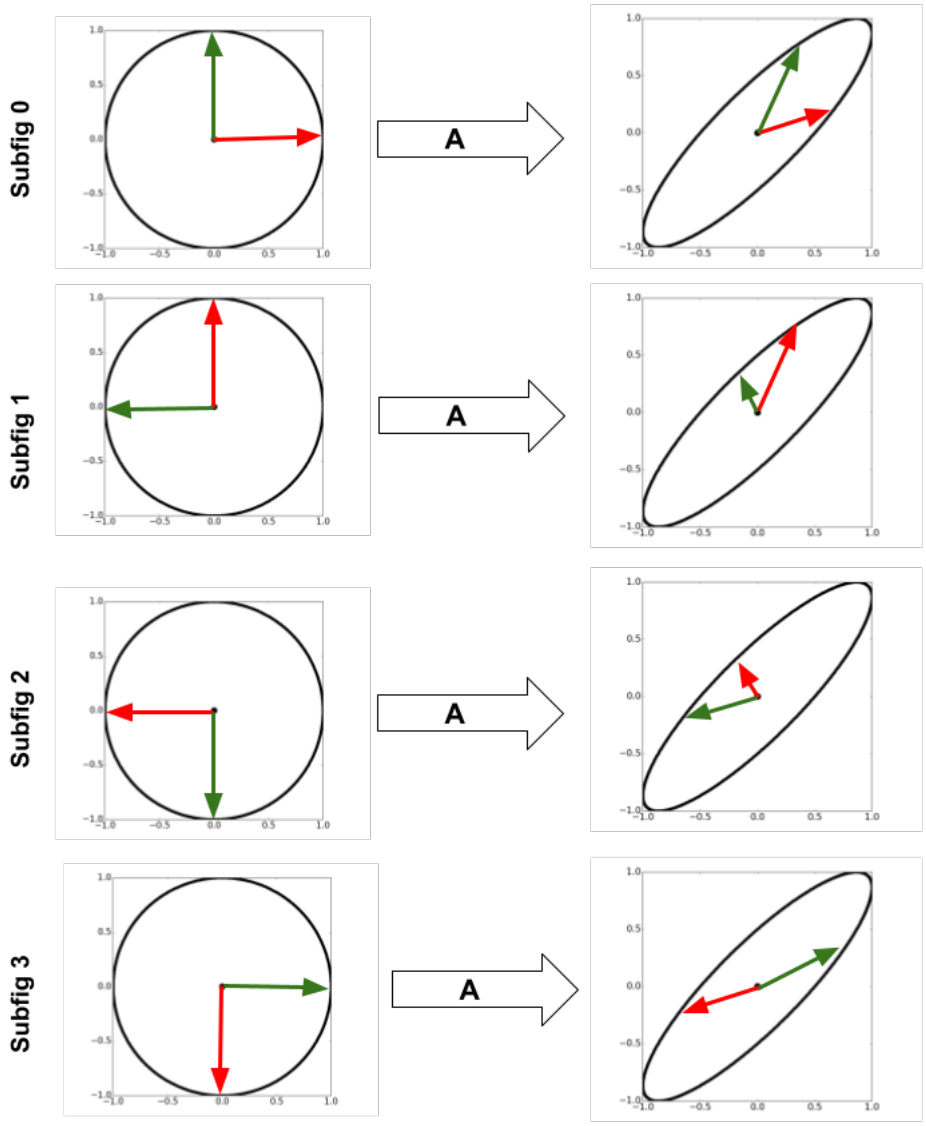


Figure 2.5: For more information, see main text, Section 2.3.2

Let's once again assume that during this rotation the angle in the RHS remains (strictly) less than 90^o (otherwise nothing remains to be proved).

- **Subfig 2:** We continue the rotation and use the same assumptions.

- **Subfig 3:** Executing a third 90^o-rotation now puts the green vector of the ortho-frame in the original position of the red vector, and the red vector in the previous position of the green vector. This means that their images are known. The image for the green vector is known in Subfig 0 and for the red vector in known in Subfig 2. And more importantly, assuming that the angles (between the image vectors) in the previous moves were always less than 90^o, it follows that it now must be in excess of 90^o. Using the argument from continuity, we can conclude that somewhere during the last rotation of the ortho-frame, its image subtended a right angle, exactly what the SVD theorem implies.

2.4. DATA ALIGNMENT

One of the common pitfalls of the SVD is associated with misaligned data. Figures 2.7-2.8 highlight the fact that SVD is *geometric*, meaning that it depends on the coordinate system in which the data is represented. In other words, SVD is only generically invariant to unitary transformations where only the inner products are preserved (see Figure 2.6). This fact may be viewed as the reason for or against this method. First, the inner product at the core of such matrix decomposition technique is essential for various insightful geometric interpretations. Furthermore, the results of the SVD contain meaningful units and dimensions. On the negative side, the SVD is liable to the alignment of the data. In fact, the SVD rank of the matrix inflates drastically when "objects" in matrix data (a certain pattern in data) translate, rotate, or scale, which severely constrains its use for the cases where data has not been heavily pre-processed. In other words, in a given set of coordinates, the SVD is unable to capture translations and rotations of the data.

2.5. COMPARING SVD AND PCA

Principal component analysis (PCA), also known as the Karhunen-Loève transform, is a popular matrix decomposition technique that is used in diverse applications such as dimensionality reduction, lossy data compression, feature extraction, and data visualization [7].

There are two commonly used definitions of PCA that lead to the same algorithm. PCA can be defined as the orthogonal projection of the data points onto a lower dimensional linear space, known as the *principal subspace*, in such a way that the variance of the projected data is maximized [8]. On a similar note, PCA can be defined as the linear projection that minimizes the average projection cost, defined as the mean squared distance between the data points and their projections [9].

2.5.1. PCA

We herein consider the latter definition of PCA mentioned above, and investigate how it relates to the SVD.

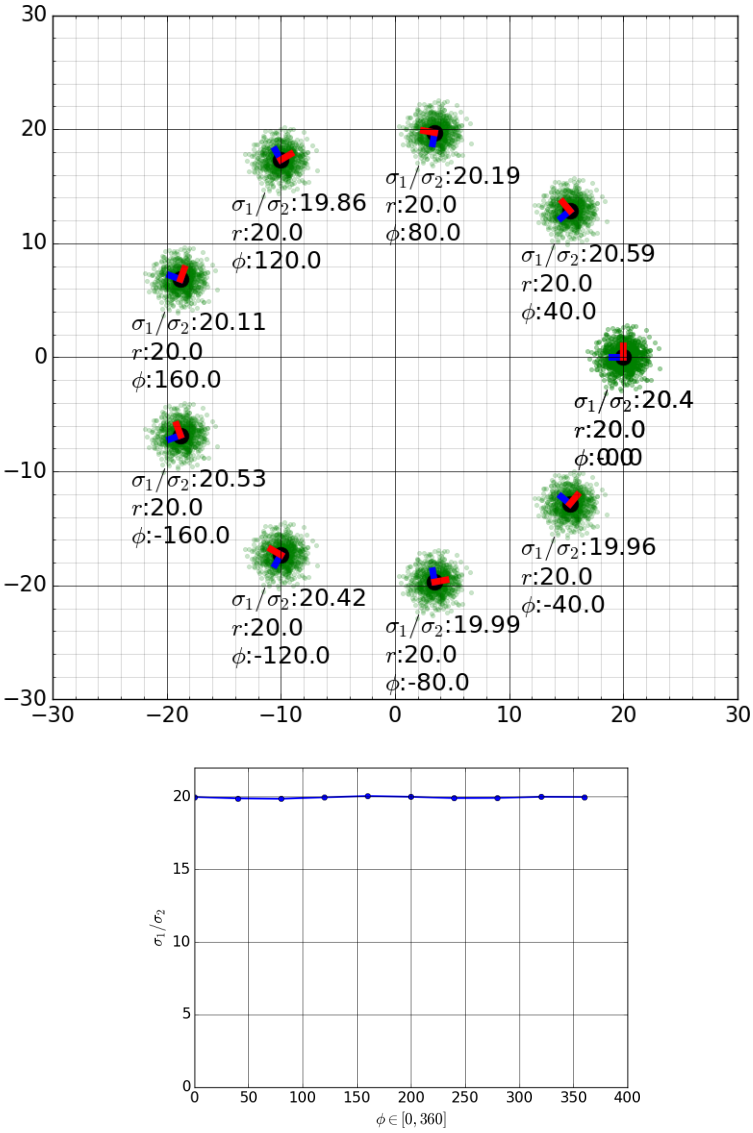


Figure 2.6: **Top:** The singular value ratio stays relatively the same if the distance from the origin is the same, regardless of the position of the cloud of points with respect to the origin (**bottom**).

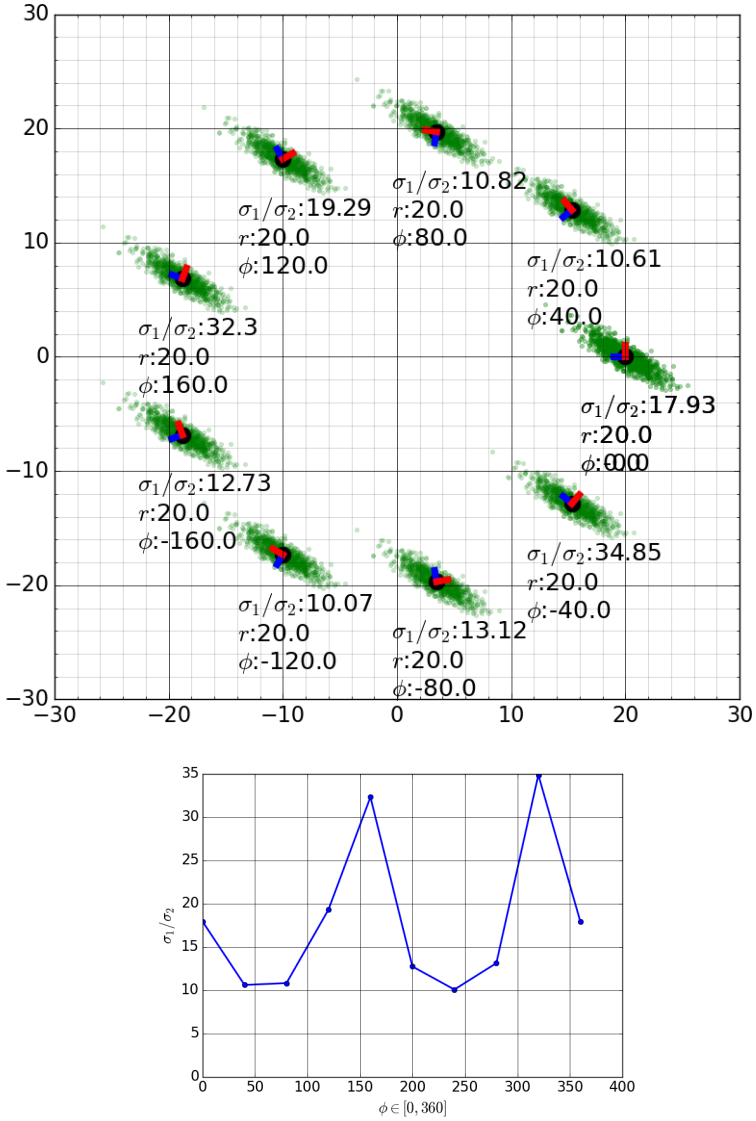


Figure 2.7: The evolution of σ_1 and σ_2 (**top**) and their corresponding ratio (**bottom**) while the cloud of points moving on a circle around the origin. As expected, both singular vectors change depending on the alignment and the position of the cloud with respect to the origin. The black vector is the direction of the first singular vector and the yellow one is the direction of the second singular vector.

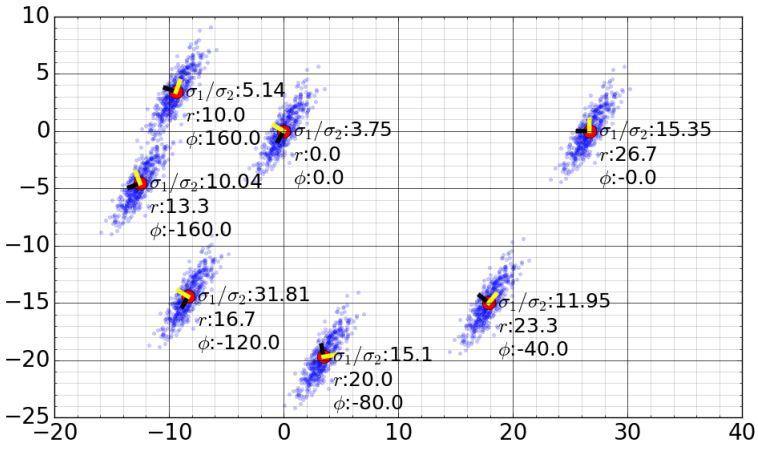


Figure 2.8: A representation on how alignment of the could and its distance from the origin affects the singular value ratio. In this figure cloud with $\phi = -120^o$ has the highest σ_1/σ_2 ratio.

Theorem 3 (PCA-type factorisation). *Assume that a $p \times q$ matrix A has rank $rk(A) = r \leq \min(p,q)$. We now define the functional $G(P,Q)$ as follow:*

$$G(P,Q) = \|A - PQ^T\|^2 \tag{2.12}$$

and the corresponding constrained optimisation problem:

$$\min_{P,Q} G(P,Q) \qquad \textit{subject to} \quad rk(P) = rk(Q) = k \quad \textit{and} \quad Q^TQ = I_k \tag{2.13}$$

where $k < r$. A solution to the above constrained minimisation problem (in $P \in \mathbb{R}^{p\times k}$ and $Q \in \mathbb{R}^{q\times k}$) is given by (using the SVD notation given in Eq. (2.10)):

$$Q = V_{(1:k)} \quad \textit{and} \quad P = U_{(1:k)}\, diag(\sigma_1,\ldots,\sigma_k) \tag{2.14}$$

hence:

$$PQ^T = \sum_{i=1}^{k} \sigma_i U_i V_i^T \tag{2.15}$$

From Eq. (2.14) this it also follows that P^TP is diagonal, but not necessarily equal to the identity. □

Note that if we drop the insistence on the diagonal form for P^TP (i.e., P need no longer be an orthogonal frame), then the solution is no longer unique. Indeed, by taking any $k \times k$ orthogonal matrix R with $R^TR = I_k = RR^T$, it is clear that $P' = PR$ and $Q' = QR$ are also solutions. In this case: $Q'^TQ' = R^TQ^TQR = I_k$ but $P'^TP' = R^TP^TPR = R^T(SS^T)R$ is in general a positive definite symmetric matrix.

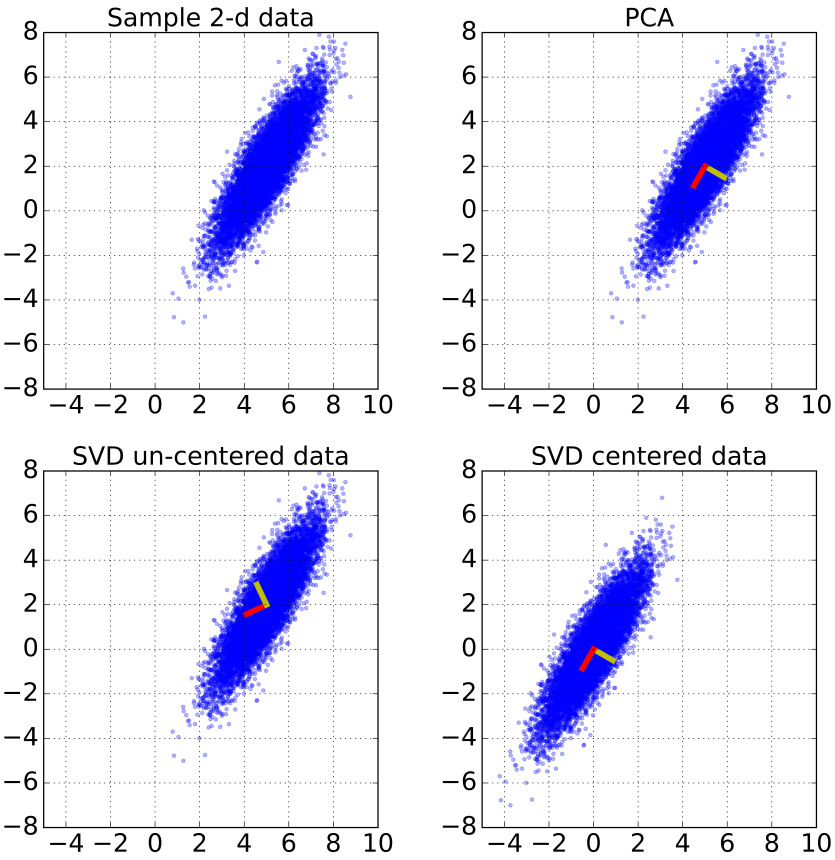


Figure 2.9: Confirmation that PCA and the SVD of the mean-subtracted data are the same (up to a sign change).

2.5.2. SVD vs. PCA

In a sense, SVD is equivalent to PCA in multivariate statistics, but in addition, is used to generate low-dimensional representations for complex multidimensional time series. SVD and PCA of a given matrix are related to one another through the covariance matrix. In other words, if $X = USV^T$ then for the covariance matrix we can write:

$$X^T X = VSU^T USV^T = VS^2V^T$$

now, if we multiply both sides by V, it yields:

$$(X^T X)V = VS^2$$

where V is the matrix of the eigenvectors of the matrix $X^T X$ and S^2 is the diagonal matrix of the eigenvalues.

- If the actual position of the point cloud matters, SVD is the better option; e.g., the first singular vector is usually close to the centre of gravity;

- When only the shape of the data cloud is of importance, then PCA is more appropriate.

Figure 2.9 implies the above-mentioned results. Shifting the data to the origin does not change how the data points are positioned *relative* to each other. That is why the results of PCA are unbiased with respect to the mean of the matrix.

2.6. ADDENDUM: ADDRESSING COMPUTATIONAL ARTEFACTS

In this thesis, we make extensive use of simulation and SVD-based computations. As a consequence, accurate sampling from various matrix distributions is important. Through comprehensive experiments, we have become aware of certain biases and artefacts that are present in the implementations of the algorithms in Matlab and Python. Unless these artefacts are remedied, they might introduce biases that invalidate ensuing results and conclusions.

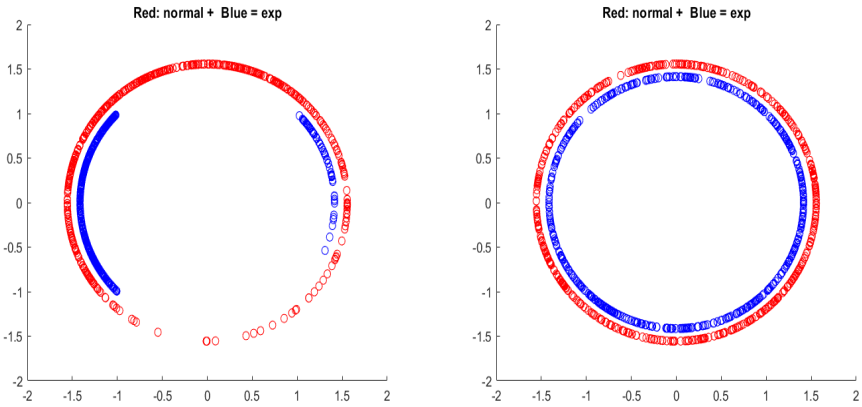


Figure 2.10: **Left:** Results (sum of the U vectors) from the original computations; **Right:** same results but with random signs.

2.6.1. SVD COMPUTATIONS IN MATLAB

Some of the unexpected results might be the consequence of computational artefacts. From the numerical evidence we have seen, it looks like the computational algorithms have a bias towards the selection of certain singular vectors. This is illustrated in the following experiment: Generate a random 2×2 matrix A according to the normal and exponential distribution. Since the columns of U (obtained by the SVD) form an orthonormal basis in the plane we can easily depict the result graphically. Rather than showing both vectors, we compute the sum as a 1-vector representation; all of these vectors are on the $\sqrt{2}$ circle (see Figure 2.10). The results for the normal distribution are shown in red, and for the exponential in blue. The raw output of the algorithm is shown on the left panel of the figure. It shows clear artefacts in the distribution (especially for the exponential results). If we randomize the U-vectors by assigning a random sign to them (i.e., ran-

domly flipping them over) the results change dramatically and agree with expectations. Similar results can be obtained in Python. Another manifestation of this is if we generate a random normal matrix A, and do the SVD $A = USV^T$ then $\det U = +1$ always while $\det V = \pm 1$ with equal probabilities.

2.6.2. SAMPLING RANDOM ORTHOGONAL MATRICES: GRAM-SCHMIDT ORTHOGONALISATION (QR DECOMPOSITION)

As pointed out earlier in Section 2.2.1, the definition of the singular vectors is determined up to a simultaneous sign change in the corresponding left- and right vector: $(U_\ell, V_\ell) \to (-U_\ell, -V_\ell)$. This indeterminacy is exploited in various algorithms to assign a preferred direction to singular vectors. One manifestation of this is apparent in various QR decomposition algorithms (in MATLAB, and also in python). Specifically: applying Gram-Schmidt orthogonolisation or using the QR decomposition in Matlab always produces frames with right-handed chirality (i.e., the determinant $\det Q = +1$).

REFERENCES

[1] A. Khoshrou and E. J. Pauwels, *Propagating uncertainty in tree-based load forecasts,* in *Electrical and Electronics Engineering (ELECO), 2017 10th International Conference* (IEEE, 2017) pp. 120–124.

[2] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, *Machine learning for fluid mechanics,* Annual Review of Fluid Mechanics **52**, 477 (2020).

[3] G. Strang, *Introduction to linear algebra*, Vol. 3 (Wellesley-Cambridge Press Wellesley, MA, 1993).

[4] R. Horn and C. Johnson, *Matrix Analysis* (Cambridge University Press, 1985).

[5] G. H. Golub and C. F. Van Loan, *Matrix computations*, Vol. 3 (JHU press, 2013).

[6] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank,* Psychometrika **1**, 211 (1936).

[7] C. M. Bishop, *Pattern recognition and machine learning* (springer, 2006).

[8] H. Hotelling, *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology **24**, 417 (1933).

[9] K. Pearson, *Liii. on lines and planes of closest fit to systems of points in space,* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**, 559 (1901).

3

PROPERTIES OF THE SVD

3.1. INTRODUCTION

In the previous chapters we argued that SVD can be used to estimate periodicity in time series, as well as construct appropriate approximations for time series. The latter is based on the assumption that a data matrix A can be decomposed as

$$A = A_0 + A_r$$

where A_0 is a low rank matrix that comprises all the useful signal information, whereas the remainder term A_r collects the (relatively small) noise. For most of the applications of SVD in various fields, it is important to understand the properties of the SVD of a matrix whose entries show some degree of random fluctuations. Therefore, in order to determine what the impact of noise will be on the singular value spectrum, it is useful to study the singular value decomposition of pure noise matrices, i.e., random matrices. This is the topic addressed in the following sections. Specifically, we address the following topics:

1. The impact of the matrix aspect ratio (number of rows versus number of columns);

2. The impact of shifts of the mean of the underlying distribution as this will become relevant in applications where a signal will in general have a non-zero mean.

3.2. SINGULAR VALUE SPECTRUM OF RANDOM MATRICES

In order to disentangle the impact of signal and noise, we first focus on the effect of pure noise (i.e., random matrices). The spectral study of random matrices (i.e., matrices for which the entries are independent, identically distributed (i.i.d.) random variables) has been a very active research domain in recent years and uncovered a number of key insights (see e.g., [2–4]).

Parts of this chapter have been published in [1].

3.2.1. PRELIMINARIES

Let N be a random $p \times q$ matrix such that the individual entries are i.i.d, random variables with zero mean and unit variance:

$$\mathbb{E}(N_{ij}) = 0 \quad \text{and} \quad Var(N_{ij}) = 1$$

As previously mentioned in Section 2.2.2, to compute the singular values of the given matrix N, we are interested in the eigenvalues of either one of the following two symmetric, quadratic matrices:

$$Q^{(q)} = N^T N \in \mathbb{R}^{q \times q} \qquad \text{or} \qquad Q^{(p)} = NN^T \in \mathbb{R}^{p \times p}$$

Denoting by N^i and N_i the i-th row or column of N, respectively, we observe that the elements of $Q^{(q)}$ and $Q^{(p)}$ can be expressed as inner products of either rows and columns of N; i.e., the inner product of columns:

$$Q^{(q)}_{ij} = N_i \cdot N_j = \begin{cases} \sum_{k=1}^{p} N_{ki} N_{kj} & i \neq j \\ \sum_{k=1}^{p} N_{ki}^2 & i = j \end{cases} \tag{3.1}$$

and similarly (inner product of rows),

$$Q^{(p)}_{ij} = N^i \cdot N^j = \begin{cases} \sum_{k=1}^{q} N_{ik} N_{jk} & i \neq j \\ \sum_{k=1}^{q} N_{ik}^2 & i = j \end{cases} \tag{3.2}$$

Since we are typically interested in what happens when we gather more data (i.e., $q \to \infty$), we will focus mostly on $Q^{(p)}$ as there are only p non-zero singular values, which correspond to the p (square roots of the) eigenvalues of $Q^{(p)}$. From here on, we will drop the superscripts for the Q-matrix as it will be clear from the context which one we are using.

It then is straightforward to compute the first moments[1]:

$$\mathbb{E}(Q_{ij}) = \begin{cases} \sum_{k=1}^{q} \mathbb{E}(N_{ik})\,\mathbb{E}(N_{jk}) = 0 & i \neq j \\ \sum_{k=1}^{q} \mathbb{E}(N_{ik}^2) = q & i = j \end{cases} \tag{3.3}$$

and similarly:

$$Var(Q_{ij}) = \begin{cases} \sum_{k=1}^{q} Var(N_{ik})\, Var(N_{jk}) = q & i \neq j \\ \sum_{k=1}^{q} Var(N_{ik}^2) = \beta^2 q & i = j \end{cases} \tag{3.4}$$

[1] Here for $Q = Q^{(p)}$, similar results hold for $Q^{(q)}$.

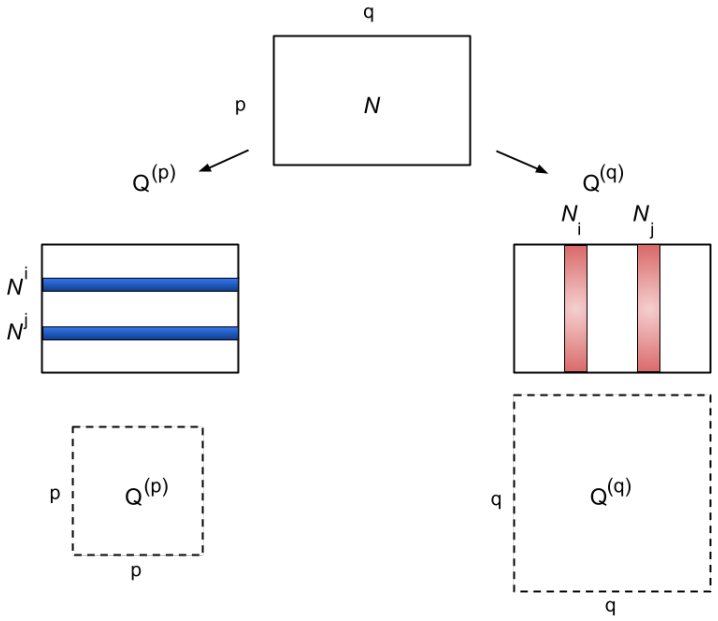


Figure 3.1: The intuition of why the singular values of N are equal to the eigenvalues of Q.

where $\beta^2 := Var(N_{ik}^2)$ is a common value since all variables are i.i.d.[2]

The above results can be recast in more compact matrix notation:

$$\mathbb{E}(Q) = qI_p \qquad \text{and} \qquad Var(Q) = q\left((\beta^2-1)I_p + \mathbf{1}_{p\times p}\right) \tag{3.5}$$

Normalised version Notice that the above results show that the normalised matrix

$$Q_n := \frac{1}{q}Q$$

has constant moments:

$$\mathbb{E}(Q_n) = I_p \qquad \text{and} \qquad Var(Q_n) = (\beta^2-1)I_p + \mathbf{1}_{p\times p} \tag{3.6}$$

3.2.2. UNIVERSALITY

From Eqs. (3.1) and (3.2), it is evident that the elements of the Q-matrices are sums of independent i.i.d. random variables. In particular, each Q_{ij} is the sum of q **independent** terms and it therefore asymptotically converges to a **normal distribution** (based on the Central Limit Theorem [5]). We hence can conclude that irrespective of the initial distribution of the N-entries, the Q-entries will converge to a normal distribution, i.e., the individual entries Q_{ij} will "forget" the original distribution when $q \to \infty$. In other words,

[2] As mentioned before, the entries of N need to be zero mean and unit variance for the results to hold.

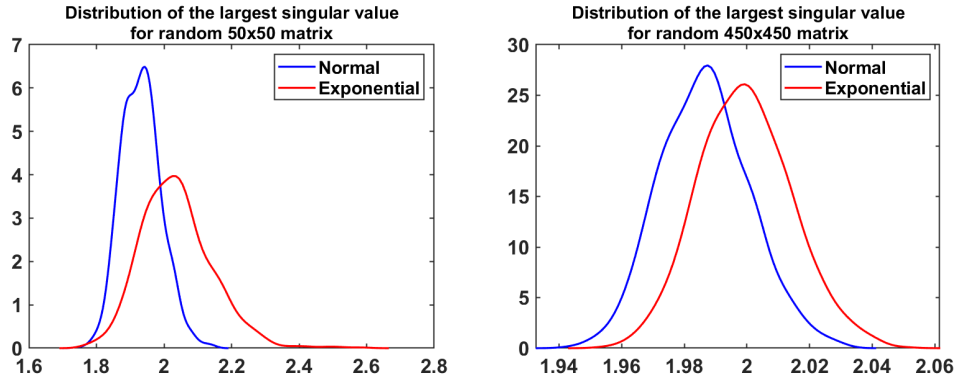


Figure 3.2: Comparison of the distribution of the highest singular values of two random matrices: normal and exponential distributions. According to the universality theorem, as the dimension grows, these distributions become closer to each other (for the sake of comparison $\sigma_1/\sqrt{q}$ has been considered, where q is the size of the square matrix. The distributions are the results of 200 iterations.).

since the singular values of N are determined by the eigenvalues of Q it follows that, as $q \to \infty$, the distribution of the singular values becomes independent of the distribution of N-matrix (apart from the first two moments). This is called the **universality property**. Figure 3.2 provides a comparison between the distribution of the first singular value of the two random matrices, one a random normal and the other random exponential. As can be seen in the case of the 50×50 matrix (the left panel) the difference between the two distributions is discernible. However, as the dimensionality grows (for the 450×450 matrices of the right panel) the distributions of the first singular values for two matrices (random normal and exponential matrices) become more similar.

In a similar way, Figure 3.3 illustrates a comparison of the singular values (averaged over 200 trials) of 50×50 random matrices for two different distributions of the individual matrix entries: standard normal and exponential (shifted to become zero-mean). These two figures combined show that as long as the mean and variance of the noise is kept constant, its actual distribution has very little influence on the distribution of the resulting singular values, assuming the size of the matrix is not too small.

In addition to the above result, we also know that rescaling the variance of the entries in a zero-mean random matrix induces the corresponding rescaling of the singular values: $\sigma_i(\alpha A) = \alpha \sigma_i(A)$. This follows immediately from the observation that $\alpha A = U(\alpha S)V^T$. In other words, the singular value ratio $SVR = \sigma_1/\sigma_2$ is not affected by a uniform increase in the noise variance. However, a shift in the mean of the noise does affect the SVR, as will be explained in the following sections.

3.3. IMPACT OF ASPECT RATIO ON SINGULAR VALUES

3.3.1. PROBLEMS WITH THE SVR APPROACH

As mentioned before, Kanjilal et. al. [6] use the singular value ratio (SVR) spectrum to find the periodicity p in the time series and consequently decompose a signal into

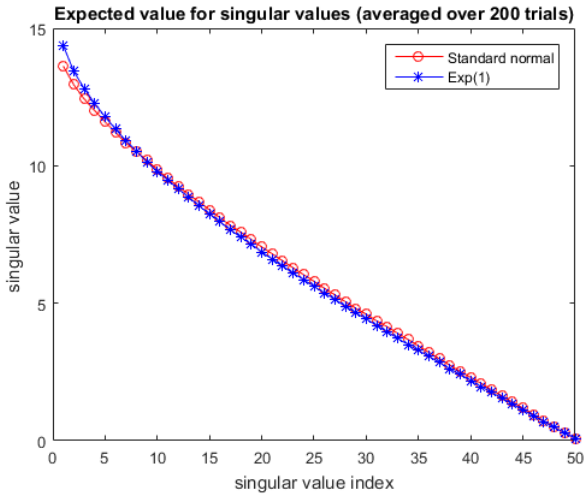


Figure 3.3: Singular values (averaged over 200 trials) for 50×50 random matrices generated by drawing i.i.d. entries from the standard normal (red) and (shifted to ensure zero mean and unit variance) exponential (blue) distributions.

its constituent periodic components. This idea of period estimation based on the SVR seems intuitive and straightforward. However, there are some complications that need to be addressed as they could bias and invalidate the results. Specifically:

- The SVR depends on the average value of the signal
- The SVR depends on the aspect ratio of the data matrix

We will illustrate these problems below.

Impact of average value In the original paper [6], it was not sufficiently appreciated how a shift in the mean value of the time series (the DC component which is the mean amplitude of the waveform) impacts the SVR. This is important as failure to understand this issue introduces a major bias in the test values and could therefore result in erroneous conclusions. The left panel of Figure 3.4 provides an example of how the mean of a signal can affect its singular value distributions. Clearly, failing to remove the mean from a noisy time series would inflate the first singular value (and only the first one) resulting in an upwardly biased value for the singular value ratio (SVR). Such a blind screening would reduce the power of an SVD method in data mining applications.

3.3.2. MOTIVATION

As mentioned before, we recast a given periodic signal in such a way that each column represents a single period. Adding more observations amounts to adding more columns, which in turn makes the corresponding matrix “fatter". This has an impact on its singular values. This is obvious from the fact that the L_2 norm (or Frobenius norm) of a matrix

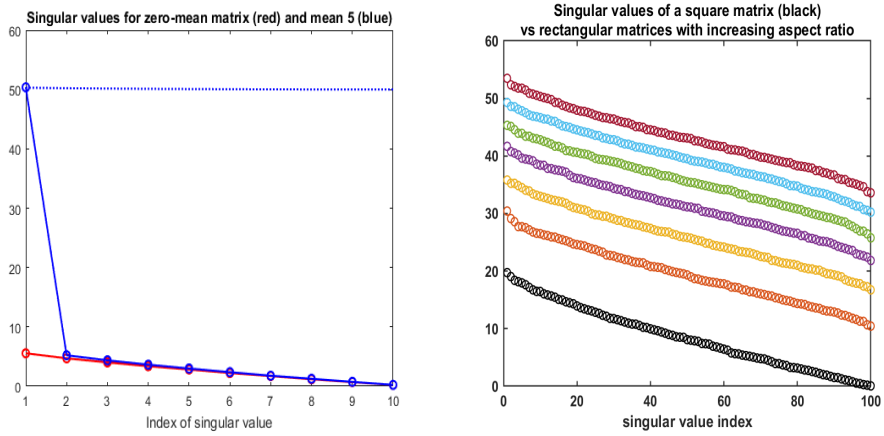


Figure 3.4: **Left:** Comparison of the singular values of a matrix (10×10) with zero-mean entries (red) and shifted mean ($\alpha = 5$). The dotted line indicates that (approximate) upper limit based on Eq. (3.19). Recall that the entries of the matrix A_0 are random numbers, but by shifting the global mean the $SVR = \sigma_1/\sigma_2$ increases, erroneously suggesting that some underlying periodic structure is present. **Right:** By proceeding from an square to rectangular matrix, all the singular values increase proportionately.

can be expressed in terms of the sum of its singular values, i.e., (see Appendix A.3):

$$\|A\|_F^2 := \sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}^2 = \sum_{k=1}^{r} \sigma_k^2$$

where A_{ij} is the entry of the matrix. If we assume the matrix to be fat ($q > p$) and full rank (i.e., $r := rk(A) = p$), it follows that increasing the number of columns q must also inflate the quadratic sum of singular values, without increasing the number of terms in the sum, which is fixed by the rank. This can happen in roughly two ways:

1. Multiplicative: all the singular values are multiplied by a factor $\alpha > 1$, i.e.,

$$\sigma_i' = \alpha(q, i)\ \sigma_i$$

2. Additive: all the singular values are shifted by a fixed amount $\alpha > 0$, i.e.,

$$\sigma_i' = \sigma_i + \alpha(q, i)$$

Notice that the first possibility would not significantly impact the SVR, whereas the second one will. Numerical experiments show that the actual mechanism at work is more akin to the second possibility: “fattening" a noise matrix shifts all the singular values upward (see Figure 3.4). To obtain crisp results we return to the case of purely random matrices. For a more detailed explanation of the underlying reasoning behind Figure 3.4, see Appendix A.6. The next section provides an explanation for this observation.

3.3.3. ASYMPTOTIC RATIO OF SINGULAR VALUES AS FUNCTION OF GROWING ASPECT RATIO

The following section investigates the evolution of the distribution of the singular values of random matrices where the ratio of the number of columns (the number of observations) to the number of rows (fixed number of dimensions of the data) increases over time.

Theorem 4. *Let A be a (full rank) random $p \times q$ matrix (where $q \geq p$) such that A_{ij} are i.i.d, $\mathbb{E}(A_{ij}) = 0$ and $Var(A_{ij}) = 1$. The singular value decomposition theorem ensures that there are (essentially unique) matrices $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{q\times q}$, both with orthonormal columns, (i.e., $U^T U = I_p$, $V^T V = I_q$). We denote the SVD decomposition as:*

$$A = USV^T \quad where \quad S = \mathbb{R}^{p\times q}$$

such that

$$S = (\Sigma_{p\times p}, 0_{p\times(q-p)}) \quad and \quad \Sigma_{p\times p} = diag(\sigma_1, \sigma_2, \ldots, \sigma_p),$$

and the singular values have been ordered in descending order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$. Then, if q (the number of columns) tends to infinity, while keeping the number of rows p constant, all the singular values σ_i will increase to infinity in such a way that the ratio of the largest to the smallest singular value tends to 1:

$$\lim_{q\to\infty} \frac{\sigma_{max}}{\sigma_{min}} = \lim_{q\to\infty} \frac{\sigma_1}{\sigma_p} = 1$$

Proof. As previously mentioned in Section 2.2.2, singular values and vectors for a matrix A are actually the eigenvalues and eigenvectors of the quadratic matrices AA^T and $A^T A$; we, therefore, focus on the quadratic random matrices of the form

$$Q = AA^T \in \mathbb{R}^{p\times p} \quad \text{where} \quad A \in \mathbb{R}^{p\times q}, \quad A_{ij} : i.i.d., \quad \mathbb{E}(A_{ij}) = 0, \quad \text{and} \quad Var(A_{ij}) = 1$$

Harking back to Eq. (3.5) we had:

$$\mathbb{E}(Q) = qI_p \qquad \text{and} \qquad Var(Q) = q\left((\beta^2 - 1)I_p + \mathbf{1}_{p\times p}\right)$$

where $\beta^2 := Var(A_{ik}^2)$ is a common value since all variables are i.i.d.

According to the CLT, we know that asymptotically (as $q \to \infty$) the entries Q_{ij} will be normally distributed. Therefore, we can use the above moment information to specify the following approximation (asymptotically, as $q \to \infty$):

- **Diagonal:** $diag(Q) \approx qI_p + \beta\sqrt{q}Z_p$, where Z_p is a $p \times p$ diagonal matrix with independent, standard normal random variables on the diagonal.
- **Off-Diagonal:** The off-diagonal part is approximated by the symmetric (zero-diagonal) matrix $\sqrt{q}M$ where

$$M_{ij} = M_{ji} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

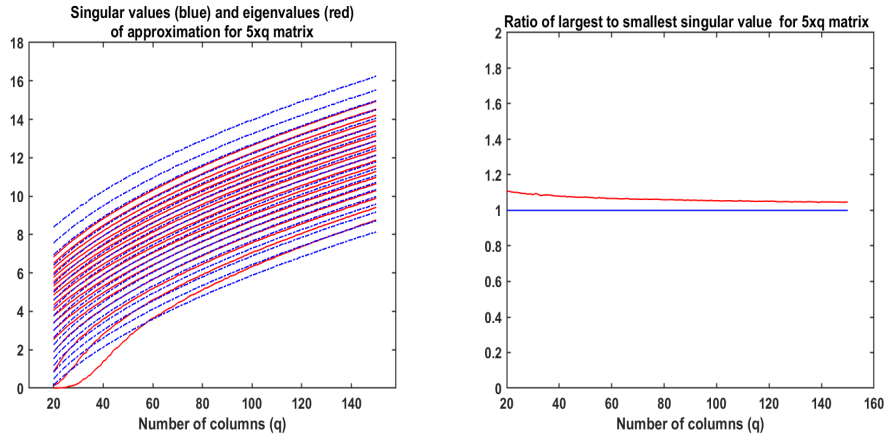


Figure 3.5: **Left:** Illustrating the approximation of the singular values of A using (the square roots of) the eigenvalues of AA^T. **Right:** Ratio of largest to smallest singular value.

Collecting all these we arrive at the following asymptotic approximation:

$$AA^T \approx qI_p + \beta\sqrt{q}\,Z_p + \sqrt{q}\,M = q\left(I_p + \frac{1}{\sqrt{q}}(\beta Z_p + M)\right) \tag{3.7}$$

Notice that $K := \beta Z_p + M$ does not depend on q. Furthermore, it is symmetric and therefore can be diagonalized.

As a consequence, K has an orthonormal basis of eigenvectors $\mathbf{k}_i$ ($i = 1, 2, \ldots, p$) and corresponding eigenvalues ν_i such that $K\mathbf{k}_i = \nu_i \mathbf{k}_i$. Collecting these eigenvectors and eigenvalues in matrices P and N respectively, we can conclude:

$$K = PNP^T \quad \text{where} \quad P \text{ is orthogonal, i.e., } PP^T = P^T P = I_p$$

Plugging this decomposition in the RHS of Eq. (3.7) we see that:

$$AA^T \approx R(q) := q(I_p + \frac{1}{\sqrt{q}}K) = q\,P\left(I_p + \frac{1}{\sqrt{q}}N\right)P^T$$

from which we can conclude that $R(q)$ has the same eigenvectors as Q (i.e., the columns of P), but the corresponding eigenvalues are given by:

$$\lambda_i(R_q) = q\left(1 + \frac{\nu_i}{\sqrt{q}}\right)$$

Notice that the eigenvalues ν_i are independent of q and therefore:

$$\lim_{q\to\infty} \frac{\lambda_i(R_q)}{\lambda_j(R_q)} = \lim_{q\to\infty} \frac{1 + \nu_i/\sqrt{q}}{1 + \nu_j/\sqrt{q}} = 1 \tag{3.8}$$

Furthermore, for q sufficiently large:

$$\sigma_i(A) \approx \sqrt{\lambda_i(R_q)} = \sqrt{q + \sqrt{q}\nu_i} \tag{3.9}$$

as can be seen in Figure 3.6. □

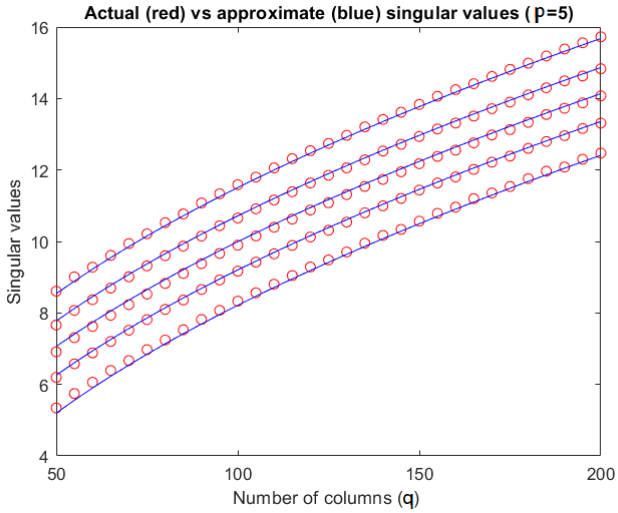


Figure 3.6: Comparison of actual singular values (red) and approximation (in blue) specified in Eq. (3.9).

3.4. IMPACT OF THE UNDERLYING PERIODIC SIGNAL

Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ represents a noisy but perfectly stationary and periodic time series with period p. We can then use the methodology explained in Section 1.2 to recast such a time series as a matrix A with size $p \times q$.

As mentioned in Section 1.2 for a pure rank-1 matrix $A_0 = \mathbf{a}\mathbf{1}_q^T$ the SVD decomposition is straightforward; all we need to do is to reduce the vectors to unit vectors:

$$A_0 = \mathbf{a}\mathbf{1}_q^T = a\sqrt{q}\left(\frac{\mathbf{a}}{a}\right)\left(\frac{\mathbf{1}_q^T}{\sqrt{q}}\right) = \sigma_1 \mathbf{u}\mathbf{v}^T \quad (a = \|\mathbf{a}\|) \tag{3.10}$$

confirming that the first (and only non-zero) singular value equals $\sigma_1(A_0) = a\sqrt{q}$.

In general, however, the data is noisy and we model that by adding independent additive noise with variance ε^2 in below:

$$A = \mathbf{a}\mathbf{1}_q^T + \varepsilon N \tag{3.11}$$

Here, similar to Eq. (1.6), N is a $p \times q$ matrix of independent, identically distributed (i.i.d.) noise variables with zero mean and unit variance. Notice that the p-dim column space in Eq. (3.11) can be interpreted as a zero-mean random cloud of q columns of matrix N, each of which is shifted by $\mathbf{a}$. The geometric intuition expounded in Section 2.3, therefore, suggests that the square of the first singular value should be shifted by $q\|\mathbf{a}\|^2$. We will now show that this intuition is indeed correct.

To investigate the behaviour of the singular values we use the fact that:

$$\begin{aligned}
\sigma^2(A) &= \lambda(A^T A) \\
&= \lambda\left((\mathbf{a}\mathbf{1}_q^T + \varepsilon N)^T(\mathbf{a}\mathbf{1}_q^T + \varepsilon N)\right) \\
&= \lambda\left(a^2\mathbf{1}_q\mathbf{1}_q^T + \varepsilon(N^T\mathbf{a}\mathbf{1}_q^T + \mathbf{1}_q\mathbf{a}^T N) + \varepsilon^2 N^T N\right)
\end{aligned}$$

where $a^2 = \mathbf{a}^T\mathbf{a} = \|\mathbf{a}\|^2$, $\sigma(.)$ is the singular value and $\lambda(.)$ indicates the eigenvalue. Since the entries of the noise matrix N are independent, zero-mean and unit variance stochastic variables, we can draw on Eqs. (3.3)–(3.5) to make the following approximation for the $q \times q$ matrix $N^T N$:

$$\mathbb{E}(N^T N)_{ij} = \sum_{k=1}^{p} \mathbb{E}\left(N_{ki}N_{kj}\right) = \begin{cases} \sum_{k=1}^{p} \mathbb{E}(N_{ki})\mathbb{E}(N_{kj}) &= 0 \quad \text{if } i \neq j \\ \sum_{k=1}^{p} \mathbb{E}(N_{ki}^2) &= p \quad \text{if } i = j \end{cases}$$

The last approximation is obtained by taking the expected values and using the fact that $\mathbb{E}(N_{ki}N_{kj}) = 1$ if $i = j$, and zero otherwise. From this, we conclude that approximately:

$$N^T N \approx pI_q$$

Similarly, because the expectation value of the cross-term vanishes, using the linearity of the expectation operator yields:

$$\mathbb{E}(N^T\mathbf{a}\mathbf{1}_q^T + \mathbf{1}_q\mathbf{a}^T N) = \mathbb{E}(N^T)\mathbf{a}\mathbf{1}_q^T + \mathbf{1}_q\mathbf{a}^T\mathbb{E}(N) = 0$$

whereas

$$Var\left(N^T\mathbf{a}\mathbf{1}_q^T + \mathbf{1}_q\mathbf{a}^T N\right) = 2a^2 I_q \tag{3.12}$$

As a consequence, to a good approximation, the singular values of A can be identified as the eigenvalues of the following matrix:

$$\sigma^2(A) \approx \lambda(a^2\mathbf{1}_q\mathbf{1}_q^T + \varepsilon^2 pI_q)$$

The structure of the matrix in the RHS allows us to arrive at some conclusions regarding the singular values.

Lemma 5. *The symmetric $q \times q$ matrix $B = a^2\mathbf{1}_q\mathbf{1}_q^T + \varepsilon^2 pI_q$ has a complete set of q orthogonal eigenvectors and corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$ (sorted in descending order):*

eigenvalue	*eigenvector*
$\lambda_1 = a^2 q + \varepsilon^2 p$	$\mathbf{1}_q$
$\lambda_i = \varepsilon^2 p \quad (i \geq 2)$	$\mathbf{e}_i - \mathbf{e}_1$

where $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ are the standard basis vectors.

Proof. The above results follow from a straightforward calculation:

$$B\mathbf{1}_q = a^2\mathbf{1}_q(\mathbf{1}_q^T\mathbf{1}_q) + \varepsilon^2 pI_q\mathbf{1}_q = a^2 q\mathbf{1}_q + \varepsilon^2 p\mathbf{1}_q = (a^2 q + \varepsilon^2 p)\mathbf{1}_q = \lambda_1\mathbf{1}_q$$

Similarly,

$$\begin{aligned} B(\mathbf{e}_i - \mathbf{e}_1) &= a^2\mathbf{1}_q\mathbf{1}_q^T(\mathbf{e}_i - \mathbf{e}_1) + \varepsilon^2 pI_q(\mathbf{e}_i - \mathbf{e}_1) \\ &= 0 \quad + \quad \varepsilon^2 p(\mathbf{e}_i - \mathbf{e}_1) \end{aligned}$$

□

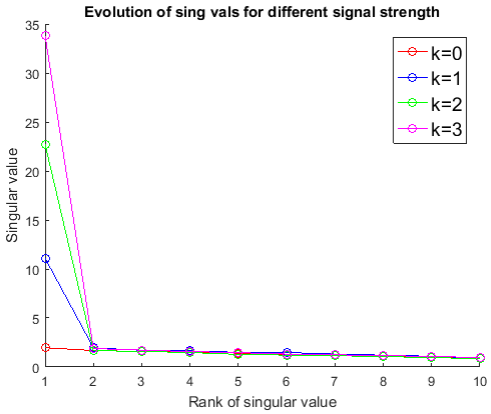


Figure 3.7: The influence of the underlying signal strength on the first singular value. The curve for $k = 0$ corresponds to pure noise (no underlying signal). Notice how increasing the signal strength results in the corresponding increments in the first singular value.

From the lemma above, we can immediately conclude that to a first approximation the following (approximate) result for the singular values of A holds.

Theorem 6. *Let us consider again the $p \times q$ matrix A in Eq. (3.11); then to a first approximation, the singular values are given by*

$$\sigma_1(A) \approx \sqrt{a^2 q + \varepsilon^2 p} \quad \textit{and} \quad \sigma_i(A) \approx \varepsilon\sqrt{p} \quad \textit{(for } i \geq 2\textit{)} \tag{3.13}$$

□

From these results we can make the following observations:

- The difference between the first and the subsequent singular values grows proportionally to $\sqrt{q}$, as it means that the more cycles that are present in the data, the more pronounced the difference is. Furthermore, in many cases the noise-level ε^2 can be neglected with respect to the strength of the signal (a^2), resulting in a further approximation:

  $$\sigma_1(A) \approx a\sqrt{q} \tag{3.14}$$

  Which indeed tallies with the geometric intuition explained in Section 2.3.

- Impact on SVR:

  $$(SVR(A))^2 = \left(\frac{\sigma_1(A)}{\sigma_2(A)}\right)^2 = \frac{a^2 q + \varepsilon^2 p}{\varepsilon^2 p} = 1 + \frac{a^2}{\varepsilon^2}\frac{q}{p}$$

  So we can conclude that the SVR is influenced by the aspect ratio of the matrix (q/p) as well as the relative size of the signal (a) versus noise (ε).

The subsequent singular values correspond to the eigenvectors which are mapped to zero by the rank-1 matrix and therefore are not influenced by the a^2 term:

$$\sigma_i(A) \approx \varepsilon\sqrt{p} \quad (\text{for } i \geq 2)$$

In other words, these lower-ranked singular values are not influenced by the signal **a** itself, just by the noise. This is illustrated in Figure 3.7 where we took a fixed noise-level $\epsilon = 0.2$ and a signal strength a which is a multiple of some basic level $a_0 = \sqrt{12.5}$ and $a = ka_0$ with $k = 0, 1, 2, 3$. The number of full cycles in each case was equal to $q = 10$. We therefore expect the first singular value for each of these signal levels to be roughly equal to $\sqrt{q}\, a_0 k \approx 11.2k$.

It is important to realize that this observation is different from the result in Section 3.5.1 where the first singular value was affected by a shift in the mean noise level. In this case, the mean $(1/p)\sum_i a_i$ of the periodic signal **a** can still be zero, but it is its L_2 norm ($a^2 = \|\mathbf{a}\|^2$) that is seen to affect the first singular value.

Slightly more general result Consider the case of a general rank-1 matrix:

$$A = \mathbf{a}\mathbf{b}^T + \varepsilon N \tag{3.15}$$

In this case the same sort of computation yields:

$$B = a^2\mathbf{b}\mathbf{b}^T + \varepsilon^2 p I_q$$

for which the largest eigenvalue corresponds to the vector **b**:

$$B\mathbf{b} = (a^2 b^2 + \varepsilon^2 p)\mathbf{b}$$

3.5. EXPLORING SOME PROPERTIES

As mentioned before, the major goal of this thesis is to investigate the applicability of the SVD in time series analysis. To this end, the impact of the underlying patterns in the signal on the SVD results was touched upon earlier in Section 3.4. Furthermore, as it is illustrated in Figures 2.4-2.8, the distance and positioning from the origin can affect the SVD results drastically. The following section provides more detailed mathematical foundations for the observed results (with a focus on time series analysis).

3.5.1. IMPACT OF ENTRIES MEAN VALUE

In the original papers [6, 7], it was not sufficiently appreciated how a shift in the mean value of the time series (the DC component) impacts the SVR. This is important as failure to understand this issue introduces a major bias in the test values and could therefore result in erroneous conclusions. To address this issue, we compare the singular values of a zero-mean $p \times q$ random matrix A_0 and its mean-shifted version: $A = A_0 + \alpha$ which is shorthand for $A = A_0 + \alpha\mathbf{1}_{p\times q} = A_0 + \alpha\mathbf{1}_p\mathbf{1}_q^T$. Using the connection between singular values and eigenvalues expounded in the previous section, we can express any singular

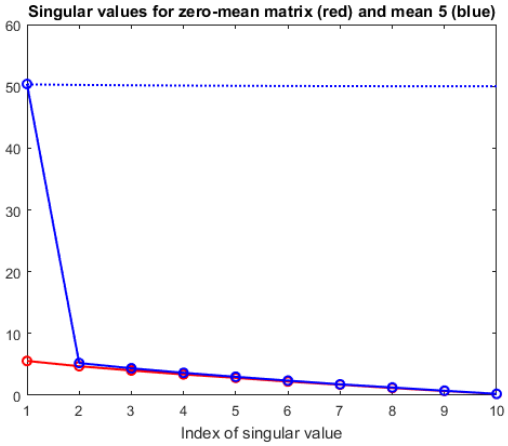


Figure 3.8: Comparison of the singular values of a matrix (10 × 10) with zero-mean entries (red) and shifted mean ($\alpha = 5$). The stem-plot of the singular values against their index i is also called the singular spectrum or eigenspectrum. The dotted line indicates the (approximate) upper limit based on Eq. (3.19). Recall that the entries of the matrix A_0 are random numbers, but by shifting the global mean the $SVR = \sigma_1/\sigma_2$ increases, erroneously suggesting that some underlying periodic structure is present.

value $\sigma(A)$ as:

$$\begin{aligned}\sigma^2(A) = \lambda(AA^T) &= \lambda((A_0 + \alpha \mathbf{1}_p \mathbf{1}_q^T)(A_0^T + \alpha \mathbf{1}_q \mathbf{1}_p^T)) \\ &= \lambda\left(A_0 A_0^T + \alpha(A_0 \mathbf{1}_q \mathbf{1}_p^T + \mathbf{1}_p \mathbf{1}_q^T A_0^T) + \alpha^2 \mathbf{1}_p \mathbf{1}_q^T \mathbf{1}_q \mathbf{1}_p^T\right) \\ &= \lambda\left(A_0 A_0^T + \alpha q(R \mathbf{1}_p^T + \mathbf{1}_p R^T) + \alpha^2 q \mathbf{1}_p \mathbf{1}_p^T\right) \end{aligned} \tag{3.16}$$

where $R = (1/q) A_0 \mathbf{1}_q$ is a $p \times 1$ column matrix for which each element is the mean of the corresponding A_0 row. However, recall that the entries of A_0 are independent zero-mean stochastic variables. Hence, unless the matrix dimensions are very small, it follows that $R \approx 0$ can be neglected. We, therefore, derive the approximation:

$$\sigma^2(A) \approx \lambda\left(A_0 A_0^T + \alpha^2 q \mathbf{1}_p \mathbf{1}_p^T\right) \tag{3.17}$$

Next, we make use of the standard results on Rayleigh quotients for eigenvalues which states that the dominant eigenvalue of a symmetric, positive definite matrix M is the solution to the maximization problem: $\lambda_1 = \max_{\mathbf{x} \neq \mathbf{0}} \left(\frac{\mathbf{x}^T M \mathbf{x}}{\mathbf{x}^T \mathbf{x}}\right) = \max_{\|\mathbf{u}\|=1} (\mathbf{u}^T M \mathbf{u})$. Furthermore, if a unit vector $\mathbf{u}_1$ realizes the above maximum, then the second largest eigenvalue is obtained as the solution of the constrained optimization problem:

$$\lambda_2 = \max_{\|\mathbf{u}\|=1} (\mathbf{u}^T M \mathbf{u}) \qquad s.t. \quad \mathbf{u} \perp \mathbf{u}_1$$

and so on for the successive eigenvalues.

Combining this with the approximation derived in Eq. (3.17), we get the following approximation for the first singular value of A:

$$\begin{aligned}\sigma_1^2(A) \approx & \max_{\|\mathbf{u}\|=1} \mathbf{u}^T \left(A_0 A_0^T + \alpha^2 q \mathbf{1}_p \mathbf{1}_p^T\right) \mathbf{u} \\ = & \max_{\|\mathbf{u}\|=1} \left(\mathbf{u}^T A_0 A_0^T \mathbf{u} + \alpha^2 q \mathbf{u}^T \mathbf{1}_p \mathbf{1}_p^T \mathbf{u}\right) \\ = & \max_{\|\mathbf{u}\|=1} \left(\mathbf{u}^T A_0 A_0^T \mathbf{u} + \alpha^2 q (\sum_{i=1}^{p} u_i)^2\right) \end{aligned} \tag{3.18}$$

This derivation shows that

$$\sigma_1^2(A) \le \max_{\|\mathbf{u}\|=1} \left(\mathbf{u}^T A_0 A_0^T \mathbf{u}\right) + \alpha^2 pq = \sigma_1^2(A_0) + \alpha^2 pq \tag{3.19}$$

since from the Cauchy-Schwartz inequality it follows:

$$\left(\sum_{i=1}^{p} u_i\right)^2 \le \left(\sum_{i=1}^{p} u_i^2\right)\left(\sum_{i=1}^{p} 1\right) = p \qquad \text{since} \quad \|u\| = 1$$

However, in general, the unit vector $\mathbf{u}$ that maximizes the Rayleigh quotient will not necessarily also maximize $(\sum u_i)^2$. In fact, for higher singular values, the number of orthogonal constraints on u increases proportionally, suggesting that on average $\sum u_i \approx 0$, and therefore $\sigma_i^2(A) \approx \sigma_i^2(A_0)$. Another argument could be that higher singular vectors are comprised of the noise in the data (low magnitude random-structured data points added to the original signal). Therefore, adding or removing the mean value will not affect their compounds in higher singular vectors and correspondingly the singular values.

This is indeed what is seen in numerical experiments (Figure 3.8). Notice that the first singular value is very close to the maximal value obtained in Eq. (3.19) which is derived if optimizing both terms in Eq. (3.18), independently and simultaneously had been done.

Clearly, failing to remove the mean from a noisy time series would inflate the first singular value (and only the first one) [3] resulting in an upwardly biased value for the singular value ratio (SVR). This would reduce the power of an SVD method in data mining applications such a blind screening. In the next section, we will investigate the impact of a genuine underlying periodic signal.

3.5.2. IMPACT OF THE DRIFT ON THE FIRST SINGULAR VALUE

Consider a signal that has a periodic component $\mathbf{a}$ and a drift component $\mathbf{k}$. For simplicity, we assume that

$$\mathbf{k} = (-k : k)^T \qquad \text{where } q = 2k+1.$$

Now, we consider a $p \times q$ matrix of the form:

$$A = \mathbf{a}\mathbf{1}_q^T + \beta \mathbf{1}_p \mathbf{k}^T$$

This represents a time series with drift (see Figure 3.9). Since each column (or row) is the

[3] It may lead to wrong impressions about the importance of the first singular value with respect to the others!

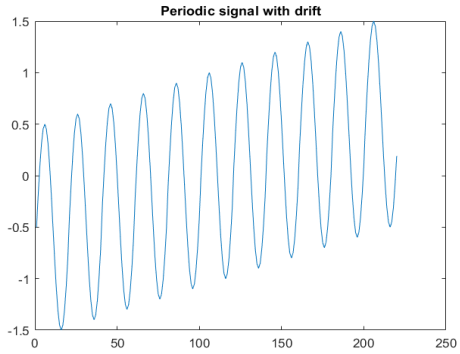


Figure 3.9: Periodic signal with drift. In this case **a** is a sine wave, while $k = 5$, and $\beta = 0.1$.

linear combination of two fixed columns, it has rank two (or less), and therefore its SVD is of the form:

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T$$

These two decomposition terms are very similar, but the second one is more specific in the sense that the vectors involved are orthogonal (and unit length). The main outcome of the figure is that pre-processing (removing the mean value) in the time series analysis is important. Only looking at the singular values may lead to biased results. One must investigate the evolution of the singular vectors to have a better understanding of the underlying patterns in the time series.

DERIVATION FOR THE SIMPLEST CASE

Let us first look at the case when there is no drift $\beta = 0$ and we assume that the base signal is zero mean, i.e., $\sum_i a_i = 0$. In that case

$$A = \mathbf{a}\mathbf{1}_q^T$$

and we get

$$\mathbf{u}_1 = \mathbf{a}/\|\mathbf{a}\| \quad \text{and} \quad \mathbf{v}_1 = \mathbf{1}_q/\sqrt{q} \quad \text{and} \quad \sigma_1 = \sqrt{q}\|\mathbf{a}\|$$

Without loss of generality, if we then introduce a small amount of drift (i.e., $\beta \ll 1$) we can simply use the fact that 1_p is orthogonal with respect to **a** since $\mathbf{a}^T 1_p = \sum_i a_i = 0$ (signal must be zero mean value, otherwise it will not work). As a consequence we can take $\mathbf{u}_2 = \mathbf{1}_p/\sqrt{p}$ which yields the decomposition:

$$A = \sqrt{q}\|\mathbf{a}\|\mathbf{u}_1\mathbf{v}_1 + \beta\sqrt{p}\mathbf{u}_2\mathbf{k}^T$$

Finally, we notice that **k** is also orthogonal to $\mathbf{v}_1$ since $\sum_i k_i = 0$, and therefore $\mathbf{v}_2 = \mathbf{k}/\|\mathbf{k}\|$. Now, recall that

$$\|\mathbf{k}\|^2 = 2\sum_{\ell=1}^{k} \ell^2 = 2\frac{k(k+1)(2k+1)}{6} \approx 2k^3/3 \approx q^3/12.$$

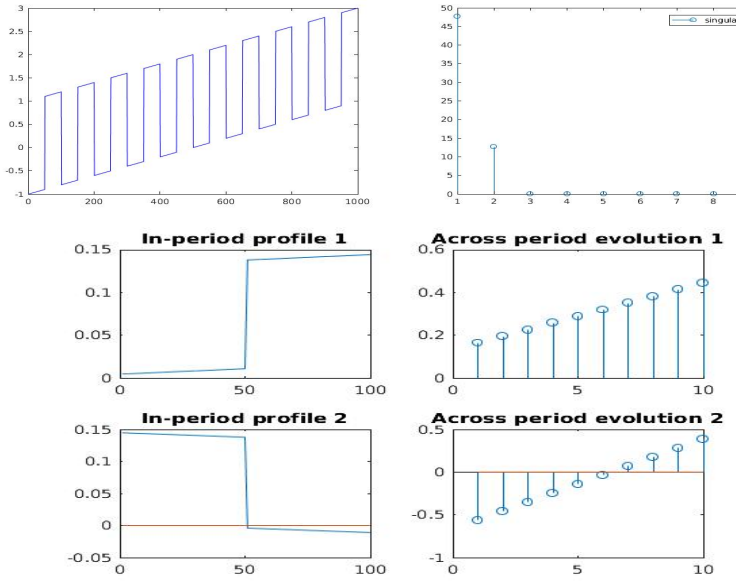


Figure 3.10: An example of a drifted time series in the absence of noise, and $\mu \neq 0$. Perhaps we could say that u_2 or $v_2$as an indication for the direction of that drift!

Plugging all this into the formula above we see that the singular values are equal to:

$$\sigma_1 = \sqrt{q}\|\mathbf{a}\| \qquad \text{and} \qquad \sigma_2 = \frac{1}{2}|\beta|\sqrt{\frac{pq^3}{3}}. \tag{3.20}$$

where $|.|$ is the absolute value operator. From Eq. (3.20) it transpires that the first singular value is determined by the periodic patterns, whereas the second one is determined by the drift. Of course, there are other ways to derive similar results and also provide an argument why $|\beta|$ is used:

$$\begin{aligned} \|\mathbf{a}\mathbf{1}_\mathbf{q}{}^T + \beta \mathbf{1}_\mathbf{p}\mathbf{k}^T\|_F^2 = \|\mathbf{a}\mathbf{1}_\mathbf{q}{}^T\|_F^2 + \|\beta \mathbf{1}_\mathbf{p}\mathbf{k}^T\|_F^2 + 2 < \mathbf{a}\mathbf{1}_\mathbf{q}{}^T, \beta \mathbf{1}_\mathbf{p}\mathbf{k}^T >_F \quad = \\ q\|a\|^2 + p|\beta|\|k\|^2 + 2\beta < \mathbf{a}\mathbf{1}_\mathbf{q}{}^T, \mathbf{1}_\mathbf{p}\mathbf{k}^T >_F \end{aligned} \tag{3.21}$$

One can argue that the largest term in the above formula is associated with the first singular value, and the second term is with the second; this explains the switches of the two terms in the numerical analysis. The third term is (close to) zero; as in our example $\mathbf{k} = [-k : k]$ is symmetric and $\mathbf{a}\mathbf{1}_\mathbf{q}{}^T$ is a rank−1 matrix.

Numerical verification Here, we also can argue that a zero mean random process with drift will exhibit a similar singular value spectrum; therefore, SVR by itself (without considering u and v profiles) can lead to erroneous results (Figure 3.12).

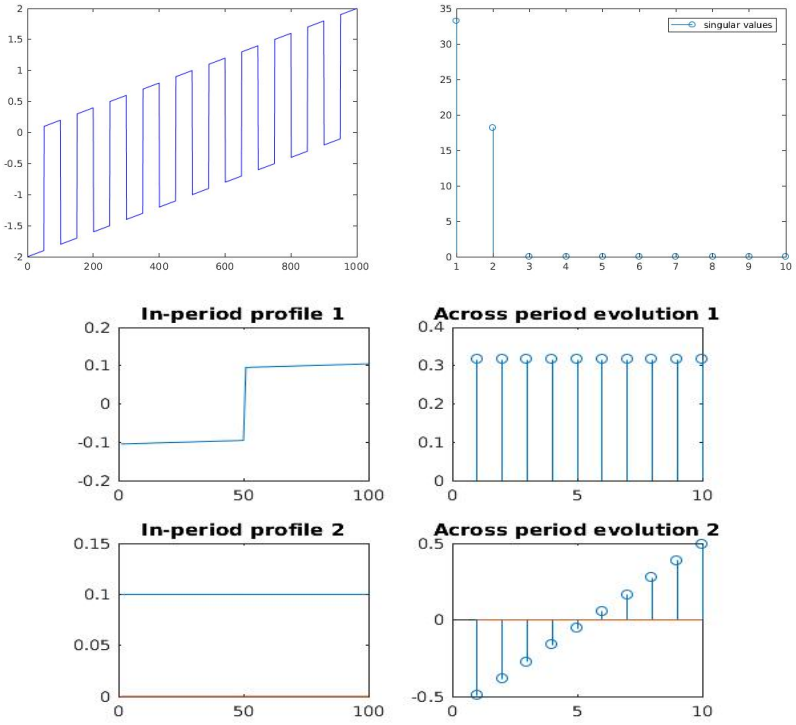


Figure 3.11: when there is a drift and in the absence of noise $\mu = 0$. notice the changes in the u and v profiles.

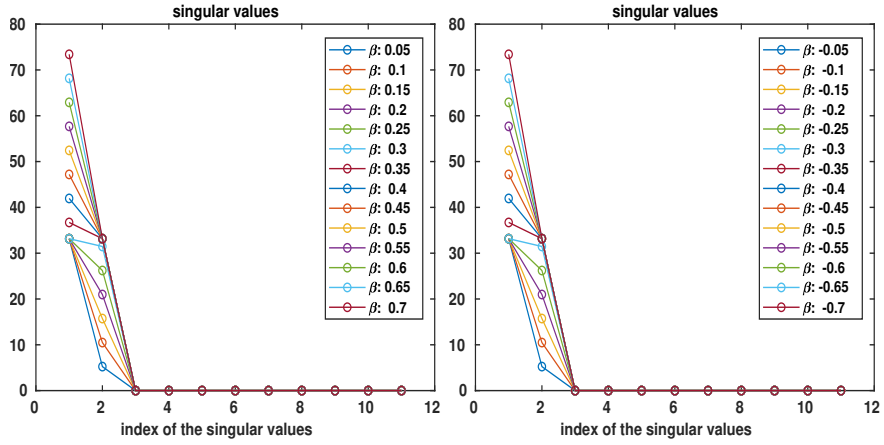


Figure 3.12: Influence of drift on 2nd singular value for the input in the figure above. Notice how the first singular value is unaffected (the first singular value is unaffected as long as the drift is small. from some point upward, the drift will change u_1, and accordingly σ_1).

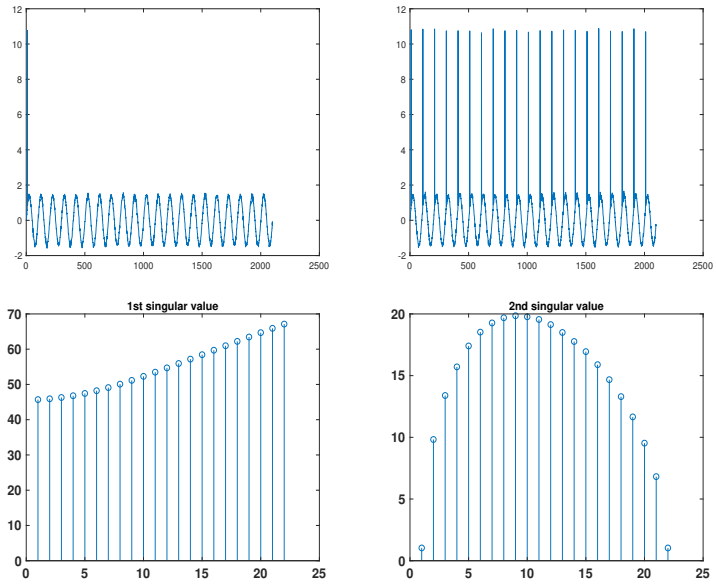


Figure 3.13: **Top:** An example of the frequency of the occurrence of an extra event. **Bottom:** The evolution of the second singular value. As expected when the occurrence is too often, the rank of the matrix decreases.

3.6. CONCLUSION

In this chapter, we have argued that the well-known singular value decomposition (SVD) (which is usually applied to matrix problems) can also be successfully applied to identify periodic patterns (profiles) in time series. Furthermore, these profiles are completely defined by the data and do not require the specification of user-defined parameters, apart from the period (which itself can be estimated using this approach). As such, this methodology offers a purely data-driven approach to adaptive signal approximation and based on that, outlier detection.

Moreover, we have shown that a judicious comparison of the V-coefficients and residuals allows one to distinguish between different ways in which data points can be atypical or salient. From a data mining perspective, this opens up new ways of analyzing time series in a data-driven, bottom-up fashion. However, it then becomes essential to thoroughly understand how the spectrum of time series is influenced by various characteristics of the signal and noise.

REFERENCES

[1] A. Khoshrou and E. J. Pauwels, *Data-driven pattern identification and outlier detection in time series,* in *Science and Information Conference* (Springer, 2018) pp. 471–484.

[2] T. Tao and V. H. Vu, *Random matrices: The distribution of the smallest singular values,* ArXiv: 0903:0614 (2009).

[3] D. Paul and A. Aue, *Random matrix theory in statistics: A review,* Journal of Statistical Planning and Inference **150**, 1 (2014).

[4] H. H. Nguyen, V. Vu, *et al.*, *Random matrices: Law of the determinant,* The Annals of Probability **42**, 146 (2014).

[5] M. Rosenblatt, *A central limit theorem and a strong mixing condition,* Proceedings of the National Academy of Sciences of the United States of America **42**, 43 (1956).

[6] P. P. Kanjilal and S. Palit, *On multiple pattern extraction using singular value decomposition,* IEEE transactions on signal processing **43**, 1536 (1995).

[7] L. H. L. J. Z. Ying and Q. Liangsheng, *Improved singular value decomposition technique for detecting and extracting periodic impulse component in a vibration signal,* Chinese Journal of Mechanical Engineering **17**, 1 (2004).

4

REGULARISED MATRIX FACTORIZATION

4.1. INTRODUCTION AND MOTIVATION

Singular Value Decomposition (SVD) and its close relative, Principal Component Analysis (PCA), are linear matrix factorisation techniques that are widely used in applications as varied as dimension reduction and clustering [3], matrix completion [4] (e.g., for recommender systems), dictionary learning [5] and time series analysis [6]. In a surprising turn of events, (deep) matrix factorisation also plays a role in the implicit regularisation that enables acceptable generalisation in deep learning [7].

Although these factorisation techniques are both conceptually simple and effective, it is well-known that they are sensitive to noise and outliers in input data. As a consequence, some modifications of the original algorithms have been proposed to alleviate the effect of these disturbances [8, 9]. Candes et al. [10] introduce *Robust PCA (RPCA)* which aims to separate signal from outliers by decomposing any given matrix into the sum of a low-rank approximation and a sparse matrix of outliers. An extension of this work for the inexact recovery of the data is presented in [11]. Another example of sparse PCA using low rank approximation is proposed in [12].

Adding a regularisation term is another versatile way to tackle the problem of noisy input. For instance, Dumitrescu et al. [13] show how a regularized version of the K-SVD algorithm can be adapted to the Dictionary Learning (DL) problem. Although, the presence of noise in the input is not the only reason to invoke regularisation. Recent research [14] shows that in many real-world data sets, not only do the observed data lie on a (non-)linear low dimensional manifold, but this also applies to the features. Similar to our approach, He et al. [15] consider a given matrix A where the columns are interpreted as data points and the rows are features. The neighbourhood structure of both the data points and the features then gives rise to distinct graphs (the so-called data and feature

Parts of this chapter have been published in [1, 2].

graphs) and hence, to corresponding graph Laplacians (L_d and L_f respectively). The resulting regularised PCA is referred to as the *graph-dual Laplacian PCA* (gDLPCA), which for a given data matrix A is obtained by minimising the below functional:

$$J(V,Y) = \|A - VY\|^2 + \alpha \operatorname{Tr}(V^T L_d V) + \beta \operatorname{Tr}(Y L_f Y^T) \qquad \text{subject to } V^T V = I \tag{4.1}$$

where $\|.\|$ and Tr are the $L-2$ norm and the trace operators, respectively. The ability of the graph dual regularization technique to incorporate both data and feature structures has deservedly attracted considerable attention in dimensionality reduction applications [15–17].

In their abstract form, SVD and PCA amount to two different but related types of matrix factorisation. More precisely, given a general (data) matrix A, the aim is to approximate it as a product of simpler (i.e., lower-rank) matrices. Specifically:

- PCA-type decomposition: $A \approx PQ^T$ where the columns of Q are orthonormal, i.e., $Q^T Q = I$;
- SVD-type decomposition: $A \approx PBQ^T$ where B is diagonal, while P and Q are unitary matrices, i.e., $P^T P = I$, $Q^T Q = I$.

The approximation in the above equations is measured in terms of the Frobenius (matrix) norm which for an arbitrary matrix $X \in \mathbb{R}^{p\times q}$ is defined as:

$$\|X\|_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} x_{ij}^2 = \operatorname{Tr}(XX^T) = \operatorname{Tr}(X^T X) = \|X^T\|_F^2. \tag{4.2}$$

In the remainder of this chapter, we will drop the subscript F. We herein take the functional Eq. (4.1) as a starting point and investigate the two factorisation approaches mentioned above (invoking Eq. (4.2) to recast the trace as a norm):

- PCA-type decomposition ($A \approx PQ^T$) by minimising the regularisation functional:

$$\|A - PQ^T\|^2 + \lambda \|DP\|^2 + \mu \|GQ\|^2 \tag{4.3}$$

- SVD-type decomposition ($A \approx PBQ^T$) by minimising the regularisation functional:

$$\|A - PBQ^T\|^2 + \lambda \|DP\|^2 + \mu \|GQ\|^2 \tag{4.4}$$

The minimisation of the functional Eq. (4.3) was discussed in [15], however, their proposed solution contains an error which we correct in this chapter. In addition, we also provide an algorithm to solve functional Eq. (4.4), which somewhat surprisingly is quite different from the one for Eq. (4.3).

The remainder of this chapter is organised as follows: In Sections 4.2 and 4.3 we derive algorithms for minimisation of the regularised version of PCA-type and SVD-type factorisations, respectively. Section 4.4 discusses how the gradient descent method can be implemented by drawing on some elementary facts from Lie-group theory. Finally, we conclude this chapter by giving some pointers to potential extensions.

4.2. REGULARISATION FOR PCA-TYPE FACTORISATION

4.2.1. REGULARISED PCA

The following theorem outlines an obvious generalisation to the regularised version of the minimisation problem.

Theorem 7 (**Regularised PCA**). *Let A be a $p \times q$ matrix of rank $r \leq \min(p,q)$. For $k \leq r$, let $P \in \mathbb{R}^{p\times k}$ and $Q \in \mathbb{R}^{q\times k}$ full rank matrices (i.e., of rank k). Furthermore, for arbitrary strictly positive integers d and g, we introduce regularisation matrices $D \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{g\times q}$, as well as weights $\lambda, \mu \geq 0$. We now define the following functional F in the variables P and Q:*

$$F(P,Q) := \|A - PQ^T\|^2 + \lambda \|DP\|^2 + \mu \|GQ\|^2 \tag{4.5}$$

and pose the corresponding constrained optimisation problem:

$$\min_{P,Q} F(P,Q) \qquad \textit{subject to} \qquad Q^T Q = I_k. \tag{4.6}$$

Introducing short-hand notation $L := D^T D \in \mathbb{R}^{p\times p}$ and $M := G^T G \in \mathbb{R}^{q\times q}$ (both symmetric and positive semi-definite), the solution of the constrained optimisation problem (4.6) is constructed as follows:

- *The k columns of the $q \times k$ matrix Q are the eigenvectors of the $q \times q$ matrix:*

$$K := A^T (I_p + \lambda L)^{-1} A - \mu M$$

  corresponding to the k largest eigenvalues;

- *Furthermore: $P = (I_p + \lambda L)^{-1} AQ$*

For the sake of completeness, let us reiterate that the condition $Q^T Q = I_k$ is not restrictive but necessary to eliminate arbitrary rescalings. In passing, we point out that the result above corrects an error in [15] where it is incorrectly stated that $P = AQ$.

Proof. Since the variable P in the functional (4.5) is unconstrained, we can identify the optimum in P (for fixed Q) by computing the gradient:

$$\frac{1}{2}\nabla_P F = (PQ^T - A)Q + \lambda D^T DP \tag{4.7}$$

and solving for P:

$$\nabla_P F = 0 \quad \Rightarrow \quad P\underbrace{Q^T Q}_{I_k} - AQ + \lambda LP = 0 \quad \Rightarrow \quad (I_p + \lambda L)P = AQ. \tag{4.8}$$

This condition needs to hold at the solution point. By first re-writing $F(P,Q)$ formula as the trace of matrices and then plugging in Eq. (4.8), we have:

$$\begin{aligned} F(P,Q) &= \mathrm{Tr}\left[(A - PQ^T)(A^T - QP^T)\right] + \lambda \,\mathrm{Tr}(P^T LP) + \mu \,\mathrm{Tr}(Q^T MQ) \\ &= \mathrm{Tr}\left[AA^T - AQP^T - PQ^T A^T + PQ^T QP^T\right] + \lambda \,\mathrm{Tr}(P^T LP) + \mu \,\mathrm{Tr}(Q^T MQ) \end{aligned}$$

Considering the fact that the trace operator is invariant under transposition as well as cyclic permutation, we arrive at:

$$\begin{aligned} F(P,Q) &= \mathrm{Tr}\left[AA^T - 2(I_p+\lambda L)PP^T + PP^T\right] + \lambda\,\mathrm{Tr}(P^TLP) + \mu\,\mathrm{Tr}(Q^TMQ) \\ &= \mathrm{Tr}\left(AA^T - PP^T - 2\lambda LPP^T\right) + \lambda\,\mathrm{Tr}(P^TLP) + \mu\,\mathrm{Tr}(Q^TMQ) \\ &= \mathrm{Tr}(AA^T) - \mathrm{Tr}(PP^T) - 2\lambda\,\mathrm{Tr}(LPP^T) + \lambda\,\mathrm{Tr}(P^TLP) + \mu\,\mathrm{Tr}(Q^TMQ) \\ &= \mathrm{Tr}(AA^T) - \mathrm{Tr}(P^TP) - \lambda\,\mathrm{Tr}(P^TLP) + \mu\,\mathrm{Tr}(Q^TMQ) \\ &= \mathrm{Tr}(AA^T) - \mathrm{Tr}\left[P^T\underbrace{(I_p+\lambda L)P}_{AQ}\right] + \mu\,\mathrm{Tr}(Q^TMQ). \end{aligned} \tag{4.9}$$

Extracting P and its transpose from Eq. (4.8):

$$P = (I_p+\lambda L)^{-1}AQ \quad \Rightarrow \quad P^T = Q^TA^T(I_p+\lambda L)^{-1} \qquad \text{as } L \text{ is symmetric.} \tag{4.10}$$

hence:

$$F(P,Q) = \mathrm{Tr}(AA^T) - \mathrm{Tr}\left[Q^T\left(A^T(I_p+\lambda L)^{-1}A - \mu M\right)Q\right]. \tag{4.11}$$

Therefore, in order to minimize F, one must maximize the right-most term in Eq. (4.11), as $\mathrm{Tr}(AA^T)$ is a constant value. This is achieved by selecting for Q, eigenvectors corresponding to the k largest eigenvalues of $(A^T(I_p+\lambda L)^{-1}A - \mu M)$. Once Q is determined, P is obtained via Eq. (4.10).

As a concluding remark, we point out that the matrix $I_p+\lambda L$ is always invertible. Indeed, since $L = D^TD$ is positive semi-definite and symmetric, it has a complete set of eigenvectors with corresponding non-negative eigenvalues, i.e., $L = W\Lambda W^T$, where W is orthogonal (i.e., $W^TW = WW^T = I_p$) and $\Lambda \geq 0$. Hence, the matrix

$$(I_p+\lambda L) = W(I_p+\lambda\Lambda)W^T$$

has a complete set of strictly positive eigenvalues, and is therefore invertible. □

Some illustrative numerical experiments can be found in [18].

4.2.2. SOME SPECIAL CASES

- $\boxed{\lambda = 0 \text{ and } \mu = 0}$: In that case, Q comprises the first k eigenvectors of $K = A^TA$ and $P = AQ$, which means that we end up with the standard SVD, as expected. Some numerical experiments can be found in [19].

- $\boxed{D = I_p \text{ and } \mu = 0}$: These conditions correspond to what is assumed in [13] where a regularized K-SVD problem is addressed. In the aforementioned work, the authors consider a special case, where $\mu = 0$ and $D = I_p$. Since this implies that $L = D^TD = I_p$ and $\mu M = 0$, the matrix K simplifies to:

$$K = \frac{1}{1+\lambda}A^TA$$

The eigenvectors of K are therefore the right singular vectors of A (i.e., the eigenvectors of A^TA). Hence $Q = V_{(1:k)}$, and as a result:

$$P = \frac{1}{1+\lambda} AQ \quad and \quad AQ = U_{(1:k)}\, diag(\sigma_1, \ldots, \sigma_k).$$

In particular, for $k = 1$ (the rank-1 reconstruction), we obtain:

$$Q = V_1 \quad \text{and} \quad P = \frac{\sigma_1}{1+\lambda} U_1$$

which is the result that can be found in [13]. The experiments are available in [20].

4.3. REGULARISATION FOR SVD-TYPE FACTORISATION

Having discussed the PCA-type factorisation, let us next turn our attention to the SVD-type factorisation which looks for an approximation of the form:

$$A \approx PBQ^T \quad \text{subject to:} \quad Q^TQ = I_k, \quad \|P_i\| = 1 \quad \forall i \in \{1, 2, \ldots, k\}, \text{ and } B \text{ diagonal.}$$

Loosely speaking, since the columns of P and Q are of unit length, they only pin down the structure of A, whereas the diagonal matrix $B = diag(\beta_1, \beta_2, \ldots, \beta_k)$ captures the *amplitude* of the corresponding structures. Similar to SVD, discussed in Chapter 2, the columns of Q are orthonormal, i.e., we again insist on $Q^TQ = I_k$. However, unlike before, the columns of P are now only required to have unit length. In respect of the SVD-type matrix factorisation technique, Theorems 8 and 9 provide alternative solutions to the lower-rank matrix approximation problem. For notational convenience, Theorem 8 first addresses a simplified case of functional (4.4) where $\mu = 0$. Following that, Theorem 9 discusses a more general case of the SVD-type factorisation.

Theorem 8 (**Regularised SVD**). *Let A be a $p \times q$ matrix of rank $r \leq \min(p, q)$. Moreover, for $k \leq r$, let $P \in \mathbb{R}^{p\times k}$ and $Q \in \mathbb{R}^{q\times k}$ of rank k, while $B \in \mathbb{R}^{k\times k}$ diagonal (i.e., $B = diag(\beta_1, \beta_2, \ldots, \beta_k)$). Furthermore, for an arbitrary positive integer d, we introduce a regularisation matrix $D \in \mathbb{R}^{d\times p}$, as well as weight $\lambda \geq 0$. Finally, we introduce the short-hand notation $L := D^TD \in \mathbb{R}^{p\times p}$ (symmetric and positive-definite). We are now in a position to define the following functional F of the variables P, Q and B:*

$$F(P, Q, B) = \|A - PBQ^T\|^2 + \lambda \|DP\|^2, \tag{4.12}$$

and the corresponding constrained optimisation problem:

$$\min_{P,Q,B} F(P, Q, B) \quad \textit{subject to:} \quad Q^TQ = I_k, \quad \|P_i\| = 1 \ \forall i \in \{1, 2, \ldots, k\}, \ \textit{and } B \textit{ diagonal.} \tag{4.13}$$

This problem is solved by the solution specified below in Algorithm 1.

Proof. Since B is unconstrained, we can determine its optimal value by computing the derivative with respect to B and equating it to zero:

$$\nabla_B F(P, Q, B) = \nabla_B \|A - PBQ^T\|^2. \tag{4.14}$$

Algorithm 1: Regularised SVD method (version 1: $\mu = 0$)

Input: A, k, λ, D
Output: P, B, Q
Initialization
while *no convergence* **do**

1. Determine the $q \times k$ matrix $Q = [Q_1, Q_2, \ldots, Q_k]$ (with orthonormal columns, i.e., $Q^T Q = I_k$) such that the sum of the smallest eigenvalue of each of k symmetric matrices $S(Q_i)$ (see Eq. (4.21)) is minimal, i.e.,:

$$\min_Q \psi(Q) = \min_Q \sum_{i=1}^{k} \lambda_1(Q_i) \qquad \text{such that } Q^T Q = I_k$$

where $\lambda_1(Q_i) = \min(eig(S(Q_i))$. To this end we use gradient descent (see Section 4.4).

2. For each Q_i as determined above, take P_i to be the eigenvector $W_1(Q_i)$ corresponding to the smallest eigenvalue $\lambda_1(Q_i)$. Construct the $p \times k$ matrix $P = [P_1, P_2, \ldots, P_k]$.

3. Finally, set $B = diag(\beta_1, \ldots, \beta_n)$ where $\beta_i = (P^T AQ)_{ii}$.

end

Expanding the norm in terms of a trace (cf. Eq. (4.2)), then using the invariance of a trace under transposition, we arrive at (recall $Q^T Q = I_k$):

$$\begin{aligned}
\|A - PBQ^T\|^2 &= \mathrm{Tr}\left[(A - PBQ^T)(A^T - QBP^T)\right] \\
&= \mathrm{Tr}(AA^T) - 2\mathrm{Tr}(AQBP^T) + \mathrm{Tr}(PB^2P^T) \\
&= \|A\|^2 - 2\,\mathrm{Tr}(P^T AQB) + \mathrm{Tr}(B^2 P^T P) \\
&= \|A\|^2 - 2\sum_{i=1}^{k} (P^T AQ)_{ii}\,\beta_i + \sum_{i=1}^{k} (P^T P)_{ii}\,\beta_i^2 \\
&= \|A\|^2 - 2\sum_{i=1}^{k} (P^T AQ)_{ii}\,\beta_i + \sum_{i=1}^{k} \beta_i^2 \qquad (4.15)
\end{aligned}$$

The last simplification is obtained due to the fact that $\|P_i\| = 1 \Rightarrow (P^T P)_{ii} = 1$. Therefore the gradient of the functional F with respect to B is calculated as follow:

$$\frac{\partial}{\partial \beta_i} \|A - PBQ^T\|^2 = 2\,(\beta_i - (P^T AQ)_{ii}).$$

For given P and Q matrices, we then find the optimal B by insisting that the resulting gradient vanishes, which yields:

$$\beta_i = (P^T AQ)_{ii} \qquad \forall\, i \in \{1, 2, \ldots, k\}. \qquad (4.16)$$

Plugging this optimal choice back into Eq. (4.15) the functional (4.12) simplifies to

$$\|A - PBQ^T\|^2 = \|A\|^2 - \sum_{i=1}^{k} \beta_i^2 \tag{4.17}$$

In order to recast Eq. (4.17) in terms of P and Q (and consequently eliminate B), we observe that for an arbitrary matrix H, we have $H_{ij} = \mathbf{e}_i^T H \mathbf{e}_j$, where $\mathbf{e}_i = (0,0,\ldots,1,\ldots,0)^T$ vectors are the standard bases (similarly for $\mathbf{e}_j$). Hence, using the fact that the diagonal of a matrix is unchanged under transposition, we conclude that:

$$\beta_i = \begin{cases} (P^T AQ)_{ii} & = \mathbf{e}_i^T P^T AQ \mathbf{e}_i = P_i^T AQ_i \\ (Q^T A^T P)_{ii} & = \mathbf{e}_i^T Q^T A^T P \mathbf{e}_i = Q_i^T A^T P_i \end{cases}$$

where P_i, Q_i are the i-th columns of P and Q, respectively. i.e., $P = [P_1, P_2, \ldots, P_k]$ and $Q = [Q_1, Q_2, \ldots, Q_k]$. As a consequence:

$$\sum_{i=1}^{k} \beta_i^2 = \sum_{i=1}^{k} P_i^T AQ_i Q_i^T A^T P_i. \tag{4.18}$$

As a final step, we introduce the notation $L = D^T D$ to recast the regularisation term as:

$$\|DP\|^2 = \mathrm{Tr}(P^T LP) = \sum_{i=1}^{k} \mathbf{e}_i^T P^T LP \mathbf{e}_i = \sum_{i=1}^{k} P_i^T L P_i. \tag{4.19}$$

Plugging Eqs. (4.18) and (4.19) into Eq. (4.12), we obtain the following simplified form for the functional F (assuming that we eliminate B using its optimal value):

$$F(P,Q) = \|A\|^2 + F_1(P,Q), \quad \text{where} \quad F_1(P,Q) = \sum_{i=1}^{k} P_i^T (\lambda L - AQ_i Q_i^T A^T) P_i \tag{4.20}$$

Introducing the notation $S(Q_i) := \lambda L - AQ_i {Q_i}^T A^T$, we derive:

$$F_1(P,Q) = \sum_{i=1}^{k} P_i^T S(Q_i) P_i \tag{4.21}$$

Since each $S(Q_i)$ is a symmetric $p \times p$ matrix, it can be diagonalised with respect to an orthonormal basis, i.e., there is an orthogonal $p \times p$ matrix W (such that $W^T W = WW^T = I_p$) and a diagonal matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ (ordered $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$), both depending on Q_i such that

$$S(Q_i) = W(Q_i)\Lambda(Q_i)W(Q_i)^T$$

i.e., the columns of W are the eigenvectors of $S(Q_i)$, with the corresponding eigenvalues on the diagonal of Λ. By introducing the notation $\lambda_1(S(Q_i))$ to denote the smallest eigenvalue of $\Lambda(Q_i)$, we obtain the minimal value $P_i^T S(Q_i) P_i = \lambda_1(Q_i)$, by choosing P_i to be the (unit) eigenvector $(W_1(Q_i))$ corresponding to the smallest eigenvalue. As a consequence, the solution strategy boils down to steps in Algorithm 1.

This choice of P, Q and B solves the constrained minimisation problem (4.13). Notice that due to the fact that P and B matrices are determined after finding Q, this optimisation problem can essentially be translated into a search problem in the space of Q matrices. Some illustrative numerical experiments are available at [21]. □

Below we proceed further by giving a slightly more general version of the previous theorem where $\mu \neq 0$, thus re-establishing the symmetry between P and Q.

Theorem 9 (**Regularised SVD, symmetric version**). *Similar to the previous theorem, let A be a $p \times q$ matrix of rank $r \leq \min(p,q)$. For $k \leq r$, let $P \in \mathbb{R}^{p\times k}$ and $Q \in \mathbb{R}^{q\times k}$ of rank k, while $B \in \mathbb{R}^{k\times k}$ diagonal (i.e., $B = diag(\beta_1, \beta_2, \ldots, \beta_k)$). Furthermore, for arbitrary strictly positive integers d and g we introduce regularisation matrices $D \in \mathbb{R}^{d\times p}$, and $G \in \mathbb{R}^{g\times q}$, as well as weights $\lambda, \mu \geq 0$. Finally, we introduce the short-hand notation $L := D^T D \in \mathbb{R}^{p\times p}$ and $M := G^T G \in \mathbb{R}^{q\times q}$ symmetric and positive-definite. We are now in a position to define the following functional F in the variables P, Q and B:*

$$F(P,Q,B) = \|A - PBQ^T\|^2 + \lambda \|DP\|^2 + \mu \|GQ\|^2 \tag{4.22}$$

and the corresponding constrained optimisation problem:

$$\min_{P,Q,B} F(P,Q,B) \qquad \textit{subject to:} \qquad Q^T Q = I_k, \quad \|P_i\| = 1, \ \forall i \in \{1,2,\ldots,k\} \quad \textit{and } B \textit{ diagonal.} \tag{4.23}$$

A solution to this problem is proposed in Algorithm 2.

Proof. Following the notations and results introduced above and in Theorem 8, let us add:

$$\|GQ\|^2 = \mathrm{Tr}(Q^T M Q) = \sum_{i=1}^{k} Q_i^T M Q_i$$

Hence, the functional (4.22) can be recast as:

$$F(P,Q) = \|A\|^2 + F_2(P,Q), \quad \text{where} \quad F_2(P,Q) = \sum_{i=1}^{k} P_i^T (\lambda L - A Q_i Q_i^T A^T) P_i + \mu \sum_{i=1}^{k} Q_i^T M Q_i \tag{4.24}$$

The minimum of each term in the first summation in F_2 is equal to the smallest eigenvalue $\lambda_1(S(Q_i))$. Finding the minimum for the constrained optimisation problem (4.23) therefore amounts to finding the minimum of the functional:

$$\psi(Q) := \sum_{i=1}^{k} \left(\lambda_1(S(Q_i)) + \mu Q_i^T M Q_i\right) \tag{4.25}$$

subject to the constraint $Q^T Q = I_k$. Therefore, the minimisation problem again calls for minimisation in Q space, as the optimal choice for P (corresponding eigenvectors) follows automatically. We, therefore, arrive at the following Algorithm 2. Some illustrative numerical examples are available in [21]. □

4.4. COMPUTATIONAL ASPECTS

4.4.1. GRADIENT AND RANDOM DESCENT ON THE UNITARY DOMAIN

Gradient Descent From Algorithm 2 it becomes clear that the full regularisation problem can be reduced to a simpler constrained minimisation problem detailed in Eq. (4.25). Since the ψ-functional is smooth on a compact domain, this minimum is guaranteed to

Algorithm 2: Regularized SVD method (version 2: $\mu \neq 0$)

Input: A, k, μ, λ, D, G
Output: P, B, Q
Initialization
$L = D^T D, \quad M = G^T G$
while *no convergence* **do**

1. Recall that for any unit vector $Q_i \in \mathbb{R}^q$ we define $S(Q_i) = \lambda L - AQ_i{Q_i}^T A^T$. Since this is a symmetric $p \times p$ matrix, it has a complete set of eigenvectors and corresponding eigenvalues. Denote the smallest eigenvalue of each $S(Q_i)$ as $\lambda_1(S(Q_i))$, and the corresponding (unit) eigenvector as $\mathbf{w}_1(S(Q_i))$.

2. For a given $q \times k$ matrix $Q = [Q_1, Q_2, \ldots, Q_k]$ (with orthonormal columns: $Q^T Q = I_k$) compute the functional:

$$\psi(Q) := \sum_{i=1}^{k} \left(\lambda_1(S(Q_i)) + \mu Q_i^T M Q_i\right)$$

   and use gradient descent (on the compact *torus domain*, see Section 4.4) to find the minimum.

3. For each Q_i as determined above, take P_i to be the eigenvector $W_1(Q_i)$ corresponding to the smallest eigenvector $\lambda_1(S(Q_i))$. Construct the $p \times k$ matrix $P = [P_1, P_2, \ldots, P_k]$.

4. Finally, set $B = diag(\beta_1, \ldots, \beta_n)$ where $\beta_i = (P^T AQ)_{ii}$.

end

exist and one can use gradient descent to locate it. However, gradient descent needs to respect the constraint $Q^T Q = I_k$, i.e., the Q-columns need to constitute orthonormal bases (or *frames*). This can be achieved by applying an orthogonal transformation ("rotation") to the current Q matrix, as it will preserve orthonormality. Put differently, applying a rotation R to an orthonormal frame Q results in a new orthonormal frame (say $\bar{Q}$). In mathematical parlance, recall that all orthogonal $q \times q$ matrices with a determinant equal to 1 (rather than -1) constitute a multiplicative group denoted as $SO(q)$ and formally defined as:

$$SO(q) = \left\{R \in \mathbb{R}^{q \times q} \mid RR^T = I_q = R^T R, \quad \text{and} \quad \det(R) = 1\right\}$$

It is then straightforward to check that for any $R \in SO(q)$, it holds that if $\bar{Q} = RQ$, the condition $Q^T Q = I_k$ implies that $\bar{Q}^T \bar{Q} = I_k$.

In view of the above, it follows that we can generate the "*infinitesimal variations*" needed to compute the gradient $\nabla_Q \psi(Q)$ by applying "sufficiently small" orthogonal matrices to the current value of Q. More precisely, we draw on the fact that $SO(q)$ is actually a Lie-group [22] and that therefore each $R \in SO(q)$ can be generated by exponentiating an element from its Lie-algebra $SO(q) = \left\{K \in \mathbb{R}^{q \times q} \mid K^T = -K\right\}$ (the skew-symmetric

matrices):

$$R = \exp(tK) \equiv I_q + tK + \frac{1}{2!}t^2K^2 + \ldots + \frac{1}{n!}t^pK^p + \ldots \qquad \text{(with } K^T = -K\text{)}$$

By choosing t sufficiently small, one obtains an orthogonal transformation that is close to the identity I_q. Furthermore, it suffices to restrict the variations to orthogonal transformations that result from exponentiating a basis for the space of skew-symmetric matrices. Such a basis is provided by the $q(q-1)/2$ skew-symmetric matrices K_{ij} (where $1 \le i < j \le q$) for which the matrix element k, ℓ is given by:

$$K_{ij}(k,\ell) = \begin{cases} 1 & \text{if} \quad k = i, \ell = j \\ -1 & \text{if} \quad k = j, \ell = i \\ 0 & \quad \text{otherwise} \end{cases}$$

Worth noting that $SO(q)$ can equivalently be generated, using any random K matrix that satisfies $K = -K^T$ [23]. Given the current value Q_0, we construct nearby values for Q by looping over $K_{12}, K_{13}, K_{23}, \ldots etc$ and constructing the corresponding orthogonal matrices $R_{12}(t) = \exp(tK_{12}), \ldots, etc$. Denoting these "*infinitesimal*" rotation matrices as R_α (where $\alpha = 1, \ldots, q(q-1)/2$), we see that the partial derivatives with respect to these rotations can be estimated as:

$$\frac{\partial \psi(Q)}{\partial R_\alpha} \approx \frac{\psi(R_\alpha(t)Q_0) - \psi(Q_0)}{t} \qquad \text{(for } t \text{ sufficiently small).}$$

From these results we can select the infinitesimal rotation that results in the steepest descent.

Random Descent It is worth mentioning that since computing $\psi(Q)$ is computationally expensive (it requires determining eigenvalues), a viable alternative to computing the gradient is *random descent*: generate random rotations (by exponentiating random skew matrices) and check whether they result in a lower ψ-value. As soon as one is found, proceed in that direction, and repeat the process.

4.4.2. ILLUSTRATIVE EXAMPLE: SMOOTHING A NOISY MATRIX

One way to think about the regularisation based on the matrix product DP (a similar argument holds for GQ), is that the *rows of* D specify *filters* (with applications e.g., in image processing) that will be applied to the *columns of* P. Indeed, using the convention that A_i and A^i denote the i-th column and row of A respectively, we observe that:

$$DP = \begin{pmatrix} D^1 \\ \vdots \\ D^d \end{pmatrix} (P_1, \ldots, P_k) = \begin{pmatrix} D^1P_1 & \ldots & D^1P_k \\ \vdots & \vdots & \vdots \\ D^dP_1 & \ldots & D^dP_k \end{pmatrix}$$

Since the functional minimisation attempts to keep the norm of the resulting matrix small, this amounts to keeping the response of the filters on the P-columns sufficiently small.

To illustrate the above, let us start from the assumption, common in the literature e.g., [14, 15, 24], that the $p \times q$ data matrix A has a relatively smooth underlying structure that is corrupted by noise:

$$A = UV^T + \tau Z$$

where the $p \times q$ matrix Z has independent standard normal entries, and τ controls the size of the noise.

To recover the underlying "signals" U and V, we minimise the SVD-type regularisation functional (4.22) where the smoothness of the result is enforced by using regularisation matrices D and G that extract the second derivative as follows:

$$D = F = \begin{bmatrix} -1 & 1 & 0 & & \cdots & & 0 \\ 1 & -2 & 1 & 0 & \cdots & & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \cdots & 0 & 1 & -2 & 1 \\ 0 & & \cdots & & 0 & 1 & -1 \end{bmatrix}$$

A typical result for a rank-1 ($k = 1$) approximation is depicted in Figure 4.1, and compared to the standard SVD solution. This illustrative example is available in [25].

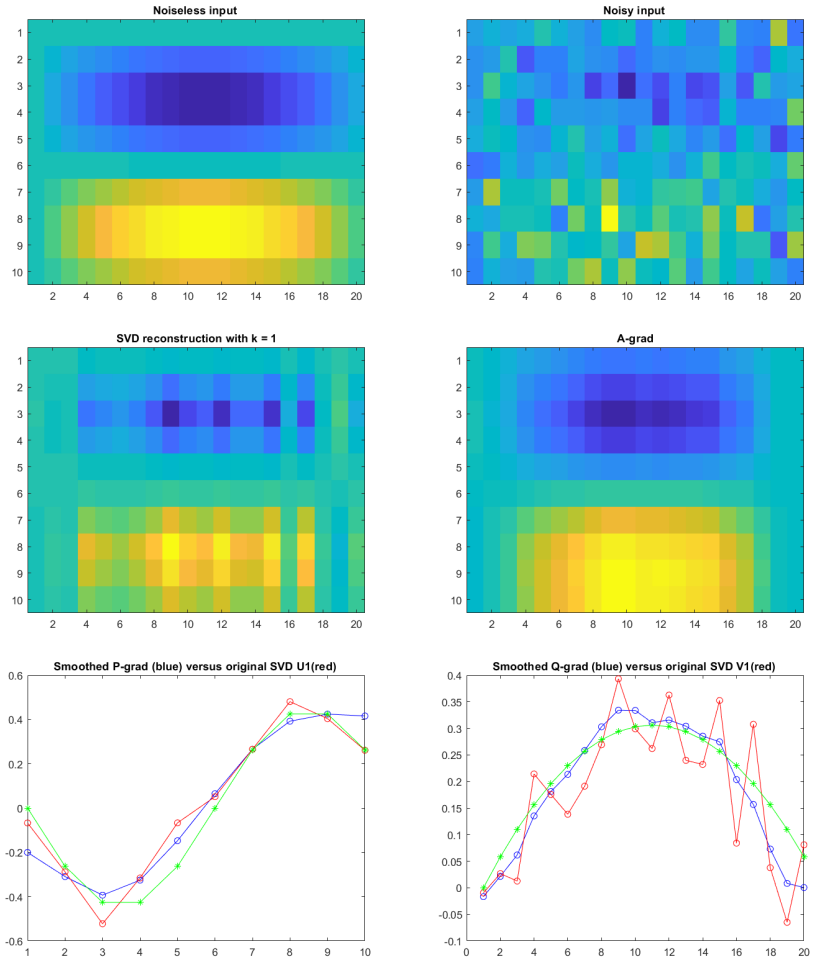


Figure 4.1: Reconstruction of noisy matrix based on RSVD. Top left: noise-less rank-1 matrix UV^T, (image) , top right: noisy input image $UV^T + \tau Z$ (high noise level), Middle left: standard rank-1 SVD reconstruction, middle right: RSVD reconstruction (D and G are 2nd derive matrices. weight parameters $\lambda = \mu = 1.5$). Bottom: comparison of standard SVD $U(:,1)$ (red) versus P (blue), and $V(:,1)$ (red) (left) versus Q (blue) (right). The actual U and V for the noiseless input signal are drawn in green.

4.5. EARLIER RESULTS, BASED ON ADHOC SMOOTHING

Throughout this thesis, the way we have constructed matrices out of time series means that rows and columns play a slightly different roles. Rows capture the in-period (e.g. in-day) variation, whereas columns represent the between-period variation. Put differently, rows capture the “fast dynamics", and columns the slow dynamics. This shows that we can look at the SVD as a (temporal) multi-scale decomposition. In this chapter, we will use this insight for two applications:

- Smoothing and image enhancing
- Forecasting (using V component to predict next day)

In below, we first introduce the second derivative operator as a pre-processing step. The use of SVD to outline the patterns in the data is presented next.

Rank-7 reconstruction In addition to appearing visually unbiased, the choice of reconstruction rank was done based on the structural similarity index (SSIM)–a popular criterion in image quality assessments [26]. SSIM measures the deviation of a reconstructed image (matrix) A_p from its original matrix A, by comparing their corresponding local means, standard deviations, and cross-covariance matrices [26]. Figure 4.2 displays the fact that the matrices reconstruction quality level off after $p = 7$, in terms of SSIM. Consequently, we opted for rank-7 as an appropriate approximation in our methodologies. It is worth noting that this figure also highlights the different levels of the volatility of the various quantities: wind (most erratic) results in the lowest similarity, while solar feed-in (most predictable) agrees best with the approximation.

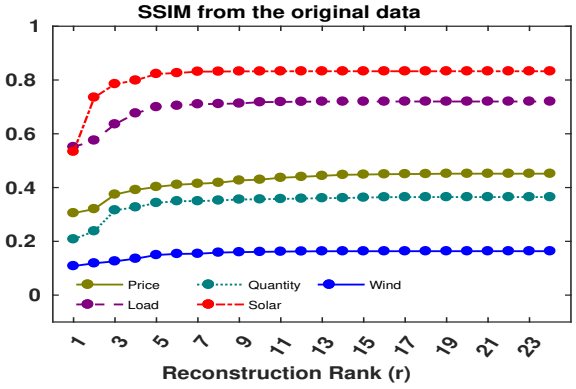
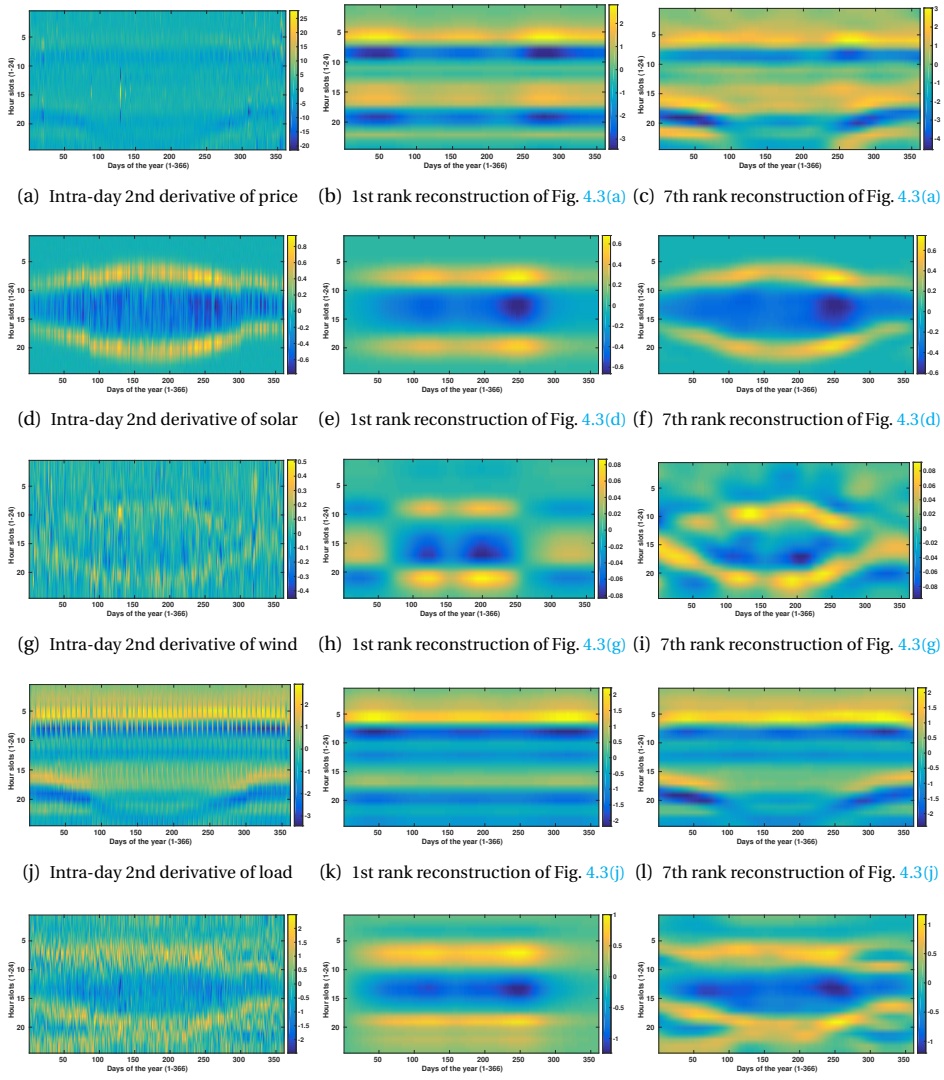


Figure 4.2: An overview of SSIM, for different reconstruction ranks ($p = 1, \ldots, 24$).

4.5.1. FINDING PEAKS AND VALLEYS

As mentioned previously, the scope of this work, in the context of the electricity market, is to explore how the inherent variability of the supply by RES can change the intra-day volatility of the day-ahead price values. To this end, within this matrix context, we add



(a) Intra-day 2nd derivative of price (b) 1st rank reconstruction of Fig. 4.3(a) (c) 7th rank reconstruction of Fig. 4.3(a)

(d) Intra-day 2nd derivative of solar (e) 1st rank reconstruction of Fig. 4.3(d) (f) 7th rank reconstruction of Fig. 4.3(d)

(g) Intra-day 2nd derivative of wind (h) 1st rank reconstruction of Fig. 4.3(g) (i) 7th rank reconstruction of Fig. 4.3(g)

(j) Intra-day 2nd derivative of load (k) 1st rank reconstruction of Fig. 4.3(j) (l) 7th rank reconstruction of Fig. 4.3(j)

(m) Intra-day 2nd derivative of quantity (n) 1st rank reconstruction of Fig. 4.3(m) (o) 7th rank reconstruction of Fig. 4.3(m)

Figure 4.3: Left: The intra-day 2nd derivative of the daily profiles in 2016, for (top to bottom) price, solar and wind feed-in, load and traded quantity. The underlying trends are magnified using rank-1 (middle) and rank-7 (right) reconstruction.

an initial pre-processing step- an additional transformation on the daily profiles of the quantities of interest (viz. price, load, traded quantity, solar and wind feed-in). More precisely, any of the above daily profiles (generically denoted by f) is regarded as a function of two variables:

- time of the day; hour slots $1 \leq h \leq 24$
- day of the year; $1 \leq d \leq 366$ (2016 is a leap year!)

Therefore, for such a function $f(h,d)$, it is feasible to obtain the corresponding intra-day 2nd derivatives with respect to the hour:

$$f_{hh} \equiv \frac{\partial^2 f}{\partial h^2} \approx \frac{f(h+1) - 2f(h) + f(h-1)}{h^2} \tag{4.26}$$

The resultant 2nd derivative profiles represent the daily dynamics of the original matrix columns; as their extreme values capture the peaks (i.e. local maxima for which $f_{hh} < 0$ and extreme) or valleys (i.e. local minima for which $f_{hh} > 0$ and extreme). Figure 4.4 pro-

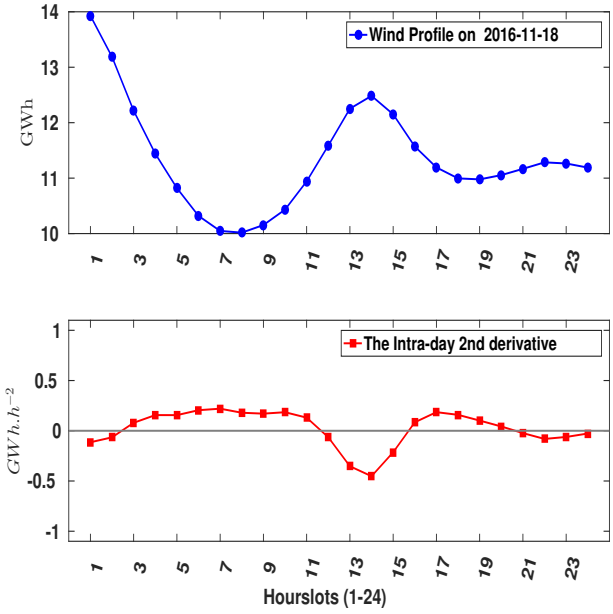


Figure 4.4: The day-ahead wind profile on Nov. 18, and its corresponding intra-day 2nd derivatives. Every sag in the lower profile corresponds to a swell in upper and vice versa.

vides an example where the upper panel contains the wind feed-in profile on Nov. 18, 2016, and the lower one corresponds to its intra-day 2nd derivative profile. We contend that comparing the evolution of the intra-day 2nd derivative profiles is useful in investigating the impact of the wind and solar energy feed-in on the price and also the traded quantity in 2016.

Applying this intra-day 2nd derivative operator to all five quantities of interest, in the matrix context, facilitates spotting some interesting features of the data. Figure 4.5

illustrates the 2nd derivatives for both solar (left) and wind feed-in (right). In the left panel, the gradual shift of the daybreak and the nightfall over the seasons is clearly visible. Closer inspection of this figure also reveals the dates of the switch to daylight saving summer time (days 87 and 304). As expected, the wind values (right panel) are more er-

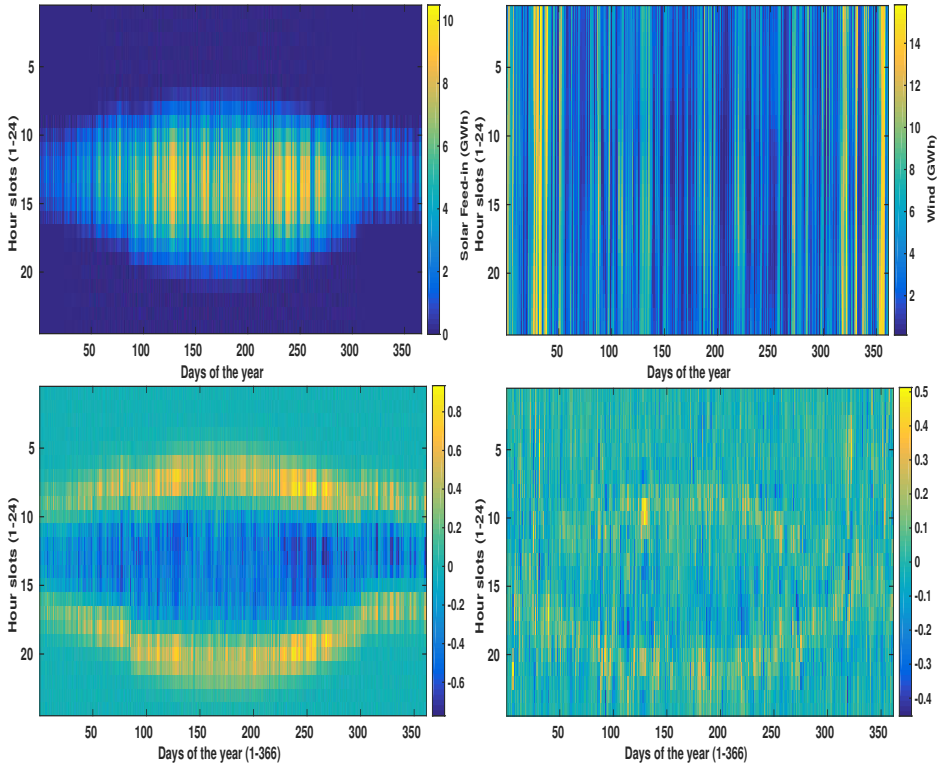


Figure 4.5: Top figures provide an overview of the solar (**left**) and wind (**right**) day-ahead values. Bottom figures contains the second derivative of the same data. Using this approach magnifies the underlying patterns in the data and also enables us to spot anomalies and abrupt changes in the data.

ratic and less seasonally determined. However, there is a striking "eye-like" shape that faintly mirrors the intra-day wind activities (see, e.g., [27]).

4.5.2. USING SVD TO HIGHLIGHT STRUCTURE

SVD is a conceptually simple and numerically stable matrix decomposition technique [28]. An outstanding feature of the SVD method is its ability in separating the fundamental "profiles" constituting a quasi-periodic time series and also indicating their relative strengths [29].

The geometrical interpretation of the SVD indicates that the columns of $U \in \mathcal{O}(24 \times 24)$ represent daily profiles, whereas the columns of $V \in \mathcal{O}(366 \times 366)$ furnish corresponding amplitudes (one for each day). Figure 4.6 shows a concrete illustration of the 24×366 price matrix: the upper left panel depicts the most dominant profile (the first column U_1)

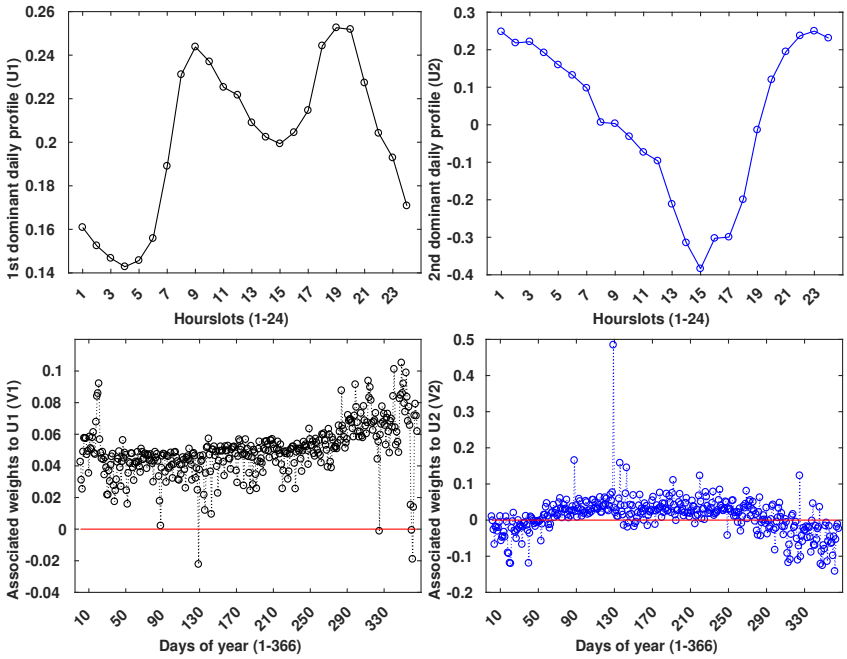


Figure 4.6: The first two most dominant U profiles (U_1 and U_2). The lower two panels contain their corresponding amplitudes (V_1 and V_2) throughout the year.

which highly resembles the overall daily profile (basically, a weighted average over the year). The expected morning and early evening price picks are clearly discernible. Their corresponding daily amplitudes are given by column V_1, as displayed in the 3rd panel (bottom left). Exceptional days with notable low or high (average) prices are clearly visible. A rank-1 approximation of the original time series could, therefore, be obtained by putting $p = 1$ in Eq. (2.5); i.e. using only the most dominant singular value σ_1, and the average daily profile U_1 in the top panel and their corresponding magnitude throughout the year using 366 values in V_1. Note that in such a rank-1 matrix, all the columns are linearly dependent, i.e., the shape of all the daily profiles are the same and only their amplitudes vary from one day to another. The U_2 profile (top right) is considered an account for seasonal variations, as it defines the first correction to U_1. Their corresponding amplitudes for this correction are specified in V_2 (bottom right). Put differently, this correction implies that any day for which the corresponding V_2 coefficient is positive (mostly during the summer) will have a lower price value between 11h and 18h than would be expected based on the (weighted) annual average U_1. In a similar fashion, increasing the reconstruction rank, by adding additional terms in the SVD expansion will improve the approximation. Figure 4.7 presents a case where the day-ahead price data for a given day (18 Jan. 2016) can be almost fully reconstructed using lower-rank approximation up to rank 7.

Figure 4.8 displays V_1 and its corresponding smoother version which was applied in the calculation of A_p. Using the smoothed version of the V_k vectors in the low-rank re-

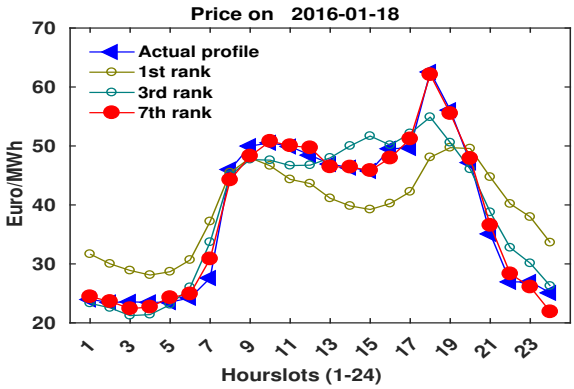


Figure 4.7: Low rank approximation of actual data (one particular day, Jan. 18, blue). Including up to 7 SVD components yields the rank-7 approximation (bold red). Lower rank approximations are also shown.

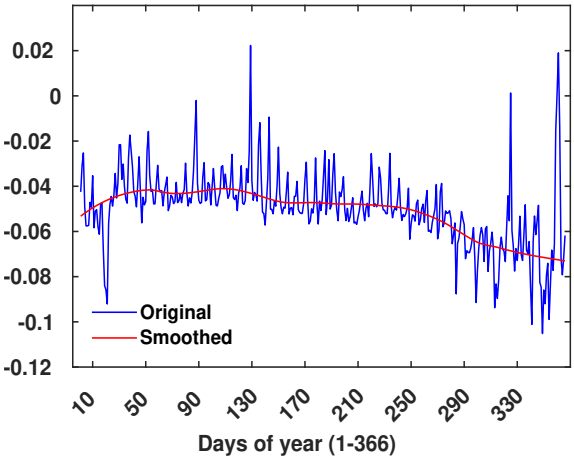


Figure 4.8: An example of V_1 and its smoother version which is used in the reconstruction step.

construction of data, is a way of accounting for outliers while magnifying the underlying patterns as depicted in Figure 4.3. The panels on the left-hand column illustrate the original 2nd derivative profiles for the various quantities of interest (price, solar and wind feed-in, load and traded quantity). The two other adjacent columns display two different lower-rank reconstructions of the same data, after smoothing their V_k columns; a rough rank-1 approximation (middle) and a much more accurate rank-7 approximation.

4.5.3. STRUCTURE-PRESERVING SMOOTHING

In the following section, we propose a straightforward method using SVD factorization to highlight the faint structures seen in Figure 4.5. Recall that the right singular vectors of V_k in the SVD, determine the amplitudes of each corresponding U_k profile for each day. Therefore, smoothing these amplitudes across the year is a way of diminishing most of

the inter-day variations without affecting the overall structure. In the current task, the smoothing was done based on the robust local regression (RLOESS) model, but alternative approaches would be equally valid. We opted for the local regression smoothing method as it alleviates the effect of outliers by assigning a lower weight to them in the regression and also allotting zero weight to data points outside six mean absolute deviations [30]. Therefore, in line with the aforementioned smoothing mechanism, matrix factorization can be reformulated as follow:

$$A_p = \arg \min_{rank(R)=p} (\|A - R\| + \lambda \, \|dV\|_s) \tag{4.27}$$

where $\|.\|_s$ is an outlier-resistant robust local regression model as a smoothing function as in [30].

In the recent applications of SVD, regularization has become an increasing trend. The following section provides an overview of Regularized SVD (RSVD) as the cornerstone of this chapter.

4.6. DATA

An exchange for the next-day power delivery contracts, where the tradings are driven by its participants, is the day-ahead electricity market. Energy trading entities, banks and financial service providers play a prominent role in increasing the liquidity of the wholesale power market [31]. These members are mainly focused on market and trade across borders, even though not necessarily own any power assets. Therefore, grid loss compensation is a great prime for TSOs to intervene in the spot market [32]. Furthermore, regulating feed-in tariff schemes for marketing zero-carbon energy sources is extensively practiced by the TSO in Germany. Consequently, understanding the market during this post-transition era is of great practical value. Figure 4.9 provides an overview of various collected sets of data for the German day-ahead market in 2016. The hourly price and the traded quantity auction values were collected from [33]. We obtained the quarterly (every 15 min) day-ahead solar and wind feed-in forecast data from [34]. In the current work, the hourly values are used instead-obtained by summing up every four quarterly values. The European Network of Transmission System Operators (ENTSO-E) was the platform for downloading the day-ahead load forecast data [35]. Using the aforementioned data, we will explore how the intra-day dynamics of day-ahead hourly price values can be affected by the fluctuations of other attributes.

As mentioned previously, an alternative way of visualizing quasi-periodic time series (with diurnal patterns) is as matrices [29]. To do so, we partition the data into diurnal periods and place each daily cycle in a column to form a matrix. Figure 4.10 illustrates the alternative representation of the data in Figure 4.9, obtained by recasting each time series into a matrix of size 24×366 (recall that 2016 is a leap year). This change of viewpoint has two important applications: 1) Representing the time series as images allow one to visually integrate patterns across longer time spans, hence improving the discriminatory power. For instance, in the bottom panel of Figure 4.10, a notable correlation between the traded quantity and the solar (eye-like horizontal shape seen in the 2nd panel from the top) as well as the wind (vertical stripes in the 3rd panel from the top) can be spotted. In the following section, we will elaborate on this and put these visual impressions

on a more sound, mathematical footing. 2) Recasting such time series as matrices also suggest drawing on matrix decomposition theorems to construct approximations which are more tightly linked to the structure of the time series.

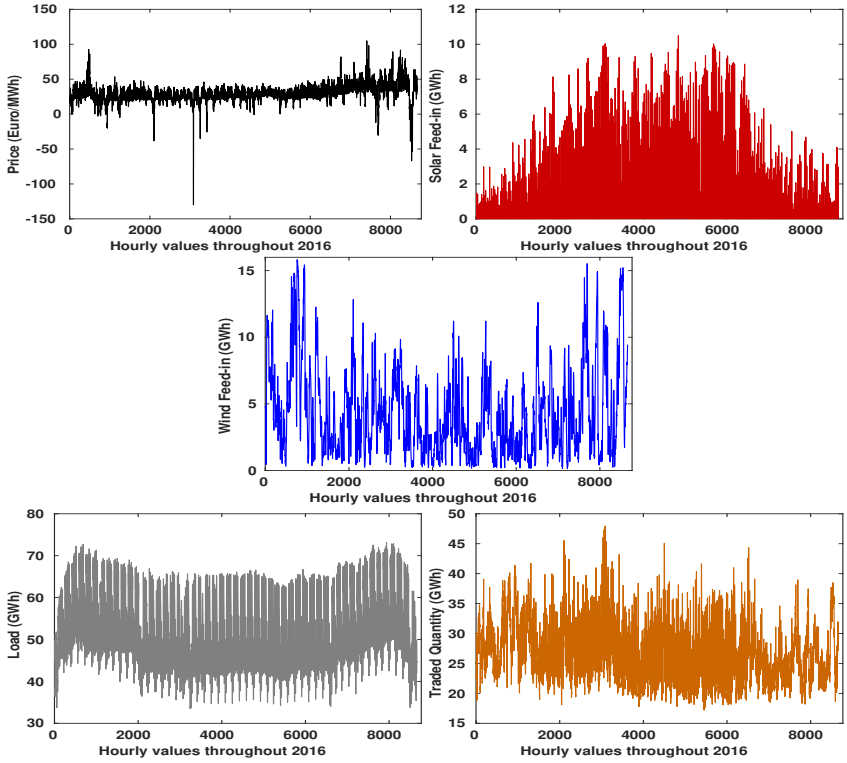


Figure 4.9: An overview of the German day-ahead market in 2016; each data point represents one hour slot. From top to bottom, the price, solar, wind, load, and traded quantity.

4.7. BACKGROUND AND LITERATURE REVIEW

In Europe, Germany is forerunning the others in the switch from conventional energies to renewables, namely wind and solar. This poses new challenges as wind and solar energy are inherently intermittent and to some extent, unpredictable. Therefore, it is of great interest to investigate what effect these changes can have on the day-ahead electricity market and the price volatility. Denny et al. [36] studied how network expansion and increasing the interconnection between Great Britain and Ireland has facilitated the integration of wind farms into the power system. Their simulation results point to a reduction in average price and its volatility in Ireland; which was considered an outcome of the large increase in the interconnection capacities. Furthermore, the high penetration of intermittent distributed energy sources enforces the transmission grid extensions and increases the cross-border interconnections capacities, to ensure grid stability. The viability of this methodology and its upshots is investigated in [37], using the projection

of the wind and solar data until 2020. Some benefits of the substantial expansion of photovoltaic (PV) installations in Germany and Italy, especially, their role in daytime peak price drop are discussed in [38]. Continuing further with the studies on the influence of renewable energy sources (RES) in Germany, a preliminary study was done in [39]. The authors argued about the recent emergence of zero or negative prices on the German day-ahead market as a piece of convincing evidence for the impact of RES. Inspired by the work in [40], the goal of our work is to determine how the intra-day price variability can be influenced by the variability of the wind and solar feed-in. To this end, the intra-day dynamics of different attributes are characterized by their second derivatives as they peak for sharp trend reversals (Section 4.5.1).

As mentioned before, in the present work, we will focus on matrices, as an alternative representation of the quasi-periodic time series data. The matrix interpretation also grants elucidation of the underlying structures in the data, using various matrix decomposition techniques. The concept of decomposing a signal by its constituent components has been addressed by researchers in the past. In [41], the singular value decomposition (SVD) technique is used to obtain the constituent periodic components of a signal. In this work, the singular value ratio (SVR) criterion was also introduced to detect periodicity and determine the period length. Furthermore, reconstructing the *principle patterns* of a given signal was another outcome of this approach. The applicability of SVD in generalized discriminant analysis is investigated in [42]. Detection and extraction of the periodic impulse components from the vibration signals with small signal-to-noise ratios (SNRs), using SVD, is investigated in [43]. Although the alternative representation of the time series has been deployed in different areas before, to the best of our knowledge, it has never been considered in the context of the energy market analysis. The present work looks into this alternative representation and also the applicability of the SVD in day-ahead electricity market analysis.

4.8. CONCLUSIONS AND FUTURE RESEARCH

As mentioned before, we use the proposed SVD-based method in the context of the day-ahead electricity market to quantify the relation between the variability of the day-ahead price values with respect to other attributes. We hence proceed as follows:

1. We compute the intra-day 2nd derivative of the daily profiles (each column of a matrix), in order to highlight peaks and valleys (concavity/convexity is a notion of intra-day variability as can be seen on the left side of Figure 4.3).

2. Next, we lower the rank of those resulting images and enhance their underlying patterns, using a smoothed SVD expansion up to rank 7. Clearly, the rank-7 reconstruction yields an acceptable approximation of the original images on the left. The resulting data are called C_p, C_l, C_q, C_s and C_w where the subscript refers to the corresponding quantity (the rightmost column of Figure 4.3).

3. Finally, a linear regression model is used to quantify the relation between the price volatility C_p with other attributes:

$$C_p = \alpha_0 + \alpha_l C_l + \alpha_q C_q + \alpha_s C_s + \alpha_w C_w \tag{4.28}$$

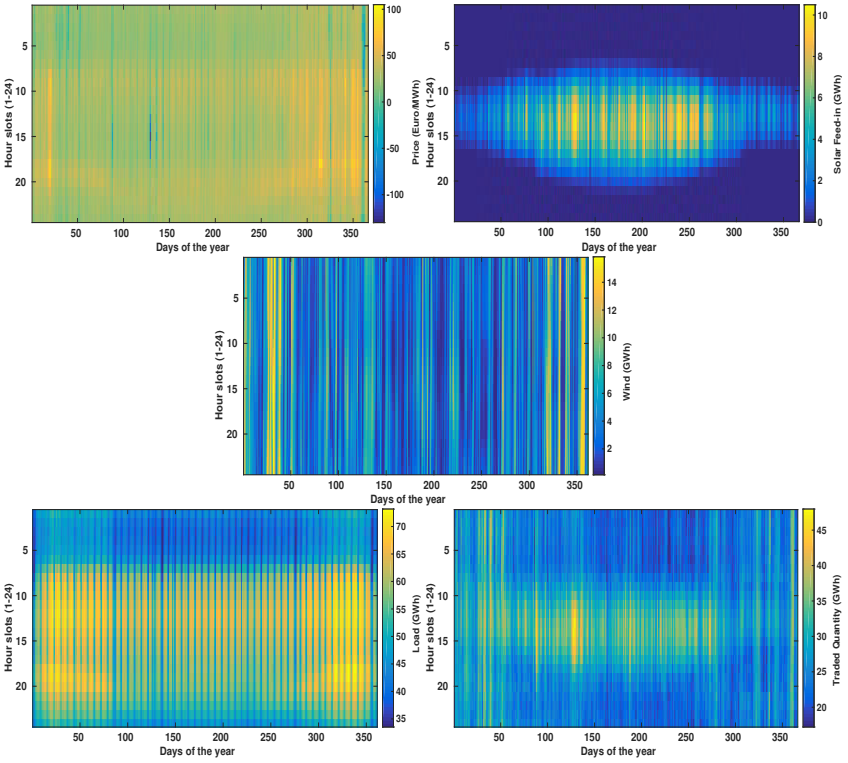


Figure 4.10: An alternative representation of the German day-ahead market by reformatting the time series in Figure 4.9 into matrices of size 24×366. Each column contains the 24 hourly values of a single day. From top to bottom, the price, solar, wind, load, and the traded quantity.

The regression analyses have been performed for three scenarios on the original (untreated) 2nd derivative data: 1) using all data; 2) using only day-time data, and 3) using only night-time data. Table 4.1 contains the results of this benchmark case where the original 2nd derivative matrices (leftmost columns in Figure 4.3) were used. Table 4.2 reflects a significant improvement of the results by using the rank-7 reconstructed matrices (rightmost column in Figure 4.3).

The significance of our results in the latter case (Table 4.2) was investigated using permutation tests where the days of the year have been shuffled, before applying the regression models. Figure 4.11 demonstrates the histogram of the results of the aforementioned randomized data along with the results of the three different scenarios in Table 4.2. The distinct high R^2 values for *24h* and *daytime* scenarios confirm a good performance of the model and show that the intra-day dynamics of the price profiles can indeed be modelled as a function of the concavity (intra-day dynamics) of other attributes. We are aware of the fact that the market mechanism is quite complex and a comprehensive study on the effects of the RES on the market demands takes into consideration of all the playing factors such as energy policy, subsidies and so on. From a data analytics point of view, however, a number of findings can be listed. The intra-day dynamics

Table 4.1: The initial regression model (before applying SVD-based technique).

TimeSlot	R^2	α_l	α_q	α_s	α_w
24h	47.28	1.26	0.44	-2.98	-1.98
day time	53.07	1.34	0.44	-2.97	-1.83
night time	15.60	0.91	0.24	N/A	-2.05

Table 4.2: Regression model applied on rank-7 reconstruction of data.

TimeSlot	R^2	α_l	α_q	α_s	α_w
24h	81.84	1.12	0.59	-2.83	-3.60
day time	86.27	1.26	0.29	-2.36	-1.27
night time	56.40	0.85	0.09	N/A	-17.59

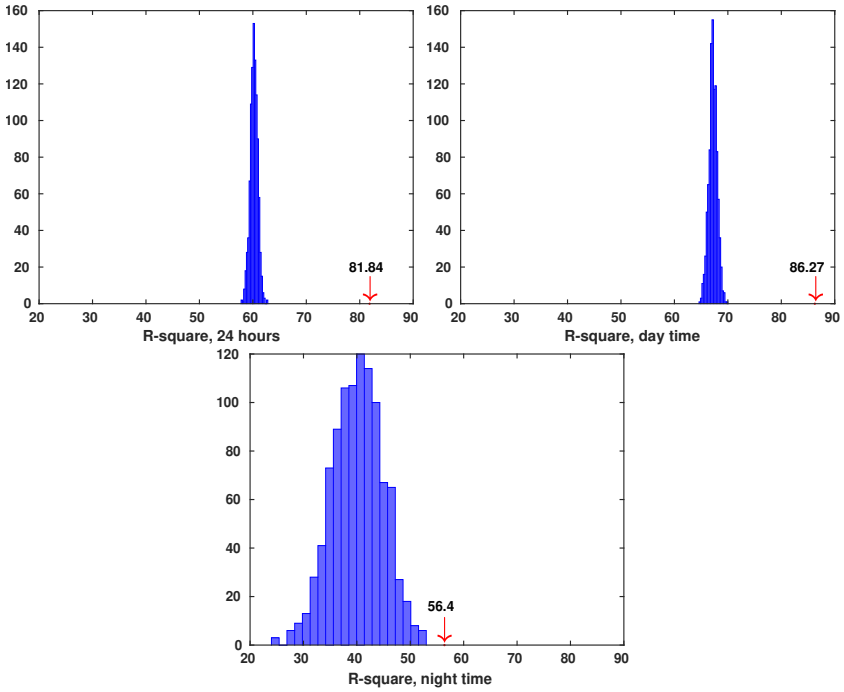


Figure 4.11: Permutation test to indicate the significance of R^2 values in all three scenarios which can be found in Table 4.2. Each histogram is the result of randomization tests, repeated 1000 times (no. of bins = 20), where the days are shuffled before applying the regression models.

(concavity) of the price is least affected by the traded quantity on the day-ahead market. Moreover, RES have a higher impact on the price dynamics than the load. During the day time, solar is the dominant attribute affecting the price dynamics, whereas, during night hours, it is the wind that affects the most. In is worth noting that in the latter case, the low R^2 value urges more extensive research to understand the time price dynamics more satisfactorily.

Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are important matrix factorisation techniques that underpin numerous applications. However, it is well-known that disturbances in the input (noise, outliers or missing values) have a significant effect on the outcome. For that reason, we have investigated regularisation in two different but related versions of the factorisation, and have detailed the solution algorithms.

An important topic for further research would be to find ways in which the gradient descent procedure in Algorithms 1 and 2 can be accelerated by taking advantage of the fact that the functional is very smooth and locally approximately quadratic. It would also be useful to derive some estimates for appropriate values for the weights λ and μ in terms of noise characteristics corrupting the underlying signal. Finally, although the P matrix in Algorithm 2 has unit-length columns, we were not able to prove that these columns are also orthogonal ($P^T P = I_k$) as is the case in standard SVD. In fact, numerical experiments seem to indicate that such a constraint is not compatible with the minimisation of the functional. This requires further theoretical elucidation.

REFERENCES

[1] A. Khoshrou, A. B. Dorsman, and E. J. Pauwels, *Svd-based visualisation and approximation for time series data in smart energy systems,* in *Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2017 IEEE PES* (IEEE, 2017) pp. 1–6.

[2] A. Khoshrou and E. J. Pauwels, *Regularisation for pca-and svd-type matrix factorisations,* arXiv preprint arXiv:2106.12955 (2021).

[3] J. J. Gerbrands, *On the relationships between svd, klt and pca,* Pattern recognition **14**, 375 (1981).

[4] M. A. Davenport and J. Romberg, *An overview of low-rank matrix recovery from incomplete observations,* IEEE Journal of Selected Topics in Signal Processing **10**, 608 (2016).

[5] I. Tošić and P. Frossard, *Dictionary learning,* IEEE Signal Processing Magazine **28**, 27 (2011).

[6] A. Khoshrou and E. J. Pauwels, *Data-driven pattern identification and outlier detection in time series,* in *Science and Information Conference* (Springer, 2018) pp. 471–484.

[7] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Implicit regularization in matrix factorization,* Advances in Neural Information Processing Systems **30** (2017).

[8] J. P. Brooks, J. H. Dulá, and E. L. Boone, *A pure l1-norm principal component analysis,* Computational statistics & data analysis **61**, 83 (2013).

[9] N. Kwak, *Principal component analysis by l_{p}-norm maximization,* IEEE Transactions on Cybernetics **44**, 594 (2013).

[10] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?* Journal of the ACM (JACM) **58**, 1 (2011).

[11] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, *Stable principal component pursuit,* in *2010 IEEE international symposium on information theory* (IEEE, 2010) pp. 1518–1522.

[12] H. Shen and J. Z. Huang, *Sparse principal component analysis via regularized low rank matrix approximation,* Journal of multivariate analysis **99**, 1015 (2008).

[13] B. Dumitrescu and P. Irofti, *Regularized k-svd,* IEEE Signal Processing Letters **24**, 309 (2017).

[14] T. Jin, J. Yu, J. You, K. Zeng, C. Li, and Z. Yu, *Low-rank matrix factorization with multiple hypergraph regularizer,* Pattern Recognition **48**, 1011 (2015).

[15] J. He, Y. Bi, B. Liu, and Z. Zeng, *Graph-dual laplacian principal component analysis,* Journal of Ambient Intelligence and Humanized Computing **10**, 3249 (2019).

[16] M. Yin, J. Gao, Z. Lin, Q. Shi, and Y. Guo, *Dual graph regularized latent low-rank representation for subspace clustering,* IEEE Transactions on Image Processing **24**, 4918 (2015).

[17] N. Shahid, N. Perraudin, V. Kalofolias, G. Puy, and P. Vandergheynst, *Fast robust pca on graphs,* IEEE Journal of Selected Topics in Signal Processing **10**, 740 (2016).

[18] *code for pca-type regularisation,* https://www.dropbox.com/s/fjtz7fh1hyf9cdm/theorem_4.m?dl=0, created: 2021, June.

[19] *code for special case:$\mu = 0$ and $\lambda = 0$,* https://www.dropbox.com/s/ngjksurfepn8dml/special_case_mu_0_lambda_0.m?dl=0 (), created: 2021, June.

[20] *code for special case: $\mu = 0$ and $d = i_n$,* https://www.dropbox.com/s/ab1rfiquiyuzuvz/special_case_mu_0_D_In.m?dl=0 (), created: 2021, June.

[21] *code for factorisation svd-type theorems,* https://www.dropbox.com/sh/f257tzsuttbp1ro/AABaJc1IVXZFQFVQKnpIGjr7a?dl=0 (), created: 2021, June.

[22] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods,* Acta numerica **9**, 215 (2000).

[23] O. King, *On subgroups of the special linear group containing the special orthogonal group,* Journal of Algebra **96**, 178 (1985).

[24] M. Gavish and D. L. Donoho, *The optimal hard threshold for singular values is* $4/\sqrt{3}$, IEEE Transactions on Information Theory **60**, 5040 (2014).

[25] *code for the numerical experiments section,* https://www.dropbox.com/sh/tcl7lag80cimibw/AAD3QNx8FSTOX-c3-wtAh-UFa?dl=0, created: 2021, June.

[26] *Structural similarity index,* http://nl.mathworks.com/help/images/ref/ssim.html ().

[27] *The royal netherlands meteorological institute,* http://www.knmi.nl/nederland-nu/weer/verwachtingen.

[28] G. H. Golub and C. Reinsch, *Singular value decomposition and least squares solutions,* Numerische mathematik **14**, 403 (1970).

[29] P. P. Kanjilal and S. Palit, *The singular value decomposition—applied in the modelling and prediction of quasi-periodic processes,* Signal processing **35**, 257 (1994).

[30] *Smooth response data,* https://nl.mathworks.com/help/curvefit/smooth.html ().

[31] *Epexspot, european power exchange,,* http://www.epexspot.com/en/market-coupling ().

[32] *Energy union and climate,* https://ec.europa.eu/commission/priorities/energy-union-and-climate_en.

[33] *Epexspot, day-ahead auction,* https://www.epexspot.com/en/product-info/auction/germany-austria ().

[34] *Taking power further,* https://www.tennettso.de/site/en/Transparency/publications/overview.

[35] *Entso-e, the european network of transmission system operators,* https://www.entsoe.eu/Pages/default.aspx.

[36] E. Denny, A. Tuohy, P. Meibom, A. Keane, D. Flynn, A. Mullane, and M. O'malley, *The impact of increased interconnection on electricity systems with large penetrations of wind generation: A case study of ireland and great britain,* Energy Policy **38**, 6946 (2010).

[37] K. Schaber, F. Steinke, and T. Hamacher, *Transmission grid extensions for the integration of variable renewable energies in europe: Who benefits where?* Energy Policy **43**, 123 (2012).

[38] K. Barnham, K. Knorr, and M. Mazzer, *Benefits of photovoltaic power in supplying national electricity demand,* Energy Policy **54**, 385 (2013).

[39] N. Adaduldah, A. Dorsman, G. J. Franx, and P. Pottuijt, *The influence of renewables on the german day ahead electricity prices,* in *Perspectives on Energy Risk* (Springer, 2014) pp. 165–182.

[40] L. Hirth, *The market value of variable renewables: The effect of solar wind power variability on their relative price,* Energy economics **38**, 218 (2013).

[41] P. P. Kanjilal and S. Palit, *On multiple pattern extraction using singular value decomposition,* IEEE transactions on signal processing **43**, 1536 (1995).

[42] P. Howland and H. Park, *Generalizing discriminant analysis using the generalized singular value decomposition,* IEEE transactions on pattern analysis and machine intelligence **26**, 995 (2004).

[43] L. H. L. J. Z. Ying and Q. Liangsheng, *Improved singular value decomposition technique for detecting and extracting periodic impulse component in a vibration signal,* Chinese Journal of Mechanical Engineering **17**, 1 (2004).

5

HYPOTHESIS GENERATION USING SVD

5.1. INTRODUCTION

Scenario-based probabilistic forecasting models have been extensively explored in the literature in recent years. A particular application of such models is in the energy sector, where e.g., having the distribution of the energy consumption for the coming days is desired. In this chapter, we put the applicability of the SVD into practice to tackle the energy forecasting problem.

A decisive variable in predicting the energy demand is the temperature data [2]. In this chapter, we propose a generic, data-driven and computationally efficient SVD-based approach to simulate temperature scenarios. The generated temperature profiles, along with other variables, are then fed into a regression algorithm to obtain a probabilistic forecast of the electricity consumption profiles (see Section 5.3).

There are mainly three practical and popular methods for generating temperature scenarios, namely fixed-date, shifted-date, and bootstrap approaches [3]. Nevertheless, these methods have mostly been used on an ad-hoc basis without being formally compared or quantitatively evaluated. As mentioned before, the predictive power of probabilistic forecasting models depends a great deal on how the methodology used in generating temperature scenarios is robust and capable of simulating the temperature data sensibly. An important distinction of the current work is the use of matrices as an alternative representation of the data. The singular value decomposition (SVD) technique is then used to generate temperature scenarios, in a robust and data-driven manner.

The strength of our proposed method lies in its simplicity and robustness, in terms of the training window size, with no need for subsetting or thresholding to generate temperature scenarios. Furthermore, to systematically account for the non-linear interactions between different variables, a new set of features is defined: the first and second derivatives of the predictors. The empirical case studies performed on the data from the

Part of this chapter has been published in [1, 2].

load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L) shows that the proposed method outperforms the top two scenario-based models with a similar setup.

The rest of this chapter is organized as follows. Section 5.2 provides a brief introduction to the data from the load forecasting track of the Global Energy Forecasting Competition [4]. Section 5.3 is devoted to our methodology. A short description of the Gradient Boosting method is presented, first, followed by our proposed models for point forecasting. After a brief recapitulation of the SVD, we explain how our proposed scenario-based load forecasting models work. The proposed method in this chapter is in fact a marriage between an SVD-based temperature scenario generator and an ensemble of trees (gradient boosting algorithm). The experimental results along with a comparison with the results of a number of benchmark models are presented in Section 5.4. We conclude this chapter in Section 5.5.

5.2. DATA

For the sake of replicability and comparability of our work with the benchmark models, a case study is constructed based on the data from the load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L) [4]. The participants in that competition had been asked to develop a short-term probabilistic load forecasting model, with a forecasting horizon of one month. The publicly available data set consists of hourly temperature values from 25 anonymous weather stations and the aggregated hourly load profiles; for a detailed description of data and the competition instructions see [5].

The electricity consumption patterns are subject to a variety of factors, such as meteorological conditions, calendar information, season, working schedules, energy cost and economic activities [6]. In the current work, however, consistent with the requirements in the GEFCom2014-L, only temperature and calendar information are taken into consideration as the available predictors (besides historic load profiles).

Temperature is believed to be a major driving force behind electricity demand. The non-linear effect of the temperature on electricity demand is hence at the centre of our attention. The left panel of Figure 5.1 provides an overview of the typical electricity consumption profiles throughout the week. This figure affirms that the consumption patterns differ notably during the weekends from the weekdays. Interestingly enough, on Friday afternoons, the demand profile gets closer to the weekends, whereas, during working hours, it is akin to other working days. Furthermore, the right panel in Figure 5.1, illustrates the evolution of daily load profiles in the year 2010; this figure was obtained by recasting the time series into a 24×365 matrix, where every column contains 24 hourly values for each daily profile [7]. As expected, in spring and fall, when the temperature is moderate, electricity demand tends to be lower than at any other time of the year (winter and summer times). It underscores the fact that electricity demand is driven by climate conditions (e.g., air conditioning usage), and also the lifestyle changes followed by that. This figure also highlights the non-linear relation between load and temperature throughout the year.

In the literature, the temperature is arguably the most dominant predictor of the load; however, in and of itself, it is not sufficient for two main reasons:

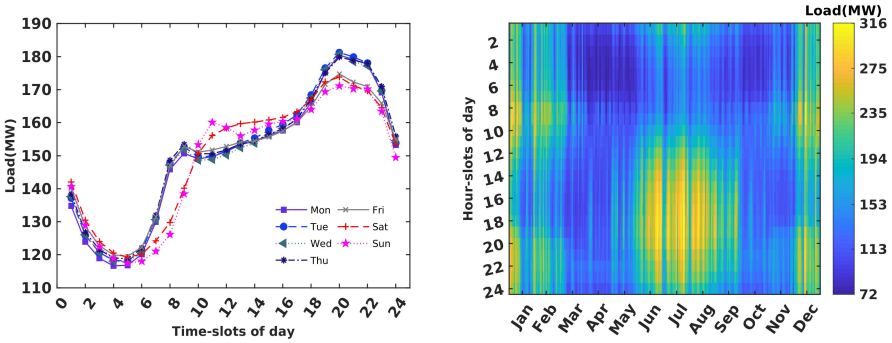


Figure 5.1: **Left:** Comparison of typical daily consumption patterns during the week. **Right:** An overall representation of the evolution of the load profiles throughout a year.

- **Diurnal human activities:** As it is seen in the left panel of Figure 5.1, the typical electricity demand behaviour changes throughout the week. These diurnal human activities are plainly not reflected in the temperature data.

- **Recency and cross effects:** Even for similar days (weekends or weekdays), the trend of daily load profiles, corresponding to similar temperature data, might not be necessarily alike; the recency and cross effects can play a vital role. For instance, the rise of temperature in early spring might not necessarily lead to high electricity consumption, in comparison with summer times, as people might appreciate the rise in outside temperature after a cold winter. This, of course, can deviate across different seasons. Figure 5.2 illustrates an overview of the trend (first derivative) changes in daily temperature and load profiles in 2010. These figures were obtained by taking the first derivatives of daily temperature and load matrices. It is seen that e.g. in early spring and summer time, with the rise of temperature, afternoon peak profiles start to disappear. Although, the overall relationship between load and temperature is clear; it is, however, non-trivial how to robustly address the non-linear effect between temperature and load profiles. Experiment results in [2] affirm that including the 1st and 2nd derivatives of the daily profiles can indeed enhance the performance of the forecasting model.

5.3. METHODOLOGY

In the present work, we opt to use an ensemble of regression trees (Gradient Boosting method) to predict day-ahead load prognoses, with a forecasting horizon of one month, given hourly temperature profiles, historical (or estimated) load profiles, and calendar information. After a brief recapitulation of an ensemble of regression trees, the detailed explanation of our methodology for the probabilistic Short Term Load Forecasting (STLF) problem is in turn provided below.

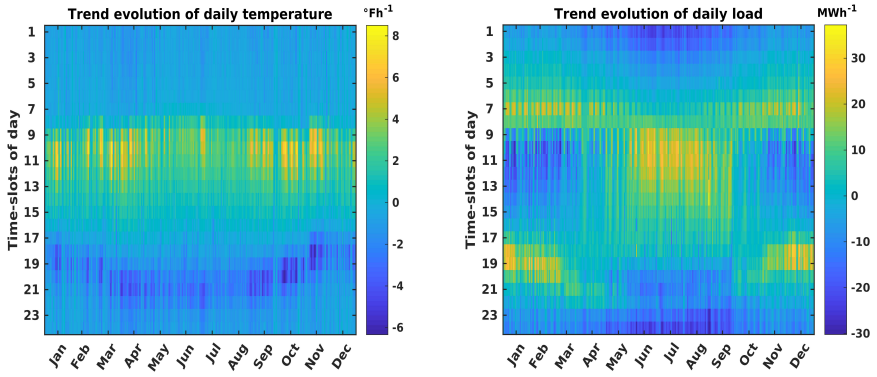


Figure 5.2: An overview of the evolution of the first derivative of the temperature (**Left**) and load (**Right**) profiles.

5.3.1. ENSEMBLE OF REGRESSION TREES

The use of "ensemble learning" methods in various classification and regression problems has taken off over the last few years. Ensembles generally rely on "resampling" techniques to obtain different training sets for each individual regression or classification model. Two popular methods for creating accurate ensembles are bootstrap aggregating (Bagging) and Boosting.

In the Bagging method, the training data for each individual model is drawn randomly, i.e., n instances with replacement-where n are the number of observations in the training set. In this approach, successive members (e.g., trees or neural networks) are independent of each other; since each member of the ensemble is trained individually using a bootstrap sample of the data set [8]. In other words, Bagging methods control the generalization error through perturbation and averaging of sub-models.

In the present work, we opt to use an ensemble of trees which is fast to train and provides more insight into the importance of the predictors. Worth noting that in this approach, to ensure that every training sample is predicted at least a few times, the number of trees needs to be large enough. Since the trees are independent of each other, the distribution function and the quantiles of each hourly forecast can be easily computed based on the output of all the trees [9]. It becomes apparent in Section 5.4.

In the Gradient Boosting method, however, the training set for each member of the ensemble depends on the performance of the previous model(s). More precisely, in order to alleviate the error in earlier models, extra weights are assigned to samples with higher prediction error rates; those are hence more likely to take part in the training of the next model [10, 11]. A comprehensive evaluation of both these techniques on 23 data sets, using two popular classifiers, i.e., decision trees and neural networks is presented in [12]. The applicability of the Gradient Boosting method in quantile regression load forecasting applications has been put into practice in [9, 13].

The goal of every typical prediction problem is to determine an estimate or approximation $\hat{\mathbf{F}}(\mathbf{x})$, of the true mapping function $\mathbf{F}^*(\mathbf{x})$ which assigns a $y \in \mathbb{R}$ to any given set of covariates $\mathbf{x} \in \mathbb{R}^p$. This process is optimized by minimizing the expected value of some

specified loss function $L(y,\mathbf{F}(\mathbf{x}))$ over the set of the joint distribution of all $\{y,\mathbf{x}\}$ pairs. In mathematical parlance, we have:

$$\mathbf{F}^*(\mathbf{x}) = \underset{\mathbf{F}(\mathbf{x})}{\arg\min}\, \mathbb{E}_{y,\mathbf{x}} L(y,\mathbf{F}(\mathbf{x})) = \underset{\mathbf{F}(\mathbf{x})}{\arg\min}\, \mathbb{E}_{\mathbf{x}}[\mathbb{E}_y(L(y,\mathbf{F}(\mathbf{x})))|\mathbf{x}] \tag{5.1}$$

where $\mathbb{E}(.)$ is the expectation operator, and $L(y,\mathbf{F}(\mathbf{x}))$ is a *loss* function, e.g., the popular choice of squared-error $\{y-\mathbf{F}(\mathbf{x})\}^2$, for regression problems. $\mathbf{F}(\mathbf{x})$ is a member of "additive" class of functions of the form:

$$\mathbf{F}(\mathbf{x};\{\lambda_k,\mathbf{a}_k\}_1^K) = \sum_{k=1}^{K} \lambda_k h(\mathbf{x};\mathbf{a}_k). \tag{5.2}$$

where K is the number of members of the ensemble model, λ_k is the coefficient of the additive model, the generic function $h(\mathbf{x};\mathbf{a})$ in Eq. (5.2) is called a *weak learner* or *base learner*-it is usually a simple parameterized function of the explanatory variables, specified by parameters $\mathbf{a}=\{a_1,a_2,\ldots\}$. In the present work, each $h(\mathbf{x};\mathbf{a}_k)$ is a small regression

Algorithm 3: The Gradient Boosting Algorithm, with a squared-error loss function.

Initialization: $\mathbf{F_0}(\mathbf{x}) = \bar{y}$

for *k=1 to K* **do**

$\tilde{y}_i = y_i - \mathbf{F}_{k-1}(\mathbf{x}_i), \quad i \in [1:N]$

$(\rho_k,\mathbf{a}_k) = \underset{\rho,\mathbf{a}}{\arg\min} \sum_{i=1}^{N} [\tilde{y}_i - \rho\, h(\mathbf{x}_i;\mathbf{a})]^2$

$\mathbf{F}_k(\mathbf{x}) = \mathbf{F}_{k-1}(\mathbf{x}) + \rho_k h(\mathbf{x};\mathbf{a}_k)$

tree as introduced in [14]. For a regression tree, the parameters $\mathbf{a}_k$ are the splitting variables, split locations and means of the terminal node of the individual trees. An overview of the Gradient Boosting algorithm, with a squared-error loss function, is presented in Algorithm 3, where the multiplier ρ_k is given by the line search:

$$\rho_k = \arg\min_{\rho} \mathbb{E}_{y,\mathbf{x}} L(y,\mathbf{F}_{k-1}(\mathbf{x}) - \rho \mathfrak{g}_k(\mathbf{x})), \tag{5.3}$$

and

$$\mathfrak{g}_k(\mathbf{x}) = \mathbb{E}_y \left[\frac{\partial \mathbf{L}(y,\mathbf{F}(\mathbf{x}))}{\partial \mathbf{F}(\mathbf{x})}\right]_{\mathbf{F}(\mathbf{x})=\mathbf{F}_{k-1}(\mathbf{x})} \tag{5.4}$$

It is worth to be mentioned that the rationale underpinning the choice of the ensemble of the trees is their ability to better handle the heterogeneous input data–which comprise both continuous and discrete variables. Additionally, tree models are effective, adaptive and modular, in that new predictors can be easily added. It is perceived that ensemble models, unlike their other ML-based counterparts, are less prone to overfitting; they promise to strike a good trade-off between bias and variance [15].

5.3.2. OUR PROPOSED FORECASTING MODELS

Generally speaking, a regression tree is an adaptive nearest neighbours-like algorithm. However, it usually shows a better performance in comparison with other counterparts

Attribute	*Description*	*Model no.*
mn	month of year: 1,2,...,12	I,II
wk	day of week: 1,2,...,7	I,II
hr	hour of day: 1,2,...,24	I,II
$\mathbf{L(d-1)}$	previous day (estimated) hourly load	I,II
$\mathbf{L(d-7)}$	previous week (estimated) hourly load	I,II
$\mathbf{T(d)}$	hourly temperature (generated profiles)	I,II
$\mathbf{L'(d-1)}$	1st derivative of $\mathbf{L(d-1)}$	II
$\mathbf{L''(d-1)}$	2nd derivative of $\mathbf{L(d-1)}$	II
$\mathbf{L'(d-7)}$	1st derivative of $\mathbf{L(d-7)}$	II
$\mathbf{L''(d-7)}$	2nd derivative of $\mathbf{L(d-7)}$	II
$\mathbf{T'(d)}$	1st derivative of $\mathbf{T(d)}$	II
$\mathbf{T''(d)}$	2nd derivative of $\mathbf{T(d)}$	II
$\mathbf{L(d)}$	hourly load (forecast target)	I,II

Table 5.1: An overview of the attributes used in our proposed models.

nearest neighbour-based methods. It tends to find the homogeneous portions of the sampling space locally, on contrary to other conventional methods which incline to treat all distances equally [16, 17].

In the present work, we follow a homogeneous forecast combination framework, i.e., we first train a single-value load forecasting model, then, vary the input data (different temperature scenarios) to obtain a series of forecasts, and accordingly, the quantiles (Section 5.4). We consider two Gradient Boosting based methods to predict the day-ahead load prognoses. In the first model, only calendar information, temperature data along with historical load data are the input variables. We proceed further in the second model to incorporate the daily dynamics of the temperature and load profiles. It is done using the first and second derivatives of the daily profiles.

Power consumption is subject to a wide range of exogenous variables, including calendar effects, electricity price and so on. In the literature, the previous consumption patterns and calendar information have been extensively used in developing various load forecasting models. However, accounting for the interaction between different variables, namely the recency and cross effects can be an onerous task; it demands some domain expertise to be done sensibly [18, 19].

A number of common deterministic (categorical) explanatory variables used in our methodologies are as follows: month of the year $\mathbf{mn} \in \{1,2,\ldots,12\}$, day of the week $\mathbf{wk} \in \{1,2,\ldots,7\}$ (starting from Sunday=1), and hour of the day $\mathbf{hr} \in \{1,2,\ldots,24\}$. As it is principled in [5] the forecasting horizon herein is one month, therefore, the estimated values for the first week of the month are being used to estimate the load profiles in the second week of the month and so on. In all cases, the aim is to predict $\mathbf{L(d)}$, 24 hourly load values for the target day $\mathbf{d}$. Below we discuss the models in more detail, but for ease of reference, Table 5.1 summarizes all the common and distinctive attributes used in the two proposed models.

Model I The first model provides us with a benchmark to measure the credibility of our proposed method in incorporating the recency and cross effects in data in Model II. Here, we introduce six different attributes to predict the hourly load values on the target day **d**. The three above-mentioned common discrete (categorical) values, namely, **mn**, **wk** and **hr**, along with the (estimated) load value for a given hour on the previous day or week (**L**(**d** − **1**) and **L**(**d** − **7**), respectively). The intuition for this choice is to reflect the diurnal and weekly patterns of human activities on electricity consumption (Figure 5.1). The last covariate **T**(**d**) is the hourly temperature forecast for the target day **d**. As it is explained in Section 5.3.3, we generate a hundred independent temperature profiles, using the singular value decomposition, to correspondingly obtain 100 independent load forecasts for each target day; the combination of these forecasts is then used to obtain the load quantiles for each hour.

Model II To reflect the lagging effect of temperature on load changes, in the second model, we add the daily dynamics of the temperature and load profiles (1st and 2nd derivatives) [1]. For a given hour slot h the corresponding first derivative of the variable $z \in \{\mathbf{L}, \mathbf{T}\}$ can be obtained by $z'(h) = 0.5[z(h+1) - z(h-1)]$; with obvious analogues for the 2nd derivative. The reasoning for doing so is that oftentimes the actual values are not as important as the general underlying trends captured by the first or second derivatives of the covariates. In other words, load value at any moment is influenced by the variations of the other attributes (namely, temperature profiles) prior to that moment. Including the derivatives are, in fact, a relatively simple and generic means to account for the recency effect in the data. In comparison with most time-varying models, where the data is typically divided into subsets (based on thresholds), or a lagging window is optimized, our proposed approach is more straightforward and user-friendly.

A major contribution herein is the use of the SVD to generate temperature scenarios **T**(**d**) for the target day **d**; it is done to determine the distribution (99 percentiles) of the load profile in our proposed probabilistic forecasting models. A brief recapitulation of the singular value decomposition (SVD) technique is provided below.

5.3.3. SINGULAR VALUE DECOMPOSITION

As previously mentioned, the SVD technique is used herein to generate new temperature profiles (matrices). To be more precise, we recast one year's worth of hourly temperature values as a matrix $T \in \mathbb{R}^{24\times 365}$ such that every column corresponds to 24 hourly values of a day. Consequently, the matrix T can conveniently be represented by a low-rank approximation. To review the SVD results see Chapter 2.

If there are only a few dominant singular values (as is the case for the temperature matrices, in Figure 5.3), the expansion of the matrix in Eq. (2.5) can be sufficiently truncated after just the first few K terms to yield A_K, an adequate lower rank approximation of A, as obtained in Eq. (2.10).

To elaborate more, Figure 5.4 illustrates the first three columns of U_k (left) and V_k (right) for temperature matrix for the year 2009. In geometrical terms, U_k columns can be interpreted as the fundamental daily profile and its successive increments; V_k values represent the corresponding scaling factors for each U_k profiles for each day. In other words, SVD decomposes the original time series into a linear combination of a num-

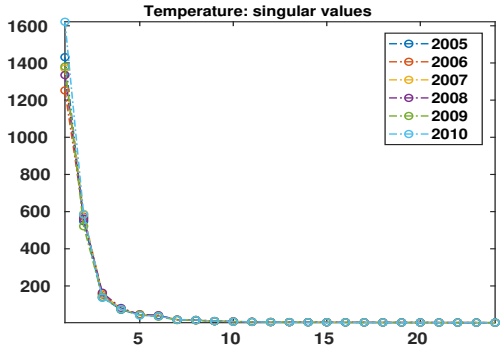


Figure 5.3: The evolution of the singular values of the temperature matrices over the years; it suggests that a reconstruction of rank-4 approximation would suffice, indicating that temperature is quite regular.

ber of (data-driven) orthonormal profiles, specified by U_k columns; each profile is then scaled up (or down) according to their corresponding weights in V_k.

For instance, U_1 in the top left panel of Figure 5.4 strikingly resembles the averaged daily temperature profile. Moreover, its corresponding V_1 profile (top right panel) outlines the evolution of that profile throughout the year; it is in agreement with the fact that temperature is higher in summer time (middle part of the graph). Recall that these profiles are weighted based on the magnitude of their corresponding singular values which are sorted in descending order from left to right (Figure 5.3). The most dominant "corrective" incremental profile U_2 and its corresponding coefficients V_2 are displayed in the middle panel of Figure 5.4. This correction hence needs to be added to the first profile to get a better approximation, i.e., $K = 2$ in Eq. (2.5). Similar interpretations are valid for the third profile (bottom panels) and so on.

It is worth noting that V_k profiles on the right-hand side of Figure 5.4 imply a distinct impression that temperatures are less variable during the summer (middle parts of the graph). In the following Section, SVD is used to simulate pragmatic temperature scenarios, in a systematic and data-driven manner. The generated profiles are accordingly fed to Models I and II to obtain the probability distribution (99 quantiles) of the load values for every given hour.

5.3.4. TEMPERATURE SCENARIO GENERATION

A common approach in probabilistic load forecasting problems is to vary the input (e.g., temperature profiles) to obtain a series of forecasts and combine them [20]. One of the major challenges, however, is how to create realistic temperature profiles, e.g., simply adding independent Gaussian noise to the hourly values of individual temperature curves results in some preposterously jagged profiles.

In the literature, a number of solutions have been proposed to simulate temperature scenarios. In [21], it is proposed to combine different weather station measurements to generate new temperature profiles. Nonetheless, it can be argued that normal weather scenarios cannot precisely be simulated by averaging the temperature profiles, as they tend to underestimate the peaks. Furthermore, such approaches are not resilient toward

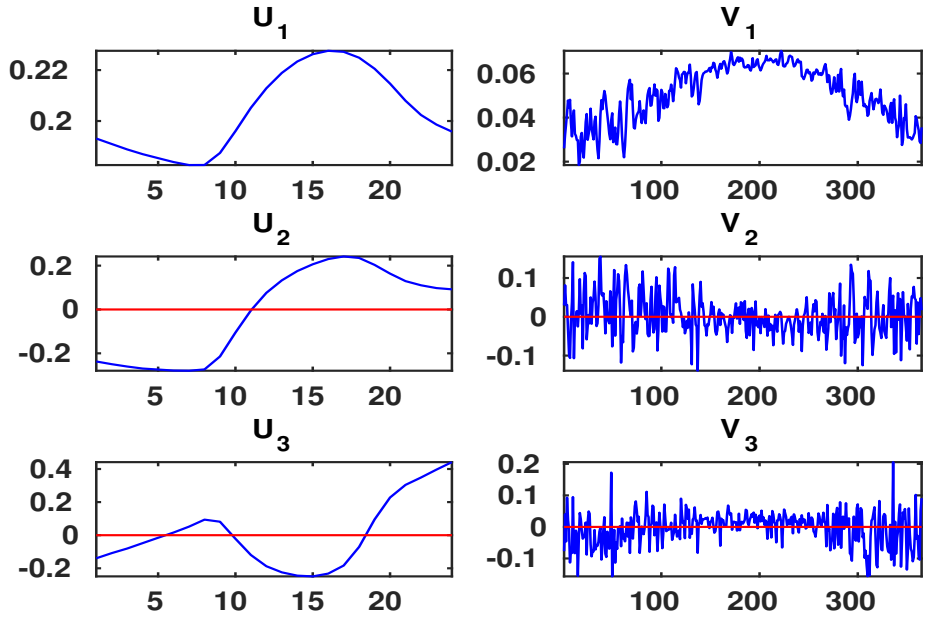


Figure 5.4: SVD-based decomposition of hourly temperature data for 2009. On the left, there are the first three U_k columns; whereas the right column displays their corresponding V_k's.

outliers, as even one instance, can change the whole profile as long as it takes part in generating new temperature scenarios. The performance of the forecasting model can consequently be diminished as a result of that. Worth noting that shifting the temperature data by one, two or even three days was initially used to generate temperature scenarios; it was later abandoned for obvious reasons. Some cumbersome approaches in terms of computational costs, such as Monte Carlo-based methods are also popular, especially among utilities, to simulate thousands of temperature profiles - an approach which is used in scenario analysis in LTLF problems [22]. In [23], new temperature scenarios are generated, again, by averaging the temperature of stations 3 and 9 (GEFCom2014-L data was used). The reason for that is mentioned to be due to the existence of a good in-sample fit with a cubic relation between the temperature records of those two stations and the load data. Besides pre-processing there are not a lot of solutions in the literature on how to generate robust and pragmatic input (temperature) scenarios.

The SVD allows us to create hundreds of sensible and realistic temperature profiles for any target day **d**, in a fairly fast and robust manner. Figure 5.3 affirms the fact that the singular values σ_k of the temperature matrices over the years have not changed much; similar conclusions can be drawn for the left singular vectors U_k. Furthermore, it is plain to see in Figure 5.4 that the V_k coefficients implicate the variability of the temperature profiles throughout the year. Since the forecasting horizon is one month, hereafter temperature matrix is referred to a month worth of temperature data for the coming month (test data in Section 5.4). We, therefore, proceed with the following steps, to create temperature scenarios:

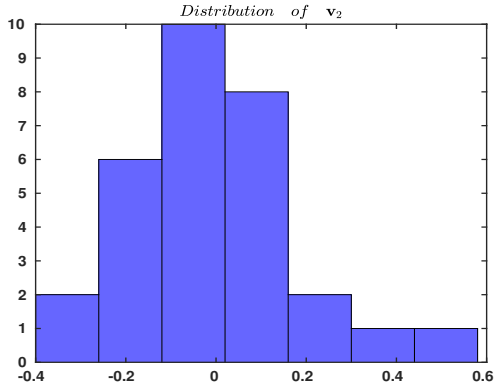


Figure 5.5: Histogram of the $\mathbf{v_2}$ coefficients for June 2011 temperature data (30 values). Note that the distribution is approximately normal with zero mean and $\mathbf{std}(\mathbf{v_2}) \approx \mathbf{0.18}$.

1. In the first step, we estimate the corresponding standard deviation $\mathbf{s}_k$ for a number of right singular vectors V_k (V_2, ..., V_4). Figure 5.5 illustrates the histogram for V_2, which shows that $\mathbf{s}_2 \approx 0.18$. Interestingly enough, similar experiments on all three V_k columns ($k \geq 2$) yield similar results; however, their contribution to the final rank K reconstructed profile is scaled up or down by the magnitude of their corresponding singular values.

2. Next, for any given day $\mathbf{d}$ of the test month, for which a number of temperature scenarios are desired, we take the actual temperature profile for that day $\mathbf{T} = \mathbf{T}(\mathbf{d})$, find the corresponding V_k coefficients (V_2^0, V_3^0, V_4^0); then blend them with zero-mean Gaussian noise: $V_k^n = V_k^0 + \mathcal{N}(0, \epsilon^2)$. These perturbed V_k coefficients are then used to generate a new (noisy) temperature scenario (reconstruct the matrix).

3. According to the scheme outlined above, for each actual daily profile $\mathbf{T}(\mathbf{d})$, a hundred temperature scenarios are generated. This new data set is then fed into the proposed prediction models. In the final step, the forecasts are duly compared to the real load values. This approach enables us to determine the distribution of the hourly load values (99 quantiles) and compute the corresponding pinball error values (Section 5.4). Figure 5.6 illustrates an example of one hundred generated temperature profiles (Left), and their corresponding daily load profiles (Right).

4. For the sake of completeness, it should be noted that V_1 is left unperturbed as this is a proxy of the average temperature on a particular day, for which the uncertainty is negligible. Similarly, there is not much to be gained from perturbing other right singular vectors (V_5 etc), as their impact on the profile is insignificant (their corresponding σ_k are small).

In [1], a preliminary study was done to investigate how the effect of the perturbation variance in temperature profiles $\mathbf{T}(\mathbf{d})$ propagates into uncertainty on the target load profile $\mathbf{L}(\mathbf{d})$.

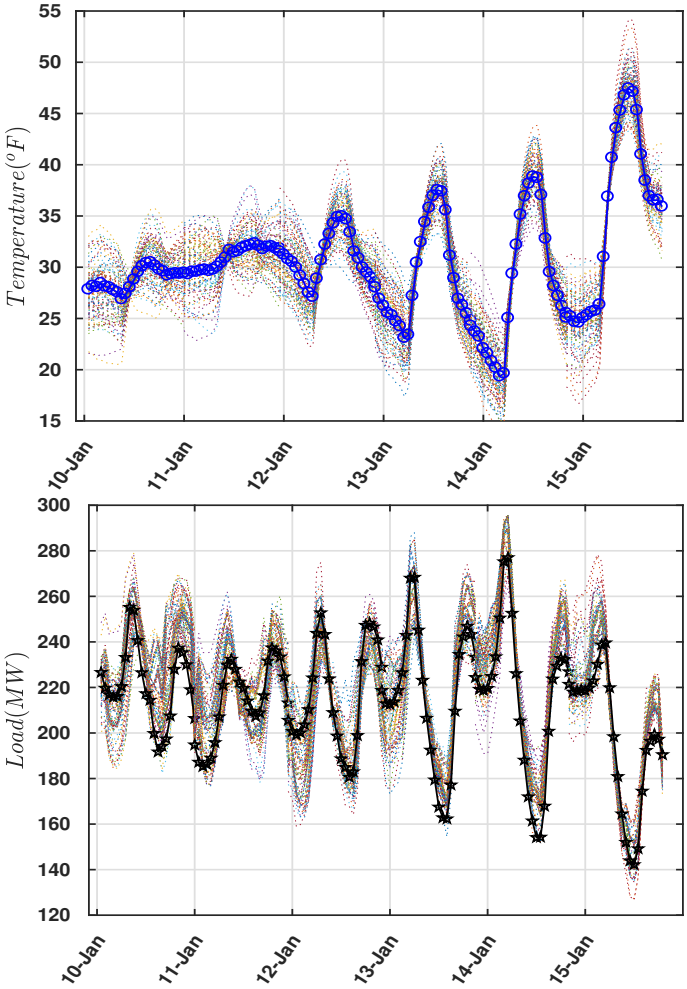


Figure 5.6: An illustrative example of 100 generated temperature scenarios for each day and their corresponding daily load profiles (obtained by Model II) for January 10-15, 2011. The spotted points are the actual values and the noise level is $\mathcal{N}(0, 0.09)$. These 100 different load values for every hour are then used to calculate the 99 quantiles.

5.4. EXPERIMENTAL RESULTS

Probabilistic forecasts provide more comprehensive information about future uncertainties than point forecasts can do [20]. As previously mentioned, the aim of the GEFCom2014-L was to estimate the quantiles of the hourly load values for a utility in the US, on a rolling basis [5]- with a forecasting horizon of one month. Furthermore, it was expected from the contestants to investigate the weather scenario generation methods for probabilistic load forecasting.

The scenario-based probabilistic forecasting methodology proposed in [22] was used by two top 8 teams (Jingrui Xie, top 3; Bidong Liu, top 8) in GEFCom2014-L. Therefore, we opt to compare our results with similar works. A least absolute shrinkage and selection operator (LASSO) estimation based method is proposed in [23] for probabilistic forecasting applications. This work is reported to outperform the methodology used by Bidong Liu [18] to win a top 8 place in GEFCom2014-L. We hence have considered the proposed method in [23] as one of the benchmark models. The reported work uses a bivariate time-varying threshold autoregressive (AR) process for the hourly load $Y_{\mathscr{L},t}$ and temperature $Y_{\mathscr{T},t}$ data ($\mathscr{D} = \{\mathscr{L}, \mathscr{T}\}$). The time series of interest are accordingly modeled for $i \in \mathscr{D}$ as follows:

$$Y_{i,t} = \phi_{i,0}(t) + \sum_{j\in\mathscr{D}} \sum_{c\in C_{i,j}} \sum_{k\in I_{i,j,c}} \phi_{i,j,c,k}(t) \max\{Y_{j,t-k}, c\} + \epsilon_{i,t} \tag{5.5}$$

where $\phi_{i,0}$ are the time-varying intercepts and $\phi_{i,j,c,k}$ are time-varying autoregressive coefficients. Furthermore, $C_{i,j}$ are the set of all considered thresholds for the load and temperature data (all set manually). $I_{i,j,c}$ are the index sets of the corresponding lags and $\epsilon_{i,t}$ is the error term. The modelling process is done in three parts: 1) choice of thresholds $C_{i,j}$; 2) choice of lag sets $I_{i,j,c}$; and 3) time-varying structure of the coefficients. For further details see [23].

Another winning team (top 3) in the GEFCom2014-L was Jingrui Xie, who developed an integrated solution for probabilistic load forecasting [21]. Her proposed methodology consists of three parts: 1) pre-processing, which includes data cleaning and temperature station selection; 2) forecasting (which focuses on the development of point forecasting models), forecast combination, and temperature scenario generation; and 3) post-processing, which embodies the residual simulation for probabilistic forecasting purposes. Inspired by the *Vanilla* model in [18], their core forecasting model is as follows:

$$\begin{aligned}\mathscr{L}_t = \beta_0 + \beta_1 \text{Trend}_t + \beta_2 \mathscr{T}_t + \beta_3 \mathscr{T}_t^2 + \beta_4 \mathscr{T}_t^3 + \beta_5 \text{Month}_t + \beta_6 \text{Weekday}_t + \beta_7 \text{Hours}_t + \\ \beta_8 \text{Hours}_t \text{Weekday}_t + \beta_9 \mathscr{T}_t \text{Month}_t + \beta_{10} \mathscr{T}_t^2 \text{Month}_t + \beta_{11} \mathscr{T}_t^3 \text{Month}_t + \\ \beta_{12} \mathscr{T}_t \text{Hour}_t + \beta_{13} \mathscr{T}_t^2 \text{Hour}_t + \beta_{14} \mathscr{T}_t^3 \text{Hour}_t\end{aligned} \tag{5.6}$$

It is in fact a multiple linear regression (MLR) model with the following main and cross effects:

- Main effects: a chronological Trend_t variable, first to third-order polynomials of the temperature ($\mathscr{T}_t, \mathscr{T}_t^2, \mathscr{T}_t^3$), and a number of categorical variables namely, Month, weekday, and Hour.

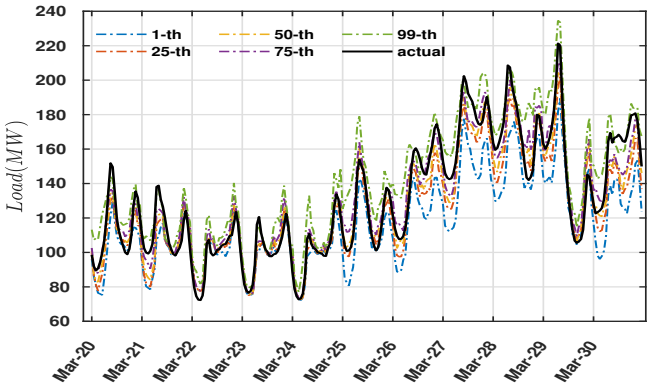


Figure 5.7: Probabilistic load forecast of 11 days from March 20, 2011 to March 30, 2011; the solid line in black is the actual value and the dash-dot lines are the forecast quantiles.

- Cross effects: similar to [18], the cross effects are incorporated using the multiplications of different attributes such as $\text{Hour}_t\text{Weekday}_t$, $\mathcal{T}_t\text{Month}_t$, $\mathcal{T}_t^2\text{Month}_t$, $\mathcal{T}_t^3\text{Month}_t$, $\mathcal{T}_t\text{Hour}_t$, $\mathcal{T}_t^2\text{Hour}_t$, and $\mathcal{T}_t^3\text{Hour}_t$.

In the next step, the residuals obtained from Eq. (5.6) are modeled using four different techniques, namely unobserved component models (UCM), exponential smoothing models (ESM), three-layer feedforward neural network (NN), and autoregressive integrated moving average models (ARIMA). Four different sets of point forecasts are accordingly generated by adding each set of residuals to the values obtained from the previous stage. The average of each four values is the final estimation for the load forecast for every given hour. In the end, 10 different temperature scenarios are generated according to [22], to obtain the 99 percentiles from the 10-point forecasts.

The regression-based models are arguably vulnerable towards outliers, especially in scenario-based applications. Due to the recency effects, outliers e.g., in temperature scenarios can affect the load forecasts for a longer time span. Our proposed SVD-based model is more robust and capable of handling this issue. As mentioned before, for every hour of the target day $\mathbf{L}(\mathbf{d})$, we obtain 100 different load values (see Figure 5.6). The results are then used to determine the distribution of the hourly load values (99 different quantiles) for any given hour, by employing linear extrapolations [24]. An illustrative example of the predicted quantiles for 11 days, March 20-30, 2011, is provided in Figure 5.7.

Model Evaluation: Pinball loss is a comprehensive index to evaluate the reliability, sharpness, and calibration of the forecasts. It is an extensively used error measure for quantile forecasts in probabilistic forecasting problems. The performance of the forecasting models in GEFCom2014 was evaluated by the overall mean of the pinball loss values. Recall that the pinball loss function can be written as:

$$\text{Pinball}(\hat{y}_{t,q}, y_t, q) = \begin{cases} (1-q)(\hat{y}_{t,q} - y_t) & \text{if } \hat{y}_{t,q} > y_t \\ q(y_t - \hat{y}_{t,q}) & \text{if } \hat{y}_{t,q} \leq y_t \end{cases} \tag{5.7}$$

Month	[23]	[21]	Model I	Model II
1	9.88	11.87	3.43	**3.23**
2	9.54	10.93	3.24	**2.89**
3	7.79	8.44	2.69	**2.56**
4	4.89	4.50	2.53	**2.30**
5	5.96	7.27	**3.33**	3.50
6	5.86	6.99	4.98	**4.66**
7	7.66	9.05	3.63	**3.42**
8	10.70	11.26	8.71	**8.58**
9	6.28	5.49	4.46	**4.05**
10	5.20	3.36	2.97	**2.76**
11	6.38	5.90	**3.50**	3.59
12	8.99	9.73	3.57	**3.36**

Table 5.2: The left two columns are the reported results in [23], and [21]. The results of our proposed two different models are presented in the right part. The results reported here are the average of 100 iterations (no. of trees is 100, and MaxNumSplits=128).

where y_t is the target hourly value of the load profile from [5], and $\hat{y}_{t,q}$ is the corresponding forecast value at the $q-$th quantile ($q \in \{0.01, 0.02, \ldots, 0.99\}$); it is obtained from one of the models specified above.

To evaluate the full predictive densities, pinball scores obtained from Eq. (5.7) are averaged over the time horizon (99 quantiles for every hour, 24 hours of the day, n days of the month). A better forecast yield a lower pinball score. For more details on the pinball loss function and the evaluation methods used in GEFCom2014, see [5].

Table 5.2 contains the results of our proposed models along with two benchmark models. It is worth noting that all the data prior to the target month have taken part in the training of each model, i.e., the first eleven months in 2011 were used for training a model to predict the load profiles in December 2011. Furthermore, the average of 100 different hourly load values (top panel in Figure 5.6) is used as a proxy for the actual load value anytime needed. The reason for that is that in the later days of the month, the earlier load profiles are needed in the form of $\mathbf{L}(\mathbf{d}-\mathbf{1})$ or $\mathbf{L}(\mathbf{d}-\mathbf{7})$. The results in Table 5.2 highlights the fact that including the derivatives (especially the 2nd derivatives) is indeed helpful in enhancing the performance of the forecasting model. Diebold-Mariano test is another well-known metric to determine whether forecasts are significantly different. Let e_{i1} and e_{i2} be the residuals for Model I and II, respectively ($i \in [1:n]$). n is the number

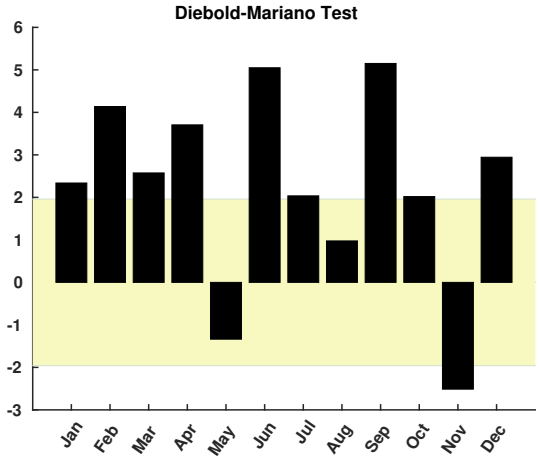


Figure 5.8: Comparison of the two models I and II, using Diebold-Mariano test ($h = 1$ and $k = 0$).

of data points, and k is the lagging variable [25].

$$
\begin{aligned}
d_i &= |e_{i1}|^2 - |e_{i2}|^2 \\
\bar{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i \\
\gamma_k &= \frac{1}{n}\sum_{i=k+1}^{n} (d_i - \bar{d})(d_{i-k} - \bar{d}), \quad n > k \geq 1 \\
DM &= \frac{\bar{d}}{\sqrt{[\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k]/n}}, \quad h \geq 1
\end{aligned}
\tag{5.8}
$$

Figure 5.8 contains the Diebold-Mariano test [26], [27] to determine whether the two models are significantly different. These results were obtained by comparing the error between the median ($q = 0.5$) of the forecasts from the two models and the actual values. The results are the average of 100 iterations, calculated according to Eq. (5.8). Suppose that the significance level of the test is $\alpha = 0.05$. For a two-tailed test, therefore, the upper and lower tails would each be 0.025. Accordingly, the upper and lower $z-$values are 1.96 and -1.96, respectively [28]. The null hypothesis of no difference between the two models (forecasts) will be rejected if the computed Diebold-Mariano statistic falls outside the range of [-1.96, 1.96]. Consistent with the results in Table 5.2, in February, June and September 2011, Models I and II are most significantly different. On the other hand, in August, when both models have the highest pinball score, the Diebold-Mariano (DM) test is low. Finally, DM tests in May and November 2011, are negative, as Model I outperforms Model II.

5.5. CONCLUSIONS

This chapter proposes two generic scenario-based probabilistic load forecasting models using an ensemble of regression trees. An important distinction of the current work is in recasting quasi-periodic time series data as matrices. The singular value decomposition technique is then used to generate temperature scenarios, in a robust, data-driven and timely manner. In the second model, we extend the first one by adding the first and second derivatives of the non-deterministic attributes (temperature and historical load data). It was done to partially account for the recency effects and interactions among the data. The empirical case studies performed on the data from the load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L) show how the proposed models outperform two benchmark scenario-based models with a similar set-up.

REFERENCES

[1] A. Khoshrou and E. J. Pauwels, *Propagating uncertainty in tree-based load forecasts,* in *Electrical and Electronics Engineering (ELECO), 2017 10th International Conference* (IEEE, 2017) pp. 120–124.

[2] A. Khoshrou and E. J. Pauwels, *Short-term scenario-based probabilistic load forecasting: A data-driven approach,* Applied Energy **238**, 1258 (2019).

[3] T. Hong *et al.*, *Energy forecasting: Past, present, and future,* Foresight: The International Journal of Applied Forecasting , 43 (2014).

[4] T. Hong, P. Pinson, and S. Fan, *Global energy forecasting competition 2012,* (2014).

[5] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, *Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond,* (2016).

[6] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, *Short-term residential load forecasting: Impact of calendar effects and forecast granularity,* Applied Energy **205**, 654 (2017).

[7] A. Khoshrou, A. B. Dorsman, and E. J. Pauwels, *Svd-based visualisation and approximation for time series data in smart energy systems,* in *Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2017 IEEE PES* (IEEE, 2017) pp. 1–6.

[8] L. Breiman, *Bagging predictors,* Machine learning **24**, 123 (1996).

[9] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. S. Kirschen, and C. Kang, *Combining probabilistic load forecasts,* IEEE Transactions on Smart Grid (2018).

[10] Y. Freund, R. E. Schapire, *et al.*, *Experiments with a new boosting algorithm,* in *Icml*, Vol. 96 (Citeseer, 1996) pp. 148–156.

[11] Y. Freund, R. Schapire, and N. Abe, *A short introduction to boosting,* Journal-Japanese Society For Artificial Intelligence **14**, 1612 (1999).

[12] D. Opitz and R. Maclin, *Popular ensemble methods: An empirical study,* Journal of artificial intelligence research **11**, 169 (1999).

[13] S. B. Taieb and R. J. Hyndman, *A gradient boosting approach to the kaggle load forecasting competition,* International journal of forecasting **30**, 382 (2014).

[14] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*, The Wadsworth and Brooks-Cole statistics-probability series (Taylor & Francis, 1984).

[15] L. Breiman, *Random forests,* Machine learning **45**, 5 (2001).

[16] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski, *Classification and regression tree analysis in public health: methodological review and comparison with logistic regression,* Annals of behavioral medicine **26**, 172 (2003).

[17] W.-Y. Loh, *Classification and regression trees,* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**, 14 (2011).

[18] B. Liu, J. Nowotarski, T. Hong, and R. Weron, *Probabilistic load forecasting via quantile regression averaging on sister forecasts,* IEEE Transactions on Smart Grid **8**, 730 (2017).

[19] P. Wang, B. Liu, and T. Hong, *Electric load forecasting with recency effect: A big data approach,* International Journal of Forecasting **32**, 585 (2016).

[20] T. Hong and S. Fan, *Probabilistic electric load forecasting: A tutorial review,* International Journal of Forecasting **32**, 914 (2016).

[21] J. Xie and T. Hong, *Gefcom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation,* International Journal of Forecasting **32**, 1012 (2016).

[22] T. Hong, J. Wilson, and J. Xie, *Long term probabilistic load forecasting and normalization with hourly information,* IEEE Transactions on Smart Grid **5**, 456 (2014).

[23] F. Ziel and B. Liu, *Lasso estimation for gefcom2014 probabilistic electric load forecasting,* International Journal of Forecasting **32**, 1029 (2016).

[24] E. Langford, *Quartiles in elementary statistics,* Journal of Statistics Education **14** (2006).

[25] *Diebold-Mariano Test Statistic,* https://nl.mathworks.com/matlabcentral/fileexchange/33979-diebold-mariano-test-statistic?focused=7267180&tab=function ().

[26] D. Harvey, S. Leybourne, and P. Newbold, *Testing the equality of prediction mean squared errors,* International Journal of forecasting **13**, 281 (1997).

[27] F. X. Diebold and J. A. Lopez, *8 forecast evaluation and combination,* Handbook of statistics **14**, 241 (1996).

[28] *Comparing predictive accuracy of two forecasts: The diebold-mariano test*, http://www.phdeconomics.sssup.it/documents/Lesson19.pdf ().

6

VOLATILITY QUANTIFICATION

6.1. INTRODUCTION

Renewable Energy Sources (RES) are assuming an increasingly pre-eminent role in German electricity production. To secure an economic and environmentally compatible supply, Germany has substantially expanded its RES-capacity, in particular wind and solar [4]. This creates new challenges as wind and solar energy are fundamentally intermittent, weather-dependent and unpredictable. It hence raises concerns about the overall reliability of the power supply and its flexibility. It is therefore of considerable interest to investigate what effect this energy transition could have on the overall trend and volatility of electricity prices. This impact could be complex because there are a number of contradicting forces at play. The marginal cost of RES is relatively low and even negative (especially if subsidized); therefore, increased penetration of wind and solar would result in a downward trend in electricity prices. Moreover, a perceived advantage of market coupling is to reduce price volatility [5]. On the contrary, the associated uncertainty regarding the availability of wind and solar energy is expected to cause spikes in the market. In other words, the integration of RES provokes the assertion that the stability of the power grid can be compromised due to the inherent intermittency of such sources. Consequently, the increased price volatility will cause additional market risks for suppliers and consumers on the market.

Volatility principally refers to random fluctuations of a time series about its mean or expected value. Generally speaking, in financial time series data analytics, volatility is measured by the standard deviation of the logarithmic return or a derivation of that [6]. In the literature, numerous methods have been introduced to determine the volatility of the time series data. Diverse methods, from applied models such as Garman-Klass and Rogers-Satchell volatility estimators to the coefficient of variation based, and formal stochastic volatility models including GARCH, Heston models and [7]. Recently, however, new concepts and notions of volatility have been explored, especially in financial data analysis. Ruiz et. al., in [8] propose the permutation entropy, topological entropy

Parts of this chapter have been published in [1–3].

and the modified permutation entropy as alternative measures for volatility quantification. In the reported work, the degree of randomness or *determinism* of a time series is considered as the notion of the volatility of data. Simonsen [9] studies different volatility features (including volatility clustering, log-normal distribution, and long-range correlations) of the Nordic day-ahead power spot market for the course of 12 years (1992 till May 2004). The aforementioned work also reports the presence of cyclic behavior of the time-dependent volatility for the quasi-periodic (with almost diurnal patterns) power market data. Additionally, the striking differences between the range of price data in different years are reported to be an obstacle in developing a generic approach to analyzing the market from different perspectives. The volatility has namely dependence on the price level, which is even more pronounced when spot prices are low. Therefore, in terms of analyzing EPEX data shifting up all the values for different years by a certain threshold in order to use the traditional financial data methodologies, does not appear to be a viable approach, as the results may vary drastically for different thresholds. A frequency domain-based method is deployed in [10] to systematically separate out the periodic components of the prices from random variations. After removing the *deterministic* part, the price volatility is determined by fitting a Wiener process to the remaining random stochastic (residual) part.

Because of the aforementioned characteristics of the EPEX price values, in the first step, we pre-process the raw data to eliminate the underlying patterns; and subsequently, focus on quantifying the volatility of data on an hourly or daily bases. The nature of this pre-processing is discussed in the following subsections.

Along with the increase in the utilization of intermittent renewable sources, short-term electricity market studies (including day-ahead, intraday and imbalance market) are becoming increasingly popular. We herein opt to focus on the day-ahead market as it represents an important and growing segment where market mechanisms are clearly visible. In particular, we focus on the following question: How can the evolution of the price volatility of the day-ahead market over the past eleven years (i.e., 2006-2016) be quantified? Inspired by the work in [11], we consider matrices as an alternative representation of the electricity market data where the time series demonstrates periodic patterns. In the next step, a popular and numerically stable matrix decomposition technique, namely the singular value decomposition (SVD), is used to disentangle the matrix of daily price profiles (one year's worth of hourly values) based on the most dominant daily profiles (left singular vectors as a notion of trend) and their corresponding variability (the right singular vectors). Accordingly new, yet easy-to-quantify, notions of hourly and daily volatility are proposed using a lower-rank matrix reconstruction (as a measure of hourly volatility) and the right singular vectors (as a measure of daily volatility). The second part of this chapter is dedicated to exploring the possible effect of RES on the overall price profiles (e.g., shifts in peak price hours, emergence of zero or negative prices).

The rest of this chapter is organized as follows. Section 6.2 focuses on the data description for the day-ahead market; it also explains the source and a brief summary of the day-ahead market mechanism. Section 6.3 provides a brief recapitulation of SVD. Sections 6.4 and 6.5 are dedicated to our methodologies and detailed description of the proposed daily and hourly volatility quantities. We conclude in Chapter in Section 6.7.

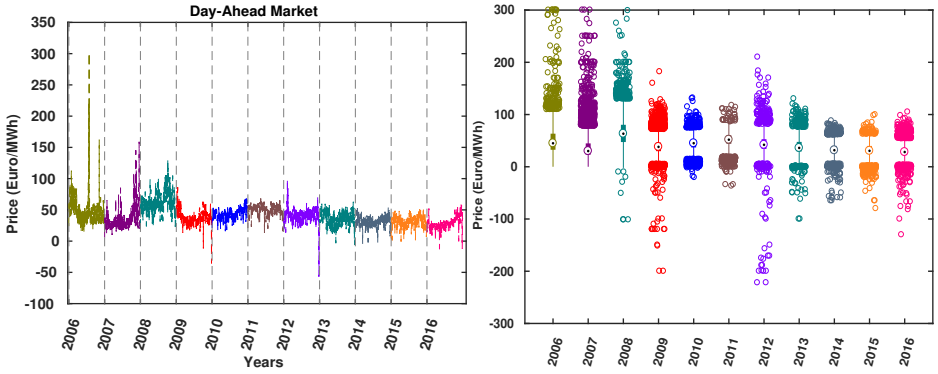


Figure 6.1: **Left**: Evolution of the German day-ahead spot prices from 2006 through 2016 (daily averages). **Right**: The corresponding box-plots of the hourly price data (for the sake of visualization, only values between $[-300,300]$ are illustrated). It is evident to see that zero or negative prices started to appear in late 2008 afterwards.

6.2. DATA

The European Power Exchange (EPEX SPOT SE) operates on the Central Western European (CWE) spot market. To guarantee a single integrated and transparent market, the EPEX SPOT SE acts as a neutral intermediary market operating service provider between the market members active in the central western European countries - viz. Switzerland, France, Germany and Austria.

This market consists of non-final consumers and big players in the energy sector such as utilities, wind and solar farm owners, hydroelectric power stations, aggregators, transmission system operators (TSOs), financial service providers and also energy trading entities that are working within the energy sector on a daily basis [12]. The following sections describe the functionality of the day-ahead market.

6.2.1. DAY-AHEAD AUCTION SPOT MARKET

An exchange for short-term (one day before the power delivery) electricity contracts is the day-ahead market. It is a single integrated market where the participants themselves propel the trading. An electricity buyer, typically a utility or TSO, determines the amount of energy (and the purchase price) it will need to fulfil its customer's requirements for the coming day. The seller, e.g., the owner of a wind or solar farm, also submits the quantity which they are prepared to deliver the next day and the price level for each hour.

These "bids" are then fed into a complex algorithm to calculate the *clearing* price. In the end, the financial and physical transactions are settled. The output of the algorithm is in fact a number of time series of prices (bounded between $[-500,3000]$), and traded volumes which are going to be exchanged, per area and period of the day, for the next day [5, 13].

Day-Ahead Spot Prices (in €/MWh) Figure 6.1 illustrates an overview of the hourly values of the price on the day-ahead market in Germany and Austria from 2006 until

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
μ	50.79	37.98	65.76	38.85	44.48	51.12	42.59	37.78	32.76	31.63	28.98
std	49.42	30.35	28.65	19.40	13.98	13.60	18.69	16.45	12.77	12.67	12.48
C_v	97.31	79.91	43.58	49.94	31.42	26.60	43.87	43.53	38.99	40.05	43.08

Table 6.1: Annual mean, standard deviation and the coefficient of variation C_v of EPEX price data, in the years 2006-2016.

2016 (data source: [12]). Interestingly enough, it is plain to see that after triggering the energy transition in 2011, occurrences of zero or negative price values are more frequent. Furthermore, a cursory glance at this time series suggests an overall downward trend in price range as well as its volatility. As mentioned before, it is reasonable to question whether the intermittency of renewables would render the price more erratic.

Loosely speaking, volatility refers to the random fluctuations of a time series about its expected value. There are various methods to define and quantify volatility, from applied models like Garman/Klass to coefficient of variation and formal Stochastic Volatility models such as GARCH, Heston models and the like e.g. [7, 14]. There are at least two reasons why it is problematic to blindly transfer standard fintech methodology to the current setting:

- While the stock market prices are only available on trading days, EPEX prices cover the whole year, 24 hours 7 days a week. Accordingly, the underlying variability of the data could wrongly be conceived as volatility whereas it in fact simply reflects the diurnal patterns of human activities.

- More importantly, EPEX prices can be zero or even negative; therefore, the standard approach to switch to logarithmic measures can be done only after shifting up all values above zero by a certain threshold. On the other hand, price volatility has a dependence on the price level, which is even more pronounced when the spot prices are low [9]. Therefore, the generalizability of conventional approaches can be questioned, as the volatility measures can vary drastically, with respect to the magnitude of the aforementioned thresholds.

Table 6.1 contains the annual mean, standard deviation and coefficient of variations of the German day-ahead market. Interestingly enough, despite a consistent reduction in the annual mean (μ) and the standard deviation (std), the coefficient of variation ($C_v = \frac{std}{\mu} \times 100$) has increased from 2015 afterwards. Figure 6.2 provides a comparison of the annual standard deviation of the logarithmic returns of the EPEX data, for three different cases. As mentioned before, since the original time series contain non-positive values, in the first step, we shift up all the hourly price values $P = \{p_1, p_2, \ldots, p_{8760}\}$ (p_{8784} for a leap year) by a positive α value ($x_t = p_t + \alpha$). Considering the upper and lower limit of the EPEX price values (recall that $p_t \in [-500, 3000]$), three different scenarios have been defined as follow: 1) $\alpha = |min(P)| + 1$; 2) $\alpha = |max(P)| + 1$; and 3) $\alpha = 501$. The logarithmic returns are then calculated as follows:

$$\beta_t = \log(x_t) - \log(x_{t-1}) \tag{6.1}$$

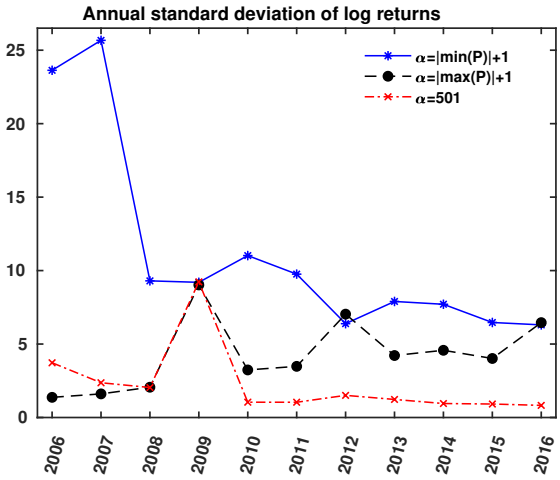


Figure 6.2: Comparison of annual standard deviation of logarithmic returns of EPEX data for different α values.

In the final step, the standard deviation of the logarithmic return values β_t is considered as a notion of annual volatility. It is plain to see that different α values have led to contradictory results. We herein define a new notion of hourly and daily volatility (using the SVD method) which is robust in terms of non-positive values, as fully elaborated in the following sections.

6.3. MATRIX DECOMPOSITION USING SVD

As previously mentioned, the SVD is applied extensively in matrix computations, but can also be put to good use in the study of time series that have exogenously induced periods. This is often the case in economics time series, where the variables of interest show cyclic patterns.

As it is explained in the previous chapters, to obtain a matrix format of the data for each year, we reshape the time series data into matrix $A_{h\times d}$, where $h = 24$ is the number of hours of the day, and $d = 365$ (366 for a leap year) is the number of days of the year [11]. The SVD method is then used to decompose the matrix $A_{h\times d}$ into a set of fundamental daily profiles and their corresponding weights during the observed time span (it becomes apparent in the following section). This method is more robust, regarding the aforementioned issues with EPEX price data, as it enables us to explore the trend and volatility of each year individually and in a data-driven manner, with no need to add offset values.

As an example, Figure 6.3 displays an overview of the German day-ahead electricity prices in 2016. As a first step, we recast each year's worth of data into a matrix $A_{h\times d}$ where each column contains 24 hourly values for a given day – it can be seen as d points in $h-$dimensional space, similar to Figure 1.15. We then obtain its corresponding three factorization matrices using the SVD expansion. As discussed in Chapter 2, if the price profiles for each day were identical (or linearly dependent), i.e. all the columns were identical (or linearly dependent), the matrix would have had a rank equal to one ($rank(A) = 1$).

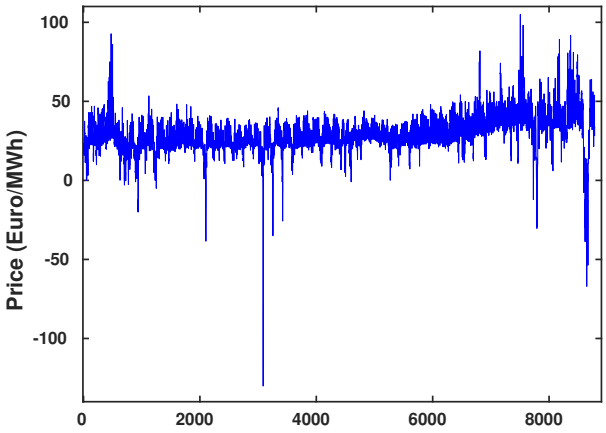


Figure 6.3: An overview of the evolution of the hourly day-ahead price values in 2016.

However, in practice, the rank of the matrix $A_{h\times d}$ exceeds one since the daily price profiles for subsequent days tend to differ, depending on the calendar information, supply availability and demand.

Figure 6.4 illustrates the first two dominant U_k and V_k profiles, corresponding to the two largest singular values σ_k ($k = 1,2$). It is plain to see that the most dominant U profile (U_1, top left), highly resembles an appropriately weighted average of daily profiles (averaged over the year). The morning and late afternoon price peaks are clearly discernible in this plot. The second column of U (U_2 bottom left) acts as a correcting factor, which needs to appropriately be added to the first profile to build up a more detailed representation of the data. The panels on the right display the corresponding V_k coefficients that specify the magnitude of the corresponding U_k profile for each day of the year. In other words, V_k profiles sensibly reflect the daily variability of their corresponding U_k profiles throughout the year.

Figure 6.5 demonstrates the evolution of the singular values for different years. It is plain to see that by considering only a few singular values (and their corresponding singular vectors) we are capable of reconstructing the price data with a good approximation.

6.4. QUANTIFYING THE DAILY VOLATILITY

The wavelet decomposition technique is used herein to quantify the daily volatility of the EPEX price data.

6.4.1. WAVELET DECOMPOSITION

In modern mathematics, wavelets are one of the most efficient and widely used tools to analyse digital signals. As the name suggests, wavelet analysis is akin to Fourier analysis which decomposes the signal of interest as a linear combination of sine waves of different frequencies and phases [15].

Wavelet analysis will not only tell us which frequencies are hidden in the signal,

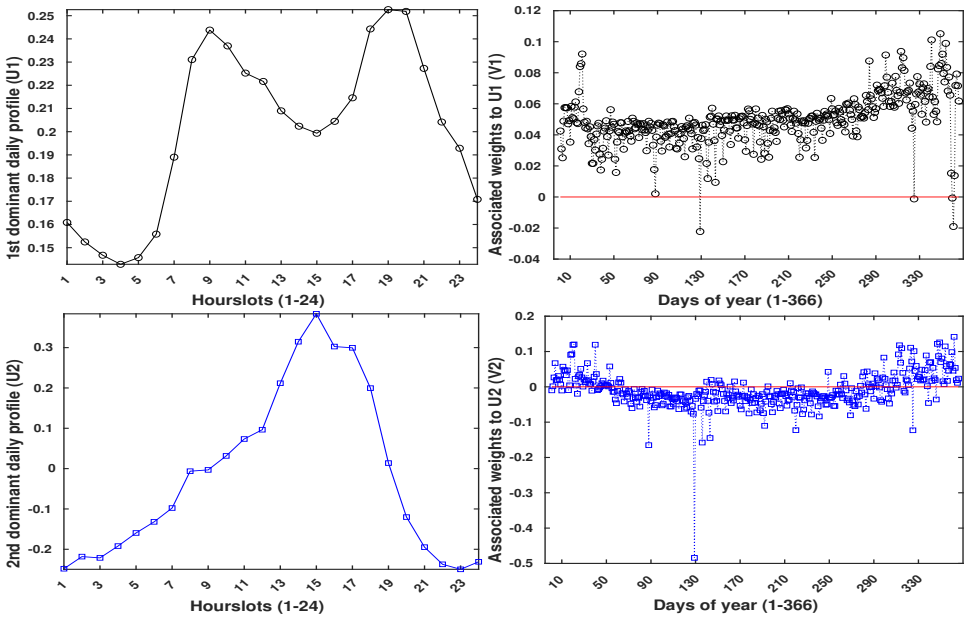


Figure 6.4: SVD-based rank-2 approximation: the first two most dominant U_k profiles are depicted on the left side. The panels on the right contain their corresponding V_k coefficients. Left column: first two columns of U-matrix representing a weighted averaged profile for each day (top) and a first order correction (again one value for each hour over a 24 hour period). Right column: corresponding amplitudes (V columns).

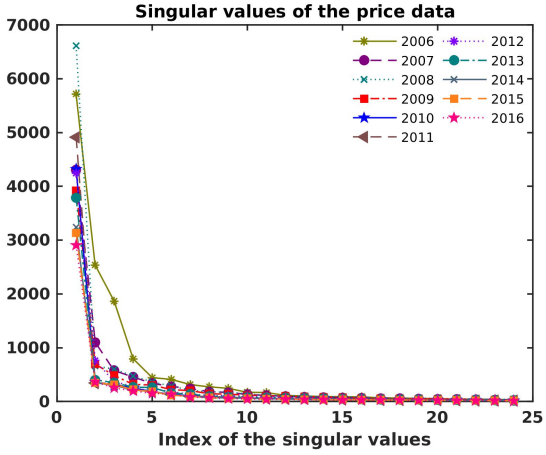


Figure 6.5: An overview of the evolution of the singular value of the price data in recent years.

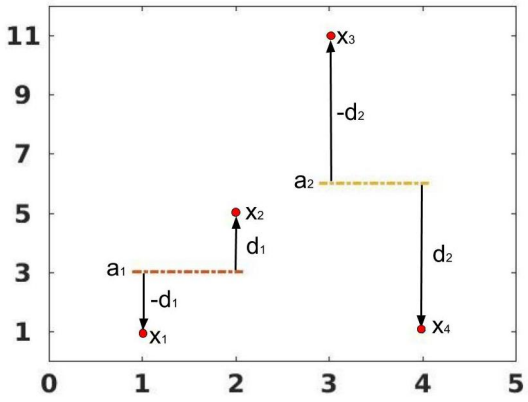


Figure 6.6: Schematic representation of the Haar wavelet decomposition.

but can also pinpoint their location in the data stream. From this, it becomes clear that wavelets hand us a useful tool to probe the data for the occurrence and location of high-frequency fluctuations [16]. The Haar wavelet is arguably the simplest wavelet and lends itself to a straightforward interpretation. It basically takes any discrete signal $\mathbf{x} = (x_1, x_2, x_3, \ldots)$ and creates an approximation $\mathbf{a}$ and detail $\mathbf{d}$ signal by running the following simple recipe:

1. Take the first two elements x_1 and x_2 and compute the approximation and detail coefficients:
$$a = \frac{x_1 + x_2}{2} \quad \text{and} \quad d = \frac{x_1 - x_2}{2}.$$
Notice that this implies $x_1 = a + d$ and $x_2 = a - d$, or more explicitly: the approximation coefficient equals the mean of the two values, and the detail coefficient is the amount of deviation between the actual value and the approximation.

2. Store the results in the approximation and detail vector, respectively:
$$\mathbf{a}(1) = a \qquad \mathbf{d}(1) = d.$$
Both vectors have a length equal to half the length of the original input $\mathbf{x}$.

3. Move on to the next pair (x_3, x_4) and continue until all $\mathbf{x}$-elements have been processed. This way we get the level-one approximation ($\mathbf{a}_1$) and detail ($\mathbf{d}_1$) coefficient (each vector of half the length of the original $\mathbf{x}$-sequence).

4. To compute the level-two approximation and detail coefficients we repeat the whole procedure but use $\mathbf{a}_1$ as input (instead of $\mathbf{x}$).

5. This can be continued until we have reached a pre-defined level.

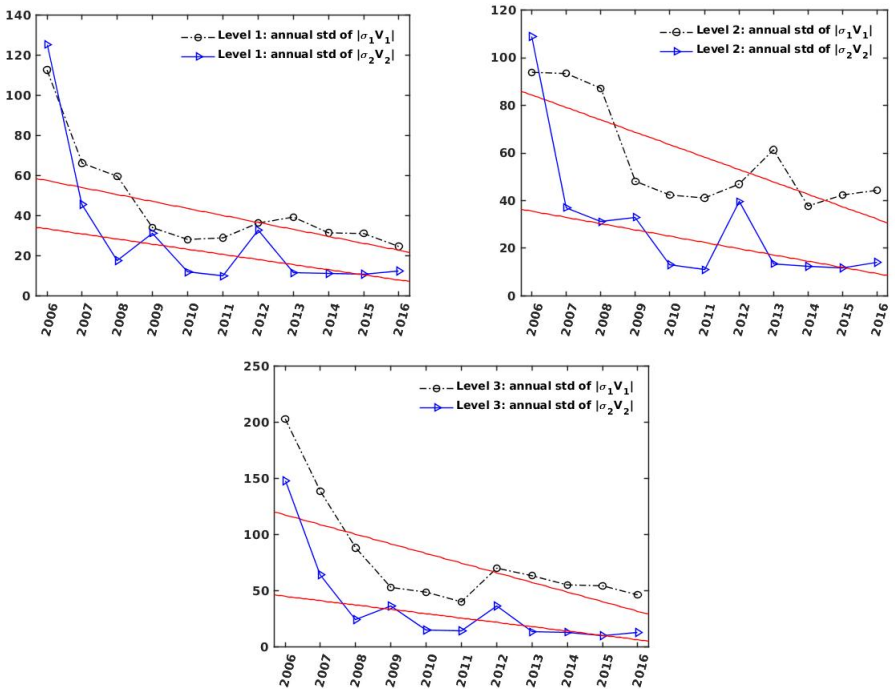


Figure 6.7: Evolution of the annual standard deviation of the Haar wavelet detail coefficients of the absolute values of v1, v2 profiles magnified by their corresponding singular values, down to three level (starting from the top left).

Figure 6.6 exemplifies the decomposition for a (short) discrete signal $\mathbf{x} = (x_1, x_2, x_3, x_4)$. In the first analysis step, the original values (the dots) are paired, and each pair is replaced by its mean or approximation (a_i, the dash-dotted lines) and the symmetric deviation d_i with respect to the corresponding mean. As a consequence, the original signal $\mathbf{x}$ can equally well be represented by the approximation vector $\mathbf{a} = (a_1, a_2)$ and the vector of detail coefficients $\mathbf{d} = (d_1, d_2)$. The next analysis step (not depicted here) would repeat the procedure, this time starting with the approximation $\mathbf{a}$ as input. As a concrete example, imagine that the time series is given by $\mathbf{x} = (1 \quad 5 \quad 11 \quad 1 \ldots)$, then

- Level 1: $\mathbf{x} = (\underbrace{1 \quad 5}_{3\pm 2} \quad \underbrace{11 \quad 1}_{6\pm(-5)} \ldots) \longrightarrow \mathbf{a}_1 = (3 \quad 6 \ldots)$ and $\mathbf{d}_1 = (2 \quad -5 \ldots)$
- Level 2: $\mathbf{a}_1 = \underbrace{(3 \quad 6}_{4.5\pm(-1.5)} \ldots) \longrightarrow \mathbf{a}_2 = (4.5 \quad \ldots)$ and $\mathbf{d}_2 = (-1.5 \quad \ldots)$

6.4.2. VOLATILITY QUANTIFICATION

It is important to realize that the level-one detail coefficients capture the highest frequency oscillations. Subsequent detail coefficients correlate with oscillations of successively lower frequencies. As mentioned before, in the present work, the right singular

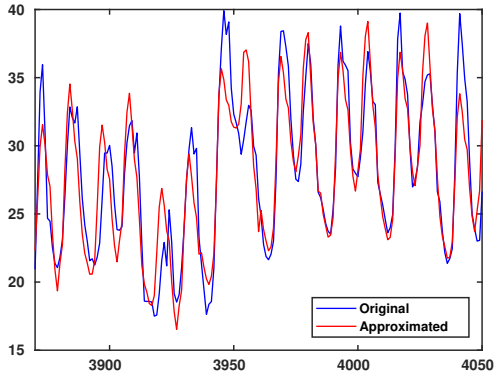


Figure 6.8: Detail of rank-2 approximation (red) superimposed on original data (blue).

vectors V_k are considered to be an indicator of the volatility of the fundamental daily price profiles (U_k) throughout a year. As it was mentioned earlier, Figure 6.5 confirms that there are only a few dominant singular values; we hence opt to consider only the first two right singular vectors (V_k). Figure 6.7 contains the standard deviation of the values of the Haar wavelet detail coefficients, down to three-level wavelength decomposition. The downward trend underscores a reduction in the volatility of the German day-ahead market in recent years. Worth noting that V_1 and V_2 coefficients have been magnified using their corresponding singular values σ_1 and σ_2. The results here corroborate with [2], where the hourly volatility of the German day-ahead market during the same period is studied.

6.5. QUANTIFYING THE HOURLY VOLATILITY

A thorough understanding of volatility is crucial for many applications in financial economic studies such as derivative pricing, corporate risk management, market efficiency, and many others. As previously mentioned in Section 6.2.1, some characteristics of the EPEX price data have made it impractical to apply the conventional econometric approaches in the time series volatility quantification. Therefore, we propose an alternative approach to hourly price volatility quantification using the residuals of the matrix reconstruction.

To get some idea of what a rank-2 approximation looks like, Figure 6.8 shows a detail of the approximation (in red) superimposed on the actual data (blue). Figure 6.9 contains a more general case of the reconstruction of price data in 2016, along with the absolute value of the corresponding residuals. In the current section, we focus on the residuals of the price after compensating for daily patterns using a rank-2 approximation; it is done to quantify the hourly volatility of the data over the years. As previously mentioned, Figure 6.5 illustrates that the singular values of the price data throughout different years follow the same pattern. Therefore, according to the results in Figure 6.5 the choice to focus on rank-2 approximation is relatively arbitrary and unimportant.

As can be seen in the top panel of Figure 6.10, the absolute value of the residuals adheres remarkably well to an exponential distribution (with a mean 2.97). It is almost

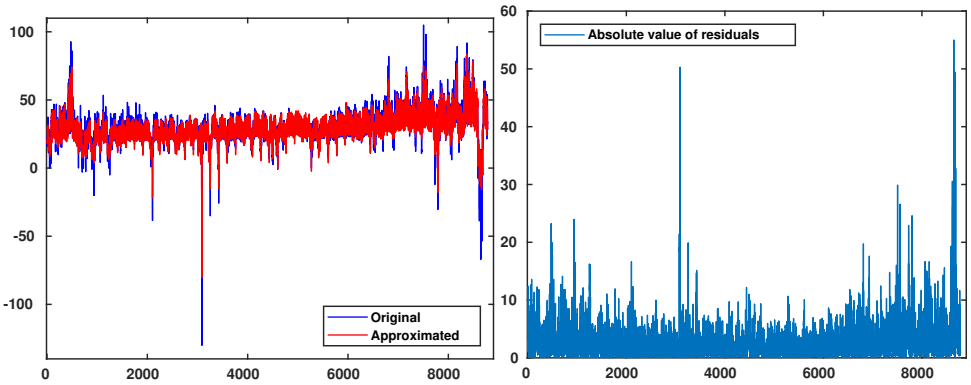


Figure 6.9: **Left**: Original and rank-2 approximation of the price data for 2016. **Right**: Absolute values of the residuals after rank-2 approximation. Residuals are a means to measure volatility.

only the top 1% that is substantially higher in value than expected. Moreover, the lower panel of Figure 6.10 highlights the fact that the higher rank approximations yield similar results. Only as expected, by increasing the reconstruction rank, the absolute value of the residuals decreases. Figure 6.11 confirms that this trend is consistent throughout the period of observation (2006-2016). Based on this observation, we propose the following approach to quantify the evolution in (annual) volatility in the years 2006 through 2016:

1. The influence of daily and seasonal variations is removed by fitting a rank-2 approximation and extracting the residual of the actual data with respect to this approximation. As volatility is influenced by both positive and negative fluctuations, we hence focus on the absolute values of the residuals. Worth noting that, the bottom panel in Figure 6.10 indicates that higher-rank reconstructions yield similar results.

2. For every year, we fit the lowest 99% of the absolute value of the residual with an exponential and compute the corresponding parameter (i.e. mean of the exponential). This value corresponds to the size of the residuals. Note that here the absolute value of the residuals is considered, which is the reason for preferring the exponential distribution over the normal distribution in our methodology.

3. Typically, the top 1% of the observed distribution is much larger than expected (based on the bulk of the distribution). We characterize these values by computing the median value of this top 1% segment separately.

The results are shown in Figures 6.12 and 6.13. The former figure shows a robust estimate for the mean of exponential distribution for each year. The estimate is based on the lowest 99% of the absolute values of the residual and is therefore robust with respect to the top 1% of extremely large values. The 99% vs. 1% is dictated by the exponential prob-plot in Figures 6.10 and 6.11 which show a clear divergence at the 99% mark. To quantify this decreasing trend, we have computed the regression line, which yields a statistically significant downward slope equal to -0.58 (with 95% confidence interval: -0.89:-0.26). The quality of this regression can be further improved by fitting a *power law*,

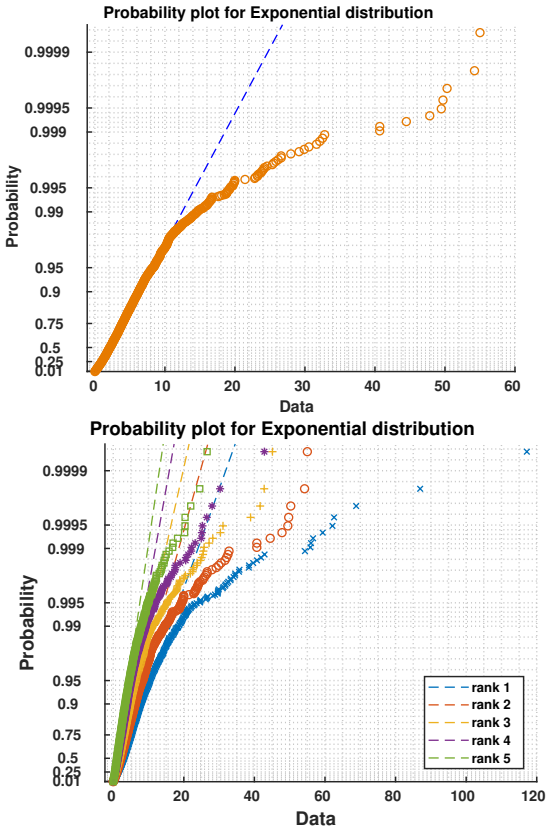


Figure 6.10: **Top**: The residuals of the rank-2 approximation of the day-ahead market prices for the year 2016: almost 99% of residuals adhere to an exponential distribution. **Bottom**: Higher rank approximations yield similar results.

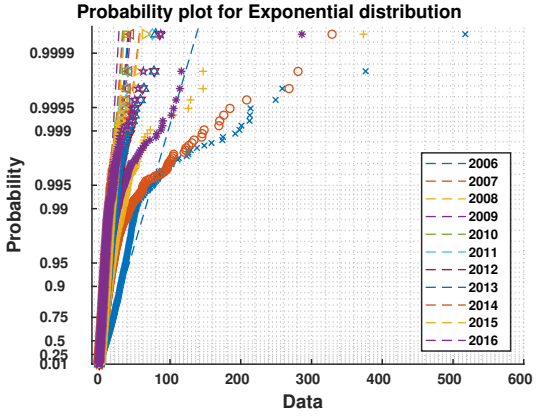


Figure 6.11: Exponential distribution of the residuals of the rank-2 approximations for different years.

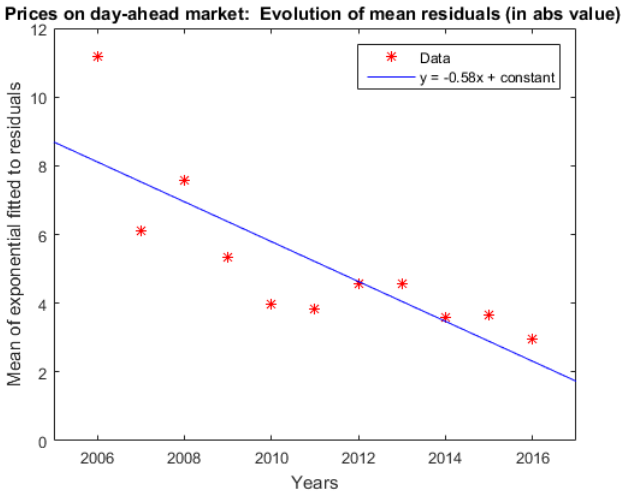


Figure 6.12: Evolution of the price volatility on the German day-ahead market in the period 2006 through 2016 . The individual data points record the estimated exponential parameter (mean) based on all but the 1% highest values of the residuals. The regression line has a slope equal to -0.58 which is highly significant (95% confidence interval: -0.89 : -0:26). See the main text for more details.

but at this point, a linear regression serves the purpose to illustrate the significant downward trend. The evolution of the top 1% is shown in Figure 6.13 where these data are represented by their median value. This figure indicates that whereas extreme residuals were not uncommon prior to 2010, these values fell significantly and have been roughly constant during the years 2012-2016. The message from both figures combined is that volatility has decreased significantly over the years 2006 to 2016.

Volatility tends to be higher in winter By scrutinizing Figure 6.9 it becomes evident that the volatility tends to be lower in summer (middle part of the graph) than in winter (extremal parts of the graph). To demonstrate that this is indeed the case, we use a measure based on the angular momentum. More precisely, if the (absolute) residual for hour slot h is given by $R(h)$ and the distance between the hour slot h and the central hour slot $h_m = n/2 = 4392$ equals $|h - h_m|$ the observed angular momentum is defined a:

$$L_{obs} = \sum_{h=1}^{n} R(h)(h - h_m)^2 \tag{6.2}$$

where n is equal to the total number of hour slots. If we re-scale the values of the hour slot in such a way that $h_m = 0$ and $-1 \leq h \leq 1$ (divided by 1000, for ease of comparison), we obtain $L_{obs} = 10.676$. A high value for L_{obs} refutes the assumption that the residuals are uniformly distributed throughout the year and favours an interpretation in which residuals (and hence volatility) are higher in the winter season. We judge the significance of this result using a permutation test. The rationale is straightforward: if the residuals are uniformly distributed over the hour slots, then a random permutation of the values

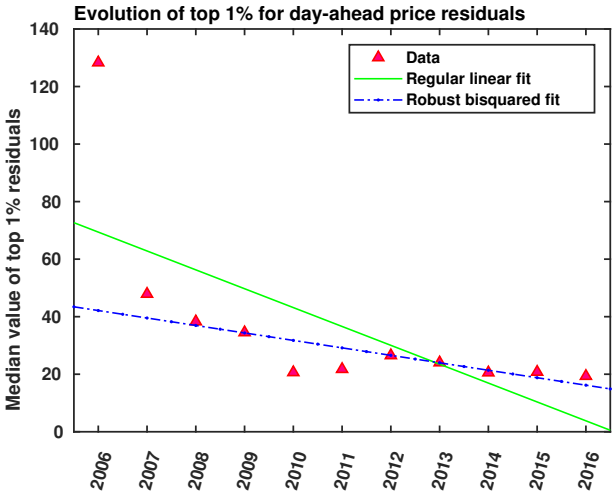


Figure 6.13: Evolution of top 1% residuals over the years 2006-2016. Each data point represents the median value of the top one percent residuals (in absolute value). As such, these values characterize the extreme deviations in day-ahead prices. The plot clearly shows that these extreme values decreased significantly before 2010, and then stayed approximately constant.

should not result in a significantly different value for L_{obs}. The results of the permutation test for 2016 are depicted in Figure 6.14. Our experiments show that similar results for all the other years can be produced.

6.6. EXTRACTING THE UNDERLYING TRENDS

This section is dedicated to a more descriptive study of the evolution of the overall trend of the day-ahead market.

6.6.1. THE EVOLUTION OF THE DAILY PROFILES

As mentioned before, the left singular vectors (U_k profiles) represent the most dominant daily profile (U_1) and its additive corrective profiles ($U_2, U_3, \ldots$). Figure 6.15 displays the evolution of the most dominant daily price profiles (U_1 magnified by σ_1) over the years. A continuous downward trend during the decade in the average value of the daily profile is noticeable. More importantly, the change in the overall shape of the daily profile is even more telling. Before 2011, the morning peak price values tended to be higher than the afternoon peak values. Whereas this trend has become reversed in recent years. Another intriguing feature of the data is the shift of the time slot (during the day) which more points to the effect of the low-cost subsidized RES on the daily price profile. Before 2011, the electricity price is most expensive at around 12h00. Evidently, the availability of solar after 2011 has pushed the prices lower and has led to morning peak prices at around 9h00. In a similar way, the afternoon peak price time slot has a shift of an hour from around 19h00 to 20h00. Furthermore, it is plain to see that the ranges of the daily profiles (difference between the maximum and minimum values) show a reduction, in recent

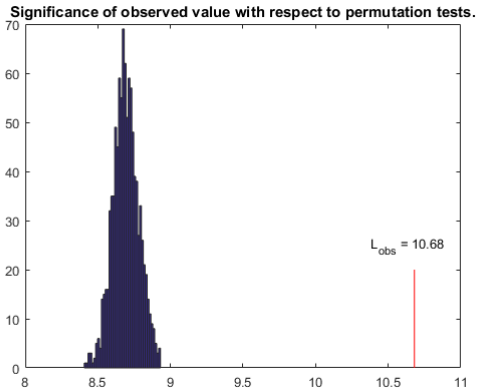


Figure 6.14: Price data for 2016: Results of the permutation test. The histogram of the L-values for 1000 random permutations of the actual data. Obviously, the actually observed value (indicated in red) is significantly larger than the values we would expect for data sets without structure (with a p-value $< 10^{-3}$). This confirms our observation that volatility shows a seasonal pattern.

years. This also can be an indication of less volatility in years. In other words, the change in timing and amplitude of daytime and evening time peak values after 2011 are striking. Before 2011, the midday peak price (around 12h00), was considerably higher than the early evening spike around 20h00. However after 2011, the first spike is not just lower than the second one, but also shows a clear shift to earlier hours. This makes sense in light of the higher contribution of renewables, and in particular solar; it is reasonable to assume that the typically high supply of solar power around midday is the reason for the drop in prices during these hours. The right panel of Figure 6.15 displays an alternative representation of the same data in the left panel; it confirms the previous findings by showing a downward trend in average daily prices from left to right. The diminishing contrast in each column indicates a smoother price profile with a lower daily spread over the years. Looking closely, the shift of the midday and afternoon peak hours is notable.

6.6.2. THE EXTREME VALUES

In the next step, the *extreme* values of the hourly prices during the years have been probed. More specifically, collecting all the hour slot values for each year in the period 2006 through 2016 yields a price distribution for each year. Extreme prices (both high and low) are characterized as prices outside the extreme 5% percentiles. So we get a representative value for high (low) prices by focusing on the values of the 95% (5% respectively) percentile for the distribution of each year's worth of hourly price values. The results are shown in Figure 6.16 where we have plotted both values (high and low) for each year. There is a pronounced continuous downward trend for the high prices, with 2008 being an obvious outlier. The lowest prices show a slight decrease over the years, as there is obviously less room for manoeuvre. The overall spread of the prices is steadily decreasing and less volatile, indicative of a more mature market.

6.6.3. THE DISTRIBUTION OF HIGH AND LOW PRICE VALUES

We herein explore the evolution in the distribution of occurrences of extreme prices over the course of the day. Recall that high (low) prices are defined as values outside the 95th (5th) percentile of price distribution for that year. Figure 6.17 (left panel) shows how the occurrence of low prices is distributed over the day (for the years 2006 through 2016). Whereas in the earlier years of the decade, there is a clear concentration of low price occurrences in the early morning (4h00-5h00), later years show a more uniform daily distribution. A similar distribution for the occurrence of high prices is shown in the right panel of Figure 6.17. It indicates a distinct shift (occurring around 2011) in the time slot of high prices. More precisely, the daytime peak is shifted from noon to the earlier hour of 9h00, while the evening is postponed, shifting from 19h00 to 20h00. In other words, before 2011, the daytime peak values were higher than the afternoon ones and occurred around noon. After 2011, high prices occur predominantly at the beginning and end of the peak period. Each column represents 440 values, which is 5% of the total number of observations during one year. Also apparent is the fact that starting in 2011, the afternoon spikes in the price profile exceed the daytime ones.

6.6.4. ZERO AND NEGATIVE PRICES

Of special interest are zero or negative prices as they reflect the effect of subsidized RES. In Germany, a significant amount of electricity is still produced by conventional sources. In 2015, e.g., lignite, nuclear energy and hard coal were responsible for producing 24, 14.2 and 18.3 % of gross electricity production, respectively [17]. The synchronization speed of these plants is slow and they can not be shut or ramped down very quickly. As a result, on some days when there is an excess of electricity production by subsidized renewable energy sources, prices may become negative and consumers can actually make a profit by consuming electricity. Figure 6.18 provides an overview of the frequency along with the magnitude of non-positive prices, in recent years. More specifically, the width (on the x-axis) of the interval assigned to each year is proportional to the number of occurrences of negative (or zero) prices in that year. The y-axis depicts the corresponding magnitude of these negative prices. Starting in 2012, the number of occurrences (length of the interval) seems to increase steadily. This trend is strikingly consistent with the growing contribution of wind and solar energy in Appendix B.1. From this graph it transpires that there is a reduction in the magnitude of the negative prices in recent years, although their number is mildly increasing.

The distribution of the growing number of instances of zero as well as negative prices in recent years, from a different perspective, is illustrated in Figure 6.19. The left panel highlights the frequency of occurrence of non-positive price values during different hours of the day, for each year. Although, before 2012, the majority of non-positive prices are happening in the early hours of the day, a cluster of non-positive price values appeared during the midday (11h00-18h00). Considering the fact that the electricity demand during these hours has not changed much (working hours), the most probable explanation for this change can be the oversupply of solar farms. As is seen in Appendix B.1.2, these are the hours with the highest solar energy availability; on average, at 13h00, solar production can become as high as 5 gigawatt-hour (GWh), which is almost 5 times more than solar feed-in at 10h00. This number can be even more during the summer. Ap-

pendix B.1.3 highlights the fact that wind and solar energy combined are responsible for almost 10 GWh feed-in at around 13h00, throughout the year. Another conspicuous observation in the left panel of Figure 6.19, is the increased frequency of the occurrences of non-positive prices in the early hours of the day in the last two years. Interestingly enough, Appendix B.1.1 presents how the wind feed-in have notably increased in 2015 afterwards. The right panel of Figure 6.19, confirms the fact the non-positive price values are most frequent during the weekends when the consumption is low. However, we witness more instances of zero or negative prices during the week, from 2011 afterwards.

6.7. CONCLUSIONS

In this chapter, we have traced the impact of the integration of renewable energy sources (RES) in Germany on the day-ahead electricity market, in terms of volatility, in the years 2006-2016. Regarding volatility quantification, there are a number of peculiarities that make conducting the empirical methods onerous. EPEX price data have the following characteristics: 1) It covers the whole year, 24 hours 7 days of a week; 2) It can have non-positive values; 3) It depends on the calendar information (working and non-working days), and 4) It shows daily upward and downward trends following the demand and also the supply availability. Therefore, there is a lot of underlying variability in data that simply reflects the diurnal patterns of human activities, and not reflecting the volatility. Furthermore, regarding the second point (non-positive values), the traditional approach in financial time series analysis to switch to logarithmic measures are impractical, without shifting up all the values by a certain threshold. On the other hand, price volatility has a dependence on the price level, which is even more pronounced when the spot prices are low. Therefore, with respect to the magnitude of the aforementioned threshold, results can vary drastically. We hence have explored an alternative approach by representing the market data as matrices rather than time series. A novel and generic notion of volatility were accordingly defined using a well-known and numerically stable matrix decomposition technique, namely the singular value decomposition (SVD), combined with Haar wavelet transforms.

Our observations indicate a price volatility reduction and also prominent changes in the day-ahead price profiles, in recent years. There is an overall downward trend in the average electricity price. This undoubtedly has a number of causes, but the increasing penetration of subsidized solar and wind power account for at least part of it. Moreover, the traditional 12h00 peak before the Energiwende (Energy switch) is flattened out and shifted to earlier hours in the morning at 9h00. In a similar manner, the afternoon peak price hours have shifted one hour, from 19h00 to 20h00. Indeed, it is possible to clearly trace the impact of solar on the change of the daily price profile over the year (this effect is most pronounced in summer). Furthermore, the effect of the growth in wind power is most transparent in the shift in the distribution of low and negative prices during the day.

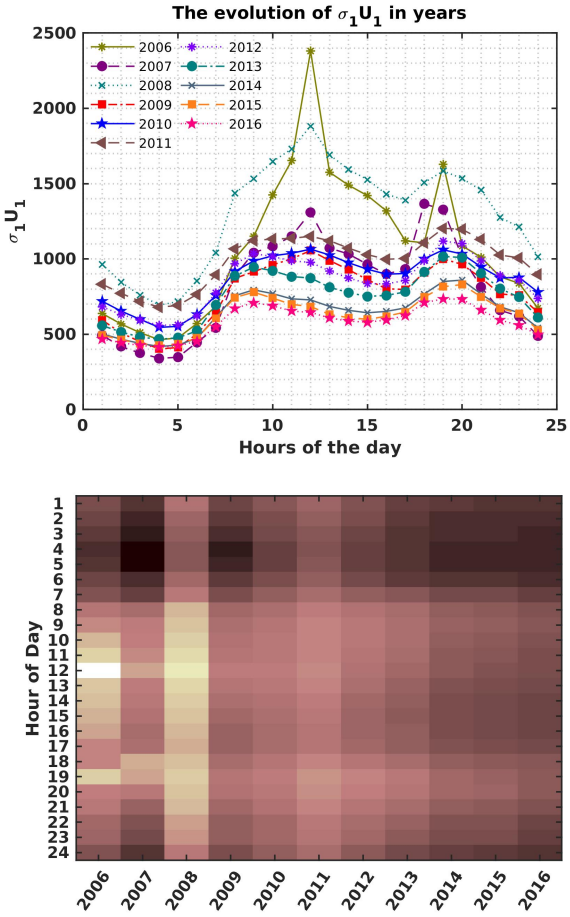


Figure 6.15: Top: Averaged daily profiles of the day-head prices. Bottom: An alternative representation, for the sake of better visualization.

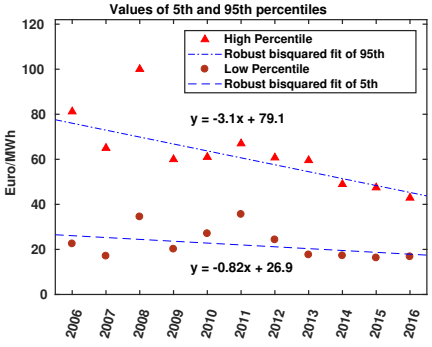


Figure 6.16: The evolution of extreme prices shows a consistent downward trend in the data.

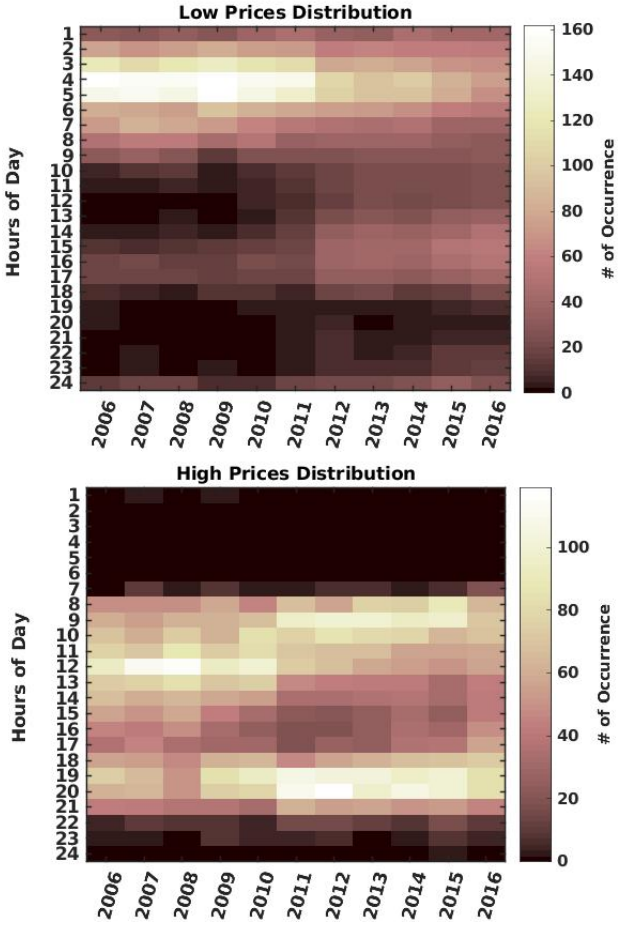


Figure 6.17: An overview of the distribution of low prices over the day (top), vs. high prices (bottom), throughout different years.

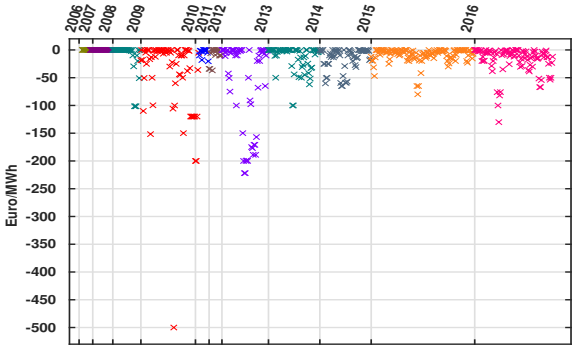


Figure 6.18: An overview of the occurrences of zero or negative prices.

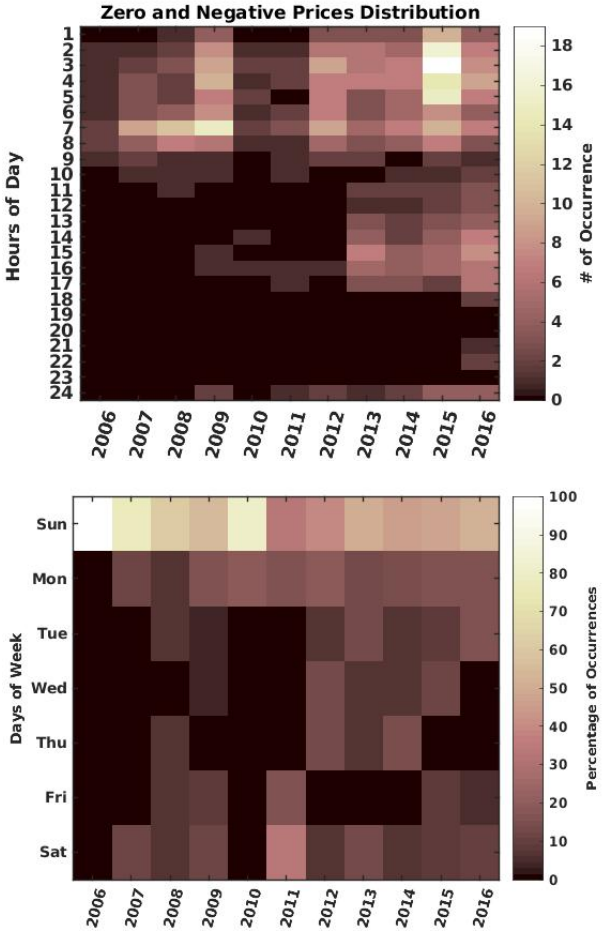


Figure 6.19: Top: The abundance of zero as well as negative prices during the period of absence of sunlight provokes the higher impact of wind than solar in this matter. Conversely, there are no negative prices in the evening hours 19h - 23h, as consumption is high, and solar input has vanished. Bottom: The percentage of the distribution of the zero or negative prices on the days of week.

REFERENCES

[1] A. Dorsman, A. Khoshrou, and E. J. Pauwels, *The influence of the switch from fossil fuels to solar and wind energy on the electricity prices in germany,* (2016).

[2] A. Khoshrou and E. J. Pauwels, *Quantifying volatility reduction in german day-ahead spot market in the period 2006 through 2016,* in *2018 IEEE Power & Energy Society General Meeting (PESGM)* (IEEE, 2018) pp. 1–5.

[3] A. Khoshrou, A. B. Dorsman, and E. J. Pauwels, *The evolution of electricity price on the german day-ahead market before and after the energy switch,* Renewable Energy **134**, 1 (2019).

[4] *Information portal renewable energy,* http://www.erneuerbare-energien.de/EE/Navigation/DE/Service/Erneuerbare_Energien_in_Zahlen/Zeitreihen/zeitreihen.htm.

[5] *Epexspot, european power exchange,,* http://www.epexspot.com/en/market-coupling ().

[6] *Financial chaos theory,* http://quantonline.co.za/Articles/article_volatility.htm.

[7] R. T. Baillie, C.-F. Chung, and M. A. Tieslau, *Analysing inflation by the fractionally integrated arfima–garch model,* Journal of applied econometrics , 23 (1996).

[8] M. d. C. Ruiz, A. Guillamón, and A. Gabaldón, *A new approach to measure volatility in energy markets,* Entropy **14**, 74 (2012).

[9] I. Simonsen, *Volatility of power markets,* Physica A: Statistical Mechanics and its Applications **355**, 10 (2005).

[10] F. L. Alvarado and R. Rajaraman, *Understanding price volatility in electricity markets,* in *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on* (IEEE, 2000) pp. 5–pp.

[11] A. Khoshrou, A. B. Dorsman, and E. J. Pauwels, *Svd-based visualisation and approximation for time series data in smart energy systems,* in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* (2017) pp. 1–6.

[12] *Epexspot, day-ahead auction,* https://www.epexspot.com/en/product-info/auction/germany-austria ().

[13] B. Cornélusse, *How the european day-ahead electricity market works,* (2014).

[14] L. C. G. Rogers and S. E. Satchell, *Estimating variance from high, low and closing prices,* The Annals of Applied Probability , 504 (1991).

[15] C. Torrence and G. P. Compo, *A practical guide to wavelet analysis,* Bulletin of the American Meteorological society **79**, 61 (1998).

[16] I. Daubechies, *The wavelet transform, time-frequency localization and signal analysis,* IEEE transactions on information theory **36**, 961 (1990).

[17] *Federal statistical office of germany,* https://www.destatis.de/EN/FactsFigures/EconomicSectors/Energy/Production/Tables/GrossElectricityProduction.html.

7

CONCLUSION

In a world replete with observations (physical as well as virtual), many data sets are represented by time series. In its simplest form, a time series is a set of data collected sequentially, *usually* at fixed intervals of time. In a number of applications, the mean and the variance of the time series is time-invariant and there is no seasonality in the data (such time series is called stationary). However, in many more applications, e.g., time series that are related to smart energy systems, the observed data often have non-stationary characteristics. For instance, whereas the electrical consumption of households is similar throughout the week, it shows a markedly different consumption pattern in the weekend.

An important research thread of the work in this thesis is the introduction of an alternative representation (as matrices) for such a time series. This offers some advantages when it comes to the analysis of these types of data. The rationale is straightforward: we then can use matrix factorization techniques to address different problems in a robust, numerically-stable manner. In particular, in this thesis, we have focused on the singular value decomposition (SVD) as a powerful, numerically stable matrix factorization technique which is then applied to time series analysis. That in turn has enabled us to look at different applications in time series analysis from a fresh perspective.

7.1. MAIN CONTRIBUTIONS

As announced in Chapter 1, one of the earliest applications of SVD in time series analysis is to detect periodicity and the number of *components* in the time series. To be more precise, in the literature, the ratio of the first to the second singular values has been introduced as a reliable measure to detect the periodicity in a time series. However, what has not been much appreciated is that the mean level of the data as well as the number of observed cycles (which determines the dimensions of the data matrix) affects the distributions of the singular values. Especially the latter case is more relevant to us, as the matrix size is not fixed and the number of columns can increase in time by adding new data. In this chapter, we also have provided an example of a complex time series where

the Fast Fourier Transform (FFT) fails to correctly determine its periodicity. We also have provided an introduction to different applications of the SVD in time series analysis.

We harked back to the SVD approach and the relevant theorems in Chapter 2. Furthermore, we have extensively studied the SVD and its geometrical interpretation to acquire a firm understanding of how it performs. We also have provided a number of examples of how the position with respect to the origin and the alignment of data points affects the singular vectors and accordingly the singular values. We also have explained how the results of the SVD and PCA are related to one another. This chapter also provides an intuitive answer to *Research Question 1*. However, an in-depth discussion of the first two research questions is provided in the next chapter. In this thesis, we make extensive use of simulation and SVD-based computations. As a consequence, accurate sampling from various matrix distributions is important. Through comprehensive experiments, we have become aware of certain biases and artefacts. We have concluded this chapter by pointing out the presence of such artefacts in the implementations of the algorithms in Matlab and Python.

For most applications of the SVD in various fields, it is important to understand the properties of SVD of a matrix whose entries show some degree of random fluctuations. Therefore, in order to determine how the noise level affects the singular value spectrum, it is essential to study the singular value decomposition of random matrices. Having provided a background in random matrices, Chapter 3 addressed *Research Question 1* and *Research Question 2* in full detail. In this chapter, we have proved some more properties of the SVD. In particular, we have provided estimates on how the drift in a simple periodic time series can affect its singular values.

The SVD and PCA techniques are both conceptually simple and effective. However, it is well-known that they are sensitive to the presence of noise and outliers in input data. In the literature, some modifications of the original algorithms of SVD and PCA have been proposed to alleviate the effect of these disturbances. In particular, one way to mitigate this sensitivity is to introduce *regularisation* terms. To this end, in Chapter 4 we first hark back to interpreting PCA in terms of low-rank approximations. We then added regularisation terms to its *functionals* and devised a solution algorithm for the new constrained optimization problem. This offers a solution to *Research Question 3a*.

We next turned our attention to the SVD factorisation and *Research Question 3b*. Algorithm 1 proposes a solution to a simpler case where only one *regularisation* term was added. In Algorithm 2 we offer a solution for the more general case. We then have shown how to tackle the computational aspects of the random and gradient descent techniques. To this end, we used ideas from Lie-group and -algebra to come up with a convenient parametrisation of the search problem. We have concluded this chapter by providing some examples of how regularisation can enable us to enhance the underlying patterns in the data.

With the increasing integration of renewable energy sources (RES) such as wind and solar energy into the power grid, balancing the grid has become more challenging. It is mostly due to the inherently intermittent nature of RES, on the one hand, and shortcomings in bulk energy storage systems, on the other. Therefore, studies on *scenario-based* probabilistic energy production and demand forecasts have gained momentum, as they are highly valuable from both a technical and an economic point of view. A particular

application of such models in the energy sector is where having the distribution of the energy consumption for the coming days is desired. The performance of such models evidently depends to a large extent on how different input (temperature) scenarios are being generated. There are mainly three practical and popular methods for generating temperature scenarios, namely fixed-date, shifted-date, and bootstrap approaches. Nevertheless, these methods have mostly been used on an ad-hoc basis without being formally compared or quantitatively evaluated. Chapter 5 provides a data-driven solution for *Research Question 4*. In this chapter, we proposed a generic framework for probabilistic load forecasting using an ensemble of regression trees. A major distinction of the current work was in using matrices as an alternative representation for quasi-periodic time series data. The SVD technique was then used to generate temperature scenarios in a robust and timely manner. The strength of our proposed method lies in its simplicity and robustness, in terms of the training window size, with no need for subsetting or thresholding to generate temperature scenarios. The empirical case studies performed on the data from the load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L) show that the proposed method outperforms the top two scenario-based models with a similar set-up.

In Chapter 6, we investigated *Research Question 5*, i.e., what effect the transition of energy to RES can have on the overall trend and also the volatility of the electricity prices. As it was exemplified in this chapter, the emergence of non-positive price values in the energy transition era has introduced new challenges in the electricity market volatility analysis. More precisely, traditional approaches to switch to logarithmic measures can only be done after shifting up all values above zero by a certain threshold. However, price volatility has a dependence on the price level, which is even more pronounced when the spot prices are low. Therefore, the aforementioned *pre-processing* step can affect the final outcome. In other words, the generalizability of conventional approaches can be questioned, as the volatility measures can vary drastically, with respect to the magnitude of the thresholds mentioned above. The first part of this chapter offers a solution to *Research Question 5a* by introducing a new notion of volatility which was obtained by reconstructing the time series using the SVD. Our observations indicate price volatility reduction, in the day-ahead market, in the years 2006-2016. The second part of this chapter addressed *Research Question 5b*; it provided shreds of evidence of the effect of renewables on daily price profiles – the emergence of non-positive prices and shifts of peak price values to hours where solar is less available.

7.2. CONCLUDING REMARKS AND FUTURE WORK

In this thesis, we have argued that the well-known singular value decomposition (SVD) (which is usually applied to matrix problems) can also successfully be applied to identify the periodic patterns in time series. Furthermore, these profiles are completely defined by the data and do not require the specification of user-defined parameters, apart from the period (which itself can be estimated using this approach). As such, this methodology offers a purely data-driven approach to adaptive signal approximation. Our findings can be used as innovative components of future smart grid systems, which are characterized by the increasing uncertainty on both the supply and demand parts.

An important topic for further research would be to find ways in which the gradient

descent procedure used in Algorithms 1 and 2 of Chapter 4 can be accelerated by taking advantage of the fact that the functional is very smooth and locally approximately quadratic. It would also be useful to derive some estimates for appropriate values for the weights λ and μ in terms of noise characteristics corrupting the underlying signal. Finally, although the P matrix in Algorithm 2 has unit-length columns, unlike the standard SVD, we were unable to confirm the orthogonality of the P vectors, i.e., $P^T P \neq I_k$. In fact, numerical experiments seem to indicate that such a constraint is not compatible with the minimisation of the functional. This requires further theoretical elucidation.

A

APPENDIX

A.1. BRIEF OVERVIEW OF MATRIX NORMS

Matrix norms come in two flavours (more details are discussed below):

- **Vector interpretation: entry-based norm** The matrix is seen as a vector (generalisation of n-tuples), and the norms are based on the values of the matrix entries.
- **Operator interpretation: operator or induced Norm** A matrix can also be interpreted as representing a linear transformation, and the norm is associated with the effect of the linear transformation on vectors.

VECTOR INTERPRETATION (ENTRY-WISE NORMS)

In this case we simply interpret a $p \times q$ matrix as an pq-tuple and use the corresponding **vector** norm. For instance, for $A \in \mathbb{R}^{p\times q}$:

- L_2 norm (squared for notational convenience):

$$\|A\|_2^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} |a_{ij}|^2 = Tr(A^T A) = Tr(AA^T)$$

  This norm is also called the Frobenius norm (which is the same as the sum of squares of the singular values.).

- L_1 norm:

$$\|A\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{q} |a_{ij}| = \mathbf{1}_p^T |A| \mathbf{1}_q$$

  where $\mathbf{1}_n$ is a column matrix of length n for which all entries are equal to 1.

- L_∞ norm:

$$\|A\|_\infty = \max_{i,j} |a_{ij}|$$

- Small entries in a **vector** contribute more to the 1−norm of the vector than to the 2−norm. That is in contrast to the contribution of large entries in a vector to the 1-norm and 2-norm.

OPERATOR INTERPRETATION (INDUCED NORM)

In this case the matrix norm is induced by the norm(s) for vectors. More specifically, the matrix norm is determined by the maximal (amplification) effect the linear transformation can have on any vector. Because of linearity we can restrict our attention to the effect on vectors of unit norm. Assuming that we are working in a finite dimensional vectorspace (V) equipped with a norm $\|\cdot\|_V$, the induced operator norm (which will be denoted as $\|\cdot\|_{(V)}$) becomes:

$$\begin{aligned} \|A\|_{(V)} &:= \max\left\{\frac{\|A\mathbf{x}\|_V}{\|\mathbf{x}\|_V} : \mathbf{x} \in V_0\right\} \\ &= \max\{\|A\mathbf{x}\|_V : \mathbf{x} \in V, \|x\|_V = 1\} \end{aligned}$$

This equation has a straightforward geometrical interpretation. The linear transformation characterized by A transforms the unit sphere into an ellipsoid. The induced norm is equal to (half) the length of the maximal principal axis of this ellipsoid.

- L_2 norm (spectral norm): is in fact the largest singular value of the matrix. The 2-norm is the square root of the sum of squared distances to the origin along the direction that maximizes this quantity.
- L_1 norm:
$$\|A\|_1 = \max_j(\sum_{i=1}^{q} |a_{ij}|) \quad where \quad j = 1 \ldots p$$
- In general, if one splits a matrix A into its column vectors: $A_{p\times q} = [A_1, A_2, \ldots A_p]$ the the one-norm of A in the maximum of the one-norms of the column vectors A_i of A.
$\|A\|_1 = \max\{\|A_i\|_1 : A_i \text{ is a column vector of A }\}$
- Similarly, in general if one splits a matrix A into its row vectors:

$$A_{p\times q} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_q \end{bmatrix}$$

  then the ∞−norm of A is the maximum of the one-norms of the row vectors A_j of A.
$\|A\|_\infty = \max\{\|A_j\|_1 : A_j \text{ is a row vector of A }\}$
- L_2 norm: is in fact the largest singular value of the matrix. The 2-norm is the square root of the sum of squared distances to the origin along the direction that maximizes this quantity.

A.2. SVD SOLVES A MATRIX NORM OPTIMISATION PROBLEM

There are some essential connections between matrix norms and SVD. In this context we will explore two:

1. The L_2 matrix norms can be expressed in terms of the singular values;
2. Using L_2 matrix norms as an objective function, SVD provides the solution to a non-convex optimization problem.

We will look at both in turn.

A.3. L_2 MATRIX NORMS EXPRESSED IN TERMS OF SINGULAR VALUES

Suppose A is an $n \times p$ matrix which has r non-zero singular values, arranged in decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$$

Theorem 10. *Both L_2 norms (Frobenius and spectral) can be expressed in terms of the singular values. More specifically:*

1. *Frobenius norm (entry-wise):* $\|A\|_F^2 = \sum_i \sigma_i^2$.
2. *Spectral norm (induced):* $\|A\|_2^2 = \sigma_1^2$.

Proof. The proofs amount to straightforward calculations:

1. Frobenius:

$$\|A\|_2^2 = Tr(A^T A) = Tr(VS^T U^T USV^T) = Tr(VS^T SV^T) = Tr(S^T S) = \sum_i \sigma_i^2$$

2. Spectral:

$$\max |Av|$$

where $|v| = 1$ is a unitary matrix. Previously we saw that v_1 maximize this argument and the result of that is σ_1^2.

□

SVD SOLVES THE LOW-RANK APPROXIMATION PROBLEM

The connection between SVD and matrix norms is given by the fact that the SVD provides the solution to the question: *for a given matrix A, find the best approximation for a pre-specified rank.* The notion of matrix norm enters when we need to specify what we mean by "best". Reformulating the problem in a more precise manner, we arrive at:

Lemma 11. *Matrices A and B are identical if and only if for all vectors* $\mathbf{v}$, $A\mathbf{v} = B\mathbf{v}$.

Related results For any matrix A, the sequence of singular values is unique and if the singular values are distinct, the the sequence of singular vectors is also unique. However, when some set of singular values are equal, the corresponding singular vectors span same subspace. Ant set of orthonormal vectors spanning this subspace can be used as the singular vectors.

Lemma 12. *Let A_p be defined as above, then the extension of the earlier expression of the norm in terms of the singular values are given by:*

- $\|A - A_p\|_2^2 = \sum_{k=p+1}^{r} \sigma_k^2$
- $\|A - A_p\|_2^2 = \sigma_{p+1}^2$

A.4. GRADIENTS FOR FROBENIUS NORM

Suppose that $A \in \mathbb{R}^{p\times k}$, $X \in \mathbb{R}^{k\times q}$ and $B \in \mathbb{R}^{p\times q}$ and define the real-valued function (based on the Frobenius norm):

$$f(X) = \|AX + B\|_F^2$$

Then we have the following gradient:

$$\nabla_X f = 2A^T(AX + B) \tag{A.1}$$

Similarly,

$$g(X) = \|XA + B\|_F^2 \quad \Longrightarrow \quad \nabla_X g = 2(XA + B)A^T \tag{A.2}$$

Applying eqs. (A.1) and (A.2) to the J_1 functional we get:

$$\frac{1}{2}\nabla_U J_1 = -(A - UV^T)V + \lambda U$$

and

$$\frac{1}{2}\nabla_V J_1 = (VU^T - A^T)U$$

A.4.1. SOME SPECIAL CASES

- For $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{a} = (\mathbf{a}^T\mathbf{x})^2$, then

$$\nabla_x f = 2(\mathbf{a}^T\mathbf{x})\mathbf{a} = 2(\mathbf{x}^T\mathbf{a})\mathbf{a}$$

- For For $f(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x}$ where Q symmetric,

$$\nabla f = 2Q\mathbf{x}$$

A.5. VARIANCE OF PRODUCT

- If X and Y are independent then:

$$Var(XY) = Var(X)Var(Y) - Var(X)(\mathbb{E}(Y))^2 - Var(Y)(\mathbb{E}(X))^2. \tag{A.3}$$

If both $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, then

$$Var(XY) = Var(X)Var(Y) \tag{A.4}$$

A.6. SINGULAR VALUES OF THE "FAT" MATRICES

- A square matrix represents a mapping from a space into itself (or another space of the same dimension). Under such a mapping the unit sphere is mapped to an ellipsoid.

- A fat matrix represents a linear mapping from a higher dimensional space into a lower one. This means that the kernel is non-trivial. It also means that the unit sphere is mapped into a solid ellipsoid (i.e. including the interior). See fig right

What we observe from these experiments is that when we increase the aspect ratio (make the matrix fatter), increasingly more unit vectors are mapped to the interior (making the outline of the ellipsoid more difficult to spot.

In addition, the image becomes less correlated (correlation coefficient of data points decreases. This means that the singular values become more alike: a perfectly spherical image would result from the identity matrix which has identical singular values. So it seems to me that the explanation for the observed behaviour of the singular values must run along the following lines:

- Increasing the aspect ratio (making the matrix fatter) results in an increase of the Frobenius norm (after all, we are summing over more matrix entries). Hence this implies that the sum of singular values has to increase. This can be done by increasing the incline of the line of singular values, or by shifting this line upwards parallel to itself (the latter is what we observe).

- The above experiments suggest that increasing the aspect ratio results in an image of the unit-sphere that is increasingly more circular which would suggest that the ratio between the smallest and largest singular value decreases — this rules out that the inclination of the singular values line increases, but does correspond to a uniform increase in all the singular values.

- Basically Figure A.1 confirm that for a set of unit vectors e_i on $n-$dimensional space i.e., $\sum_{j=1}^{n} e_{ij}^2 = 1$, if we increase the dimension $n \to \infty$ then the $e_i \to \mathcal{N}(0,\sigma)$.

Theorem 13. *If* $\mathbf{u}_i \quad (i = 1,\ldots,N)$ *are unit vectors in* n*-dimensional space* $\mathbb{R}^n$*, sampled uniformly on the unit sphere* $S^{n-1} \subset \mathbb{R}^n$*.*
Then (denoting vectors as columns):

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i \mathbf{u}_i^T \longrightarrow \frac{1}{n} I_n \qquad as \quad N \longrightarrow \infty. \tag{A.5}$$

Another way of formulating this would be:

Let $\mathbf{u}_1, \mathbf{u}_2 \sim Unif(S^{n-1})$ *uniform and independent on the unit-sphere in* $\mathbb{R}^n$*, then*

$$\mathbb{E}(\mathbf{u}_1^T \mathbf{u}_2) = 0 \qquad while \qquad \mathbb{E}(\mathbf{u}_1 \mathbf{u}_1^T) = \mathbb{E}(\mathbf{u}_2 \mathbf{u}_2^T) = \frac{1}{n} I_n$$

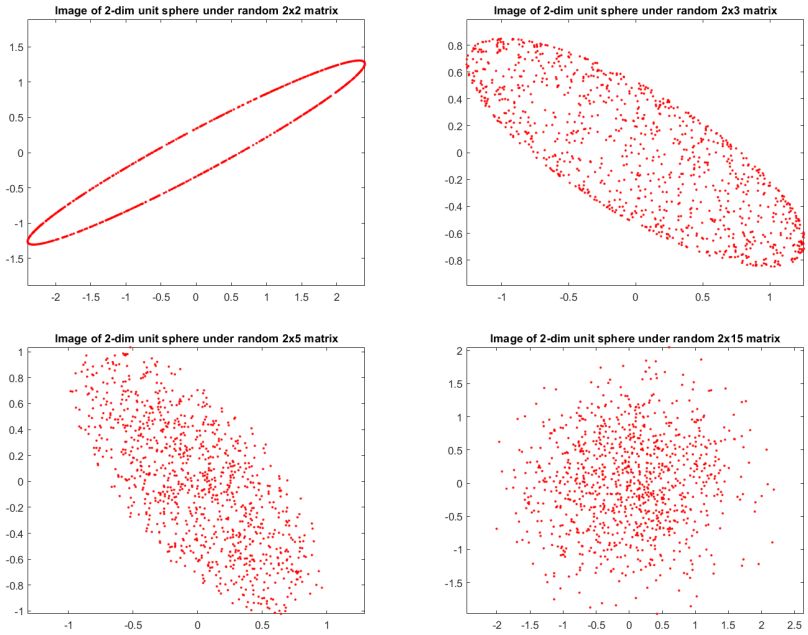


Figure A.1: Image of 1000 random points on the unit sphere under a random transformation (unit normal). Top left: Random 2×2 matrix. Top right; 2×3. Bottom right: 2×15.

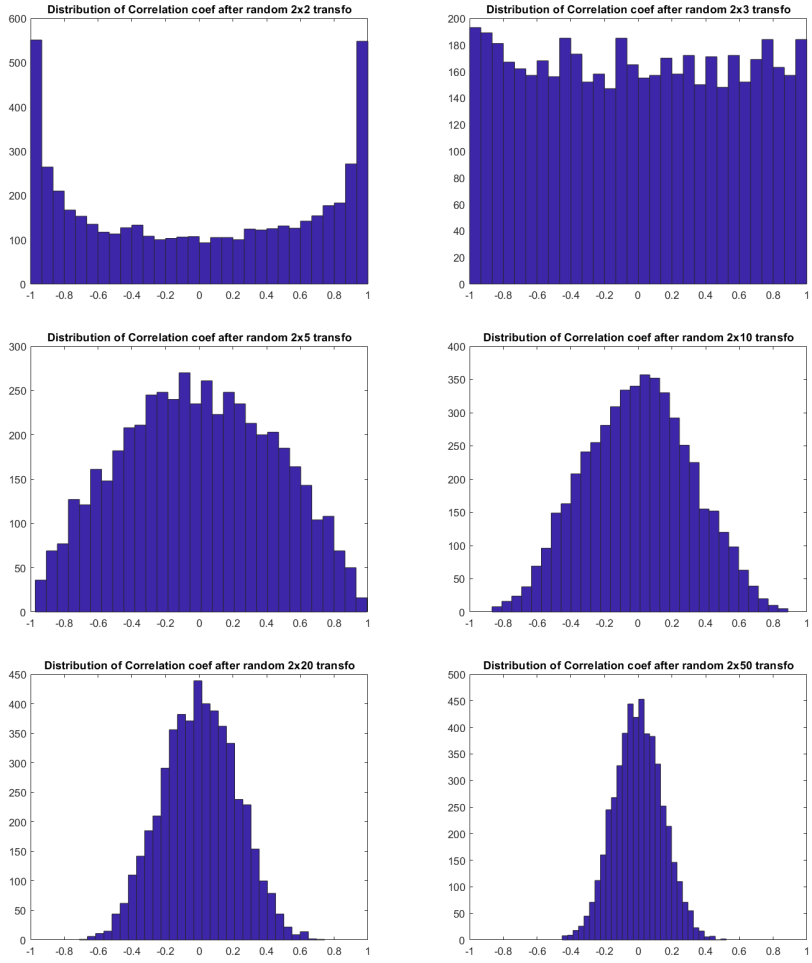


Figure A.2: Distribution of correlation coefficient of the images of 1000 uniformly distributed unit vectors after random transformation of n-dimensional space into 2-dim space ($n = 2, 3, 5, 10, 20, 50$). Clearly, higher dimensional space tend to project down to less correlated data. This confirms the progression seen in Fig. A.1.

Outline of proof

- If we would take $\mathbf{u}_i = \mathbf{e}_i$ to be the standard unit vectors (in which case $N = n$) the asymptotic result turns into an equality. This follows immediately from the observation that $\mathbf{e}_i\mathbf{e}_i^T$ is a matrix with a single non-zero entry (equal to 1) on the i^{th} diagonal position.

- Each term in the LHS sum in the LHS always has trace equal to 1 (exactly), independent of N. This follows from the fact that for any rank-1 matrix:

$$Tr(\mathbf{ab}^T) = \sum_i (\mathbf{ab}^T)_{ii} = \sum_i a_i b_i = \mathbf{a}^t\mathbf{b} \qquad \text{and hence:} \qquad Tr(\mathbf{u}_i\mathbf{u}_i^T) = \mathbf{u}_i^T\mathbf{u}_i = 1.$$

  As a consequence:

$$Tr\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i\mathbf{u}_i^T\right) = \frac{1}{N}\sum_{i=1}^{N} Tr(\mathbf{u}_i\mathbf{u}_i^T) = 1. \tag{A.6}$$

- Notice that the diagonal elements for every matrix $\mathbf{u}_i\mathbf{u}_i^T$ are just the square entries of $\mathbf{u}_i$:

$$diag(\mathbf{u}_i\mathbf{u}_i^T) = (u_{i1}^2, u_{i2}^2, \ldots, u_{in}^2)$$

  Since the distribution of the unit vectors is uniform over the sphere, the induced distribution of the (squared) entries will be independent of the position along the diagonal (if this weren't the case, then there would be a preferred direction space, which contradicts the uniformity). Hence, the sum (from 1 to N for each diagonal element, converges to the same value. This in combination with the fact that the trace needs to be one, shows that at least on the diagonal, Eq. (A.7) holds.

- The off-diagonal elements of each rank-1 matrix $(\mathbf{uu}^T)_{k\ell} = u_k u_\ell$ also have some structure.

  Hence this means that the $k\ell$ element of the LHS (where $k \neq \ell$) is a (growing) sample mean and therefore converges to the mean of the corresponding mean of the stochastic variable $U_k U_\ell$ (where we use the capital notation to indicate the stochastic component variables that result from a drawing a unit vector $\mathbf{u} = (U_1, U_2, \ldots, U_n)$ uniformly on the unit-sphere:

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i\mathbf{u}_i^T\right)_{k\ell} \longrightarrow \mathbb{E}(U_k U_\ell) \qquad \text{as} \quad N \longrightarrow \infty. \tag{A.7}$$

  Notice that if the components were independent, then this would clinch the proof because $\mathbb{E}U_i = 0$ (symmetry), assuming independence, $\mathbb{E}(U_k U_\ell) = \mathbb{E}(U_k)\mathbb{E}(U_\ell) = 0$. However, independence does not hold since $\sum_{k=1}^{n} U_k^2 = 1$. So there is a weak dependence that becomes weaker as the dimension of the ambient space grows.

- Still, we can expect $\mathbb{E}(U_k U_\ell) = 0$ because if that weren't the case, then there would a non-zero covariance between different components:

$$Cov(U_k, U_\ell) = \mathbb{E}(U_k U_\ell) - \mathbb{E}(U_k)\mathbb{E}(U_\ell) = \mathbb{E}(U_k U_\ell) \neq 0$$

but that seems to contradict the uniformity on the unit-sphere.

- Further to the above item: the dependence between the components is strongest in the low dimensional spaces and gets weaker in high dim.

The strongest dependence is for unit vectors in 2-dim space. Sampling from the unit-circle in 2-dim amounts to

$$\mathbf{u} = (u_1, u_2) = (\cos(\theta), \sin(\theta)) \qquad \text{where} \quad \theta \sim U(-\pi, \pi).$$

Hence,

$$\mathbb{E}(u_1 u_2) = \mathbb{E}(\cos(\theta)\sin(\theta)) = \frac{1}{2}\mathbb{E}(\sin(2\theta) = 0 \qquad \text{since} \quad 2\theta \sim U(-2\pi, 2\pi)$$

- I think all of the above are sufficient to construct a valid proof. At least it explains why the result is true.

Let $U = (U_1, U_2, \ldots, U_n)$ where each $U_i \in S^{n-1} \subset \mathbb{R}^n$ lies on the unit sphere. We want to prove that $\mathbb{E}(U_i U_j) = 0$ if $i \neq j$.

Proof: We can construct U as follow:

$$U = \frac{X}{\|X\|} \quad \text{when} \quad \mathscr{X} \sim \mathscr{N}(0, I_n) \tag{A.8}$$

$$\mathbb{E}(U_i U_j) = \mathbb{E}(\frac{X_i}{\|X\|}\frac{X_j}{\|X\|}) = \mathbb{E}(\frac{X_i X_j}{\|X\|^2}) \tag{A.9}$$

For the 2-dimensional case we have:

$$X = (X_1, X_2) \quad X_1, X_2 \text{ independent random variables in } \mathscr{N}(0, \sigma)$$

Let us assume the following:

$$\begin{cases} P = X_1 X_2 \\ R^2 = X_1^2 + X_2^2 = Q \end{cases}$$

Therefore we have

$$\begin{cases} Q - 2P = (X_1 - X_2)^2 \geq 0 \\ Q \geq 2P \\ Q + 2P = (X_1 + X_2)^2 \geq 0 \rightarrow Q \geq -2P \end{cases}$$

From the above formulas we can derive at the following:

$$\begin{cases} p = X_1 X_2 \\ q = X_1^2 + X_2^2 \end{cases}$$

or

$$\begin{cases} q + 2p = (X_1 + X_2)^2 \\ q - 2p = (X_1 - X_2)^2 \end{cases}$$

$\varphi(p,q) = \frac{1}{2\pi} e^{-(X_1^2+X_2^2)/2} |\frac{\partial(X_1,X_2)}{\partial(p,q)}|$ since $X_1 = X_1(p,q), X_2 = X_2(p,q)$ therefore,

$$\begin{cases} X_1 + X_2 = \sqrt{q+2p} \\ X_1 - X_2 = \sqrt{q-2p} \end{cases}$$

By summing up and subtracting the two equations, we derive:

$$\begin{cases} 2X_1 = \sqrt{q+2p} + \sqrt{q-2p} \\ 2X_2 = \sqrt{q+2p} - \sqrt{q-2p} \end{cases}$$

hence

$$|\frac{\partial(X_1,X_2)}{\partial(p,q)}| = \frac{1}{2}(\frac{1}{\sqrt{q^2-4p^2}})$$

or

$$\varphi(p,q) = \frac{1}{4\pi} \frac{e^{-q/2}}{\sqrt{q^2-4p^2}}$$

therefore we conclude:

$$\mathbb{E}(\frac{X_1 X_2}{R^2}) = \mathbb{E}(\frac{P}{Q}) \tag{A.10}$$

therefore we have:

$$\int\int \frac{p}{q}\varphi(p,q)\,\mathrm{d}p\,\mathrm{d}q = \int_0^{\infty}\int_{\frac{-q}{2}}^{\frac{q}{2}} (\frac{p}{q}\frac{e^{-q}}{\sqrt{q^2-4p^2}}))\,\mathrm{d}p\,\mathrm{d}q = \int_0^{\infty}\frac{e^{-q}}{q}(\int_{\frac{-q}{2}}^{\frac{q}{2}}\frac{p}{\sqrt{q^2-4p^2}}\mathrm{d}p)\mathrm{d}q = 0$$

since the inner integral is the integral of an odd function over a symmetric interval. We hence can conclude that for $i \neq j$:

$$\mathbb{E}(U_i U_j) = \mathbb{E}(\frac{X_i X_j}{\|X\|^2}) = 0$$

- Some random musing: this result is somewhat counter-intuitive. For each $\mathbf{u}_i \mathbf{u}_i^T$ is essentially an orthogonal projection on the corresponding unit vector. So you assume that because of the uniform distribution, all these contributions cancel out and therefore the result would be the zero mapping rather than the unit mapping.

To explain the change of correlation in the figures above, we need to investigate the correlation of the images of random unit vectors under random linear transformation, This means that we are interested in images of the form:

$$\mathbf{f} = A\mathbf{u} \qquad \text{where} \quad A \text{ is } (m \times n) \text{ matrix with } m \leq n \text{ (fat)}$$

Combining N of these results in a matrix we get:

$$F = AU \qquad \text{where} \quad U = (\mathbf{u}_1, \ldots, \mathbf{u}_N) \text{ is } n \times N \text{ and } F = (\mathbf{f}_1, \ldots, \mathbf{f}_N) \text{ is } m \times N$$

To compute the correlation (covariance) of the image vectors in F we have to compute:

$$Cov(F) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i \mathbf{f}_i^T = \frac{1}{N} A \left(\sum_{i=1}^{N} \mathbf{u}_i \mathbf{u}_i^T \right) A^T \longrightarrow \frac{1}{n} A A^T.$$

Every entry of the correlation matrix is in fact the inner product of the corresponding two columns of the matrix X. More importantly, because the correlation matrix is symmetric and semi-positive (it contains the inner products) that guarantees that there are real, non-negative eigen values.

Using the SVD decomposition to rewrite $A_{m\times n} = \mathcal{U}_{m\times m}\mathcal{S}_{m\times n}\mathcal{V}^T_{n\times n}$ we get furthermore:

$$Cov(F) = \frac{1}{n}\mathcal{U}\mathcal{S}\mathcal{S}^T\mathcal{U}^T = \frac{1}{n}\mathcal{U}\,\mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)\,\mathcal{U}^T$$

The matrix product in the RHS only depends on m, so increasing the number of columns N does indeed reduce the covariance. By increasing N, the covariance matrix is not decreasing but becoming constant on the RHS! But we need to covariance matrix to go to the unit matrix.

So basically we need to show that

$$\frac{1}{n} diag(\sigma_1^2, \ldots, \sigma_m^2) \quad \longrightarrow \quad I_n \qquad \text{as} \quad n \longrightarrow \infty.$$

This indeed would imply that the singular values tend to the same values.

A.6.1. WHY ARE SINGULAR VALUES INFLATED?

Notice that if $A = (m \times n)$ (where $m < n$) and we make the SVD decomposition $A = \mathcal{U}\mathcal{S}\mathcal{V}^T$ then the last $n - m$ columns of V constitute an orthonormal basis for the null-space $Z := \ker A$.

- Notice that:

$$(AA^T)_{ij} = \sum_{k=1}^{n} A_{ik}A_{jk} \qquad \text{(inner product of } i^{th} \text{ and } j^{th} \text{ row of } A)$$

Similarly:

$$(A^T A)_{ij} = \sum_{k=1}^{m} A_{ki}A_{kj} \qquad \text{(inner product of } i^{th} \text{ and } j^{th} \text{ column of } A)$$

As a consequence:

$$(AA^T)_{ii} = \sum_{k=1}^{n} A_{ik}^2 \sim \chi_n^2 \qquad \text{(sum of independent squared standard normals)} \tag{A.11}$$

while

$$(A^T A)_{ii} = \sum_{k=1}^{m} A_{ki}^2 \sim \chi_m^2 \qquad \text{(sum of independent squared standard normals)} \tag{A.12}$$

- Introducing additional notation: Let $\mathscr{U}^i, \mathscr{U}_i$ be the i-th row, column respectively of $\mathscr{U}$, Moreover, we denote $\mathbf{e}_i = (0,\ldots,0,1,0,\ldots,0)^T$ we can use the SVD decomposition $A = \mathscr{U}\mathscr{S}\mathscr{V}^T$ to conclude:

$$\begin{aligned}(AA^T)_{ii} &= \mathbf{e}_i^T AA^T \mathbf{e}_i \\ &= \mathbf{e}_i^T \mathscr{U} diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)\mathscr{U}^T \mathbf{e}_i \\ &= \mathscr{U}^i diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)(\mathscr{U}^i)^T \\ &= \sigma_1^2 u_{i1}^2 + \sigma_2^2 u_{i2}^2 + \ldots + \sigma_m^2 u_{im}^2 = \end{aligned} \tag{A.13}$$

$$\sum_{k=1}^{m} \sigma_k^2 u_{ik}^2 \sim \chi_n^2 \quad \text{(cf. eq. A.11)} . \tag{A.14}$$

Since we know that the columns of U are orthonormal, it follows that also the rows are orthonormal (ref??): $UU^T = I_m \Longrightarrow U^T U = I_m$ (and vice-versa). This implies that

$$\sum_{k=1}^{m} u_{ik}^2 = 1,$$

hence we can interpret eq. (A.14) as a weighted mean of the squared singular values that needs (on average) be equal to n (the expected value of χ_n^2). Hence we see that if we increase n, then the values of σ_k need to increase as well. Furthermore, since the rows and columns of U are random, this increase cannot be shouldered by a small number of the singular values, but needs to happen across the board.

- We can make a similar argument:

$$(A^T A)_{ii} = \sum_{k=1}^{m} A_{ki}^2 \sim \chi_m^2$$

But also:

$$(A^T A)_{ii} = \mathbf{e}_i^T A^T A \mathbf{e}_i = \sigma_1^2 v_{i1}^2 + \ldots + \sigma_m^2 v_{im}^2$$

But notice that the factors in weighted sum do not add up to 1 (since $m < n$). Notice that increasing n means that fewer components of v enter in the sum, and therefore the singular values need to increase in order to keep the average fixed on m.

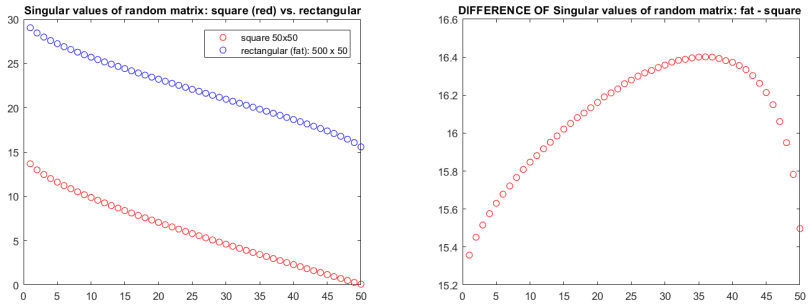


Figure A.3: The scaling factor seems to depend on the order of singular value. Left: difference between average sing values of random (standard normal) matrix for square (50 × 50) and fat (50 × 500) matrix.

A.7. PERTURBATION OF EIGEN-VALUES AND -VECTORS

A small change in the elements of a matrix can have a profound effect on a function of that matrix. Let Q be a symmetric $n \times n$ matrix with eigenvalue spectrum $\lambda(Q) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. The corresponding eigenvectors $\mathbf{u}_i$ form an orthonormal basis for $\mathbb{R}^n$:

$$Q\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad \text{and} \qquad \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}.$$

Now consider how a small perturbation of $Q + \epsilon dQ$ will affect the eigen-values and -vectors:

$$\lambda_i(\epsilon) = \lambda_i + \epsilon \mathbf{u}_i^T dQ \mathbf{u}_i + o(\epsilon^2) \tag{A.15}$$

$$\mathbf{u}_i(\epsilon) = \mathbf{u}_i + \epsilon \sum_{j \neq i} \frac{\mathbf{u}_j^T dQ \mathbf{u}_i}{\lambda_i - \lambda_j} + o(\epsilon^2) \tag{A.16}$$

Note that these formulas hold as long as the unperturbed and perturbed systems involve symmetric matrices, to guarantee the existence of N linearly independent eigenvectors.

Let Q be a symmetric matrix which therefore has a orthonormal basis of eigenvectors u_i with corresponding eigenvalues λ_i:

$$Qu_i = \lambda_i u_i.$$

Differentiating yields:

$$(dQ)u_i + Q du_i = (d\lambda_i)u_i + \lambda_i du_i.$$

or again:

$$(Q - \lambda_i) du_i = (d\lambda_i - dQ)u_i.$$

Left-multiplying by u_k^T and using the fact that Q is symmetric and therefore $u_k^T Q = \lambda_k u_k^T$ we obtain:

$$u_k^T (\lambda_k - \lambda_i) du_i = u_k^T (d\lambda_i - dQ) u_i.$$

We now consider the following two cases:

- $k = i$: in this case the LHS vanishes, and we get from the RHS:

$$d\lambda_i = u_i^T (dQ) u_i = (dQ)_{ii}.$$

- $k \neq i$: it then follows that:

$$u_k^T du_i = \frac{u_k^T (dQ) u_i}{\lambda_i - \lambda_k}.$$

Since the u_i constitute a basis we can expand:

$$du_i = \sum_{j \neq i} \epsilon_{ij} u_j \qquad \text{since (without loss of generality)} \qquad \epsilon_{ii} = 0.$$

Plugging this in the equation above we get:

$$du_i = \sum_{k \neq i} \frac{u_k^T (dQ) u_i}{\lambda_i - \lambda_k} u_k$$

A.7.1. PERTURBATION THEORY FOR MATRICES

Consider a symmetric matrix $Q = U\Lambda U^T$; differentiation yields:

$$dQ = dU\Lambda U^T + Ud\Lambda U^T + U\Lambda dU^T$$

Next we use the fact that every orthogonal matrix can be written as the exponential of a skew-symmetric:

$$U = e^K \quad (\text{where } K^T = -K) \qquad \text{and hence: } dU = UdK$$

Substituting this in the above yields:

$$U^T dQU = U^T dU\Lambda + d\Lambda + \Lambda(dU^T)U$$

$$U^T dQU = d\Lambda - (\Lambda dK - dK\Lambda)$$

Writing the above equation for diagonal and off-diagonal elements yields the correct perturbation for the eigen-values and -vectors.

B

APPENDIX

B.1. EPEX MARKET AND RES FEED-IN

The following section contains some evidences of the impact of the day-ahead estimated wind and solar feed-in on the price changes (Section 6.6), in the recent years.

B.1.1. DAY-AHEAD WIND ENERGY FEED-IN (IN GWH)

Fig. B.1 provides an overview of the evolution of the wind feed-in (day-ahead forecasts) over the years. The left panel in Fig. B.1 indicates a smooth annual growth. In a similar way, the right panel highlights the fact that the production is relatively constant around the clock. The change in the annual average of the daily profiles from 2014 afterwards is noticeable. Peak production of the wind profile is eventually moving from early afternoon to late night or early morning. This can become an issue for the stability of the grid, as there might not be enough demand during those particular hours.

B.1.2. DAY-AHEAD SOLAR ENERGY FEED-IN (GWH)

The developments in energy storage technologies and also the falling costs of harvesting solar power have made it increasingly attractive for the private households [**?**]. Fig. B.2 shows the day-time (non-zero values) solar energy feed-in forecast from 2010 to 2016. After the rapid rise in 2010 through 2013, solar feed-in has leveled off in the last two years. The panel on the right in Fig. B.2 illustrates the annual averaged solar feed-in for each time slot for years 2010-2016. Peak of solar feed-in is around 13h00; that coincides with the high demand during the day.

B.1.3. EVOLUTION OF GERMAN DAY-AHEAD PRICE DURING WINTER AND SUMMER

To illustrate the impact of solar energy on the price, we scrutinize the data separately for the summer (June through August) and winter (December through February) periods. In winters, days are shorter and the sun, if it emerges at all, traces out a lower path in the sky; therefore, a significantly smaller amount of solar energy is produced (Fig. B.3). Wind

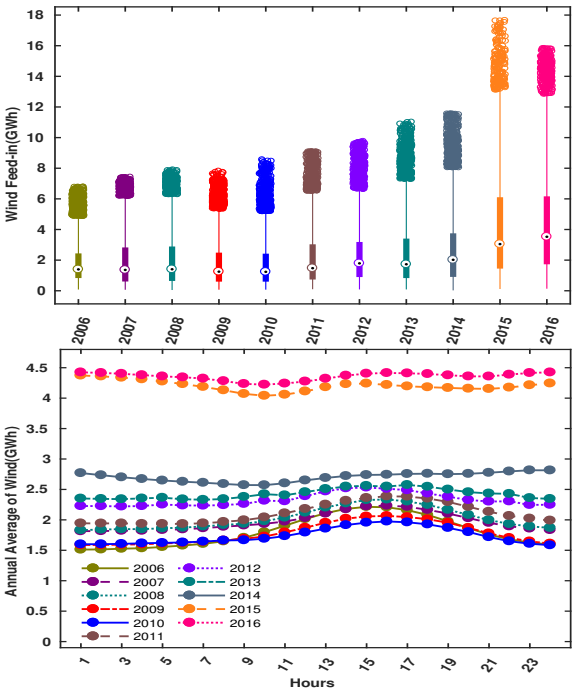


Figure B.1: Top: The consistent growth in wind energy feed-in over the years. Bottom: Annual average of the daily profiles.

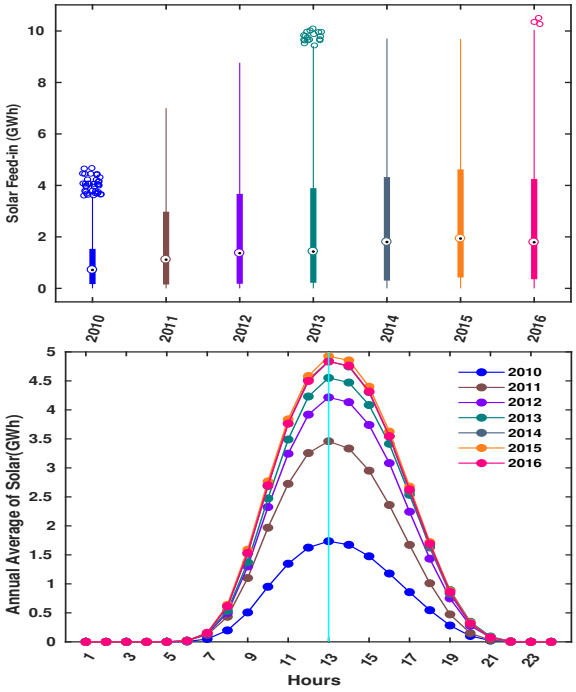


Figure B.2: Left: Smooth growth in annual solar energy feed-in (only non-zero day-time values have been considered). Right: Annual daily average of the solar feed-in.

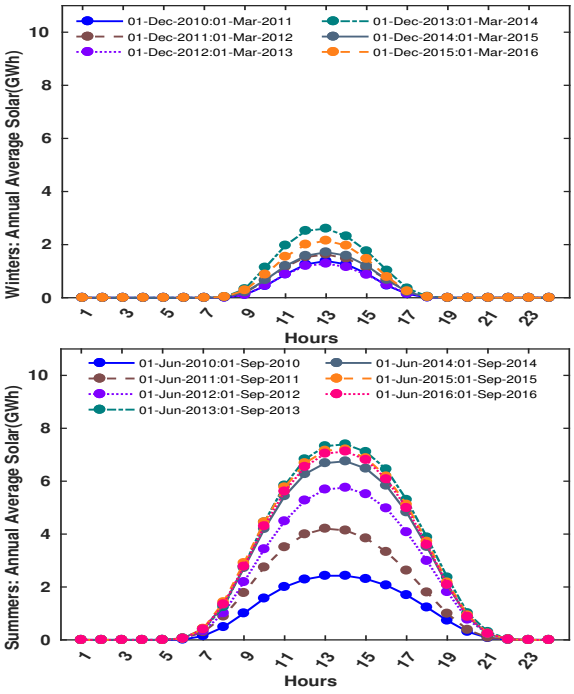


Figure B.3: Low production of solar and also shorter occurring hours in winters (left), vs. high amount and longer period of solar production during summer (right).

energy, on the other hand, is fairly constant throughout the day, but there are marked difference between the seasons (Fig. B.4). Fig. B.5 contrasts the evolution of the daily average of the price profile during winter (December through February) and summer (June through August) season. During the observed period we see that for winter time the peak at 19h00 is reduced both in size and sharpness, most likely due to the increase in the wind energy. During the summer period, the morning peak at 12h00 disappears completely over the years, in all likelihood again due to the increasing supply of wind and especially solar energy. In other words, the increasing supply of wind and solar energy is not only reducing the electricity price, but it is also changing the daily profile substantially.

Comparing the solar energy feed-in in winters and summers in Fig. B.3 and also considering the evolution of the price profiles in Fig. B.5 allow us to conclude that solar energy, especially in summer, effectively flattens the daytime price profile. Fig. B.5 highlights the evolution of the daily average of the price profiles during winter (December through February) and summer (June through August) season. Every value is the average of the prices for that specific hour, with the average ranging over the specified period. The left panel shows that during winter period the maximum values occur from 18h00 to 20h00, with peak at 19h00. Also, there is a steep increase in the morning (around 7h00). On the other hand, during summer (right panel), the price increase in the morning (5h00-9h00) is considerably flatter. Also the price-spike observed during

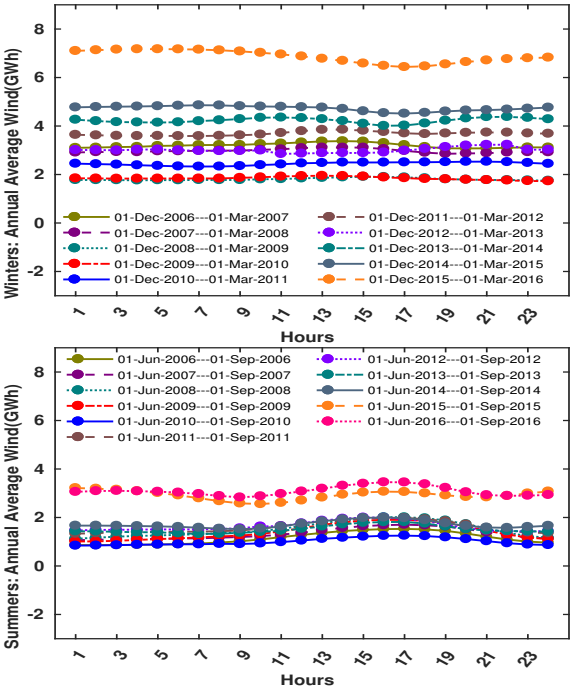


Figure B.4: Almost smooth and steady harvest of constant wind breathe in winters (left), vs. low production of wind in summer (right).

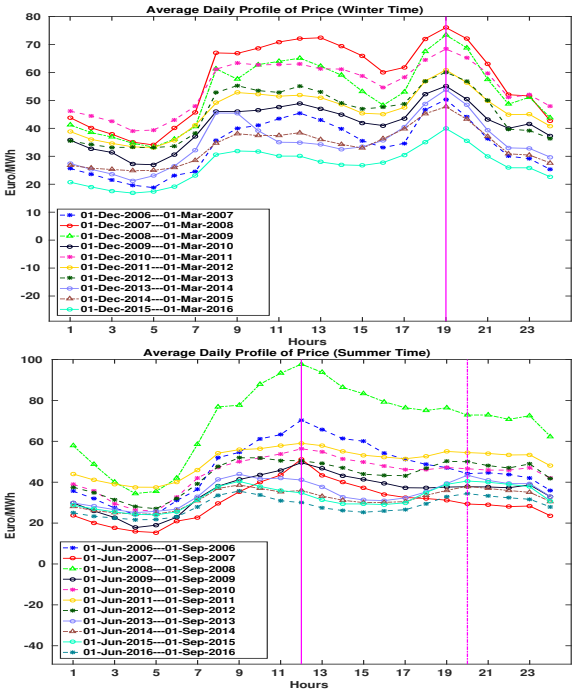


Figure B.5: The evolution of the seasonal daily average of the price profile during winter time (left) and summer time (right).

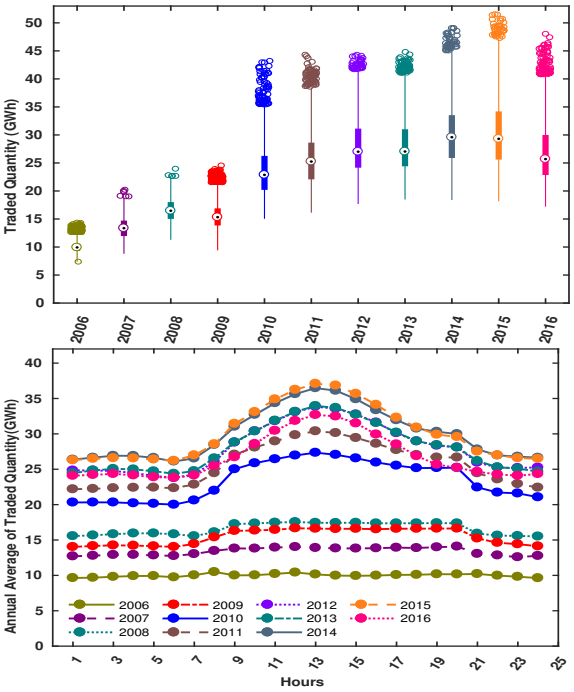


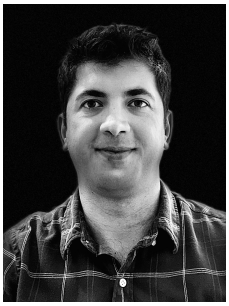
Figure B.6: Left: Boxplots for the hourly values of traded volumes on the day-ahead market. Right: Daily evolution of the traded volume for each hour slot.

winter evenings (around 19h00) is completely absent in summers. Both observations underscore the impact of solar on the price.

B.1.4. DAY-AHEAD TRADED QUANTITY (GWH)

Fig. B.6 displays the evolution of the traded quantity values on the German day-ahead market, in the recent years. Two interesting features are readily apparent. In the left panel the occurrence of a considerable number of outliers (represented as individual points near the upper part of the boxplots) point to unusually high volumes being traded. This highly resembles the wind feed-in profiles in Fig. B.1.1. The panel on the right depicts the annual averages of a typical daily profile. Again, the steady increase in the traded volume is evident. However, whereas in the first half of the decade, the traded volume is essentially constant over the course of the day, the latter part of the decade shows an increasingly more prominent bump that mirrors the average supply of solar energy, and could therefore be an indicator of surpluses generated by the renewable energy sources (in particular solar). In 2016, however, we witness a minor reduction in the traded volume, as it may be a direct outcome of warm winter combined with less solar feed-in in that year.

CURRICULUM VITÆ



Abdolrahman Khoshrou (Majid) was born and raised in the north part of Iran. He graduated from Babol Noshirvani University of Technology, Mazandaran, Iran, with a Bachelors's degree in Power Engineering in 2007.

He received a Master's degree in Information Engineering from the Technical University of Porto, Portugal, in 2015. During his master's study, next to his studies, he enjoyed working at Cyber-Physical Control Systems and Robotics lab as a Machine Learning Research Assistant. During that time, he developed a novel online unsupervised learning of Gaussian Mixture Models used for path planning of underwater vehicles.

In November 2015, he started his PhD research at the Centrum Wiskunde & Informatica (The National Dutch Mathematics and Informatics Center) in Amsterdam. This time has fostered in him a desire and passion for life-long learning and growth. Later on during his work at fast-paced start-ups such as Maistering and Sympower, he honed hands-on mentality and drive to get things done. The results of his PhD are presented in this dissertation.

LIST OF PUBLICATIONS

Journals

1. Abdolrahman Khoshrou, Eric J Pauwels. **Regularisation for PCA-and SVD-type matrix factorisations**. preprint 2021. Springer - Advances in Computational Intelligence[1].

2. Abdolrahman Khoshrou, and E.J. Pauwels. **Short-term scenario-based probabilistic load forecasting: A data-driven approach**. 2019. Elsevier - Applied Energy.

3. Abdolrahman Khoshrou, André Dorsman, Eric J. Pauwels. **The evolution of electricity price on the German day-ahead market before and after the energy switch**. 2019. Elsevier - Renewable Energy.

Conferences

1. Abdolrahman Khoshrou, Eric J. Pauwels. **Data-driven pattern identification and outlier detection in time series**. 2018. Springer, Cham - Science and Information Conference.

2. Abdolrahman Khoshrou, Eric J Pauwels. **Quantifying volatility reduction in German day-ahead spot market in the period 2006 through 2016**. 2018. IEEE - Power & Energy Society General Meeting (PESGM).

3. Abdolrahman Khoshrou, André Dorsman, Eric J. Pauwels. **SVD-based Visualisation and Approximation for Time Series Data in Smart Energy Systems**. 2017. IEEE - Innovative Smart Grid Technologies Conference Europe (ISGT-Europe).

4. Abdolrahman Khoshrou, Eric J Pauwels. **Propagating uncertainty in tree-based load forecasts**. 2017. IEEE - Electrical and Electronics Engineering (ELECO), 2017 10th International Conference.

5. André Dorsman, Abdolrahman Khoshrou, Eric J. Pauwels. **The influence of the switch from fossil fuels to solar and wind energy on the electricity prices in Germany**. 2016. ISINI conference in Groningen.

[1] Part of this work was published at Belgian-Netherlands Artificial Intelligence Conference (BNAIC) 2021.