UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E MÉTODOS QUANTITATIVOS

GUSTAVO CARVALHO DE MELO VIRGOLINO

WIND TURBINE POWER CURVE MODELING WITH GAUSSIAN PROCESSES

FORTALEZA
2020

GUSTAVO CARVALHO DE MELO VIRGOLINO

WIND TURBINE POWER CURVE MODELING WITH GAUSSIAN PROCESSES

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem e Métodos Quantitativos do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Modelagem e Métodos Quantitativos.

Orientador: Prof. Dr. Guilherme de Alencar Barreto.
Coorientador: Prof. Dr. César Lincoln Mattos.

FORTALEZA
2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

V81w Virgolino, Gustavo Carvalho de Melo.
Wind turbine power curve modeling with Gaussian processes / Gustavo Carvalho de Melo Virgolino. – 2020.
87 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Modelagem e Métodos Quantitativos, Fortaleza, 2020.
Orientação: Prof. Dr. Guilherme de Alencar Barreto.
Coorientação: Prof. Dr. César Lincoln Mattos.

1. Wind energy. 2. Wind turbine power curve. 3. Gaussian process. 4. Heteroscedastic models. I. Título.
CDD 510

GUSTAVO CARVALHO DE MELO VIRGOLINO

WIND TURBINE POWER CURVE MODELING WITH GAUSSIAN PROCESSES

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem e Métodos Quantitativos do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Modelagem e Métodos Quantitativos.

Aprovada em: 08/12/2020

BANCA EXAMINADORA

Prof. Dr. Guilherme de Alencar Barreto (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. César Lincoln Mattos (Coorientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Rafael Izbicki
Universidade Federal de São Carlos (UFSCar)

Prof. Dr. José Ailton Alencar de Andrade
Universidade Federal do Ceará (UFC)

Ad maiorem Dei gloriam.

For my beloved wife, Larissa.

ACKNOWLEDGEMENTS

I profoundly thank God almighty for blessing me with the intellectual gifts necessary to persevere on my studies and to work on this dissertation. I also thank the Roman Catholic Apostolic Church and its priests for providing me with invaluable spiritual support to pursue my goals.

I dearly thank my wife, Larissa Matias Rocha Carvalho, for her love and unconditional support, always understanding all the long hours spent on this project and pushing me to be the best I can be. I also thank my mother, Fernanda Carvalho de Melo, for inspiring me to work hard to reach my goals.

I thank my advisor, Prof. Dr Guilherme de Alencar Barreto, for accepting me as his student and guiding me on the development of this dissertation. I also thank him for introducing me to my co-advisor, Prof. D. César Lincoln Mattos, who skillfully introduced me to Gaussian processes with the proper balance between teaching me the subject and letting me explore it on my own. I thank both of them for understanding the difficulties I faced about splitting my time between research and work and helping me through this process.

I thank Delfos Intelligent Maintenance, the company to whom I work, for giving me the opportunity to pursue my masters degree in parallel with my work. Working at Delfos IM was critical to the development of my technical expertise on wind turbine power curve modeling and much of this knowledge is reflected in this dissertation.

I thank my friend José Augusto Fontenelle Magalhães, a former co-worker and co-author of a published journal paper related to this dissertation. Sharing my ideas and implementations with another Gaussian process researcher was very important to my studies about the topic.

I thank the GPflow community for the great help I received from them, which goes far beyond the great Gaussian processes modeling software provided by them. The discussions I had with its members were critical to the implementation of all the proposed models in this dissertation.

“The greatness of the human being consists in this: that it is capable of the universe.”

(Saint Thomas Acquinas)

RESUMO

Nesta dissertação, o problema de modelagem da curva de potência de turbinas eólicas é revisitado com o objetivo de propor e avaliar uma nova estrutura de modelagem semiparamétrica, probabilística e baseada em dados. Para este propósito, processos gaussianos e suas extensões heterocedásticas e robustas são combinados com funções logísticas, resultando em modelos que se assemelham à forma sigmoidal esperada para curvas de potência de turbinas eólicas, permitem previsões probabilísticas, modelam adequadamente o comportamento heterocedástico do fenômeno e são robustos a *outliers*. A metodologia de modelagem proposta é comparada a múltiplas técnicas de modelagem encontradas na literatura técnica e científica de curvas de potência de turbinas eólicas, a saber, o método de bins, regressão polinomial, redes neurais, funções logísticas e regressão via processo gaussiano. Usando um rico conjunto de dados de 1 ano de operação de uma turbina eólica, todos os modelos são comparados em múltiplos cenários relativos às principais características do problema de modelagem de curvas de potência de turbinas eólicas. Os resultados mostram que a metodologia de modelagem proposta apresenta resultados competitivos em métricas determinísticas quando comparada aos demais modelos avaliados, enquanto também exibe as propriedades probabilísticas desejadas, o que lhe confere a capacidade de representar adequadamente as incertezas intrínsecas ao problema de modelagem de curvas de potência de turbinas eólicas.

Palavras-chave: Energia eólica. Curvas de potência de turbinas eólicas. Processos gaussianos. Modelos heterocedásticos.

ABSTRACT

In this dissertation, the wind turbine power curve (WTPC) modeling problem is revisited with the objective of proposing and evaluating a new semi-parametric, probabilistic and data-driven modeling framework. For this purpose, Gaussian processes and their heteroscedastic and robust extensions are combined with logistic functions, resulting in models which resemble the sigmoidal shape expected for WTPCs, output probabilistic predictions properly modeling the heteroscedastic behavior of the phenomenon and are robust to outliers. The proposed modeling framework is compared to multiple modeling benchmarks found in both the technical and scientific WTPC literature, namely, the method of bins, polynomial regression, neural networks, logistic functions and standard Gaussian process regression. Using a rich dataset of 1-year of operational data of a wind turbine, all models are compared in multiple scenarios concerning the key features of the WTPC modeling problem. The results show that the proposed modeling framework has competitive results regarding deterministic metrics when compared to the evaluated benchmark models, while also exhibiting the desired probabilistic properties, which gives it the ability to properly represent uncertainties intrinsically found in WTPC modeling.

Keywords: Wind energy. Wind turbine power curve. Gaussian processes. Heteroscedastic models.

LIST OF FIGURES

Figure 1 – Internal components of a modern WT. 21
Figure 2 – OEM-WTPC for the model SWT-2.3-108, manufactured by Siemens. 22
Figure 3 – Typical WTPC of a pitch regulated WT. 26
Figure 4 – Dataset **A**, built using the event log to filter abnormal states. 71
Figure 5 – Dataset **B**, built using the data-based filter. 72
Figure 6 – Monthly seasonality of the wind speed distribution for dataset **A**. 73
Figure 7 – Predicted values for all models fitted to dataset **A**. 76
Figure 8 – Predicted values for all models fitted to dataset **B**. 78
Figure 9 – Predictive probability density for all GP models fitted to dataset **B**. 80

LIST OF TABLES

Table 1 – RMSE of dataset **A** for all models fitted to dataset **A**. 77

Table 2 – RMSE of datasets **A** and **B** for all models fitted to dataset **B**. 79

Table 3 – MNLPD of dataset **A** for all GP models fitted on datasets **A** and **B**. 79

Table 4 – Month-based cross-validation statistics for RMSE scores. 81

LISTA DE ABREVIATURAS E SIGLAS

WT	Wind Turbine
WTPC	Wind Turbine Power Curve
OEM	Original Equipment Manufacturer
MoB	Method of Bins
6PLE	6-Parameter Logistic Function
OLS	Ordinary Least Squares
GP	Gaussian Process
PSD	Positive Semi-Definite
SE	Squared Exponential
ARD	Automatic Relevance Determination
ELBO	Evidence Lower Bound
SVGP	Sparse Variational Gaussian Process
RMSE	Root Mean Squared Error
MNLPD	Mean Negative Log Predictive Density
Poly-9	Polynomial model with degree 9
NN	Neural Network
MLP	Multilayer Perceptron
L2P	Logistic model with 2 Parameters
L3P	Logistic model with 3 Parameters
KL divergence	Kullback-Leibler divergence
HG	Heteroscedastic Gaussian
HS	Heteroscedastic Student-T
LRHS	Locally Robust Heteroscedastic Student-T
w.r.t.	with respect to

LIST OF SYMBOLS

P	Power
v	Wind speed
ρ	Air density
ρ_{ref}	Reference air density
θ_{yaw}	Yaw misalignment angle
R	Rotor radius
$C_p(\lambda, \theta_{\text{pitch}})$	Power coefficient
λ	Tip speed ratio
ω	Rotational speed of the rotor
θ_{pitch}	Pitch angle
η_{mech}	Mechanical energy conversion efficiency on the drive train
η_{elec}	Electrical energy conversion efficiency on the generation
v_{ci}	Cut-in wind speed
v_{rated}	Rated wind speed
P_{rated}	Rated power
v_{co}	Cut-out wind speed
v_{raw}	Raw wind speed, before air density normalization
p	Normalized power
T	Ambient temperature
B	Ambient pressure
ϕ	Relative humidity
R_0	Gas constant of dry air
R_w	Gas constant of water vapor
B_w	Water vapor pressure
b_k	k-th Bin of the MoB
Δv	Width of the bins

$\bar{v}_k$	Mean wind speed of bin b_k
$\bar{P}_k$	Mean power of bin b_k
N_k	Number of observations in bin b_k
α	Lower asymptote parameter of the 6PLE model
δ	Upper asymptote parameter of the 6PLE model
β	Growth rate parameter of the 6PLE model
v_0	Location shift parameter of the 6PLE model
γ	Asymmetry control parameter of the 6PLE model
ε	Sixth paramter (no clear interpretation) of the 6PLE model
s	Scale (or inverse growth-rate) parameter of the 6PLE model
a	Intercept of the linearized L2P model
b	Coeficient of the linearized L2P model
f, g, h	Unknown functions, whose uncertainty is usually modeled as a GP
$\mathscr{X}$	Generic domain of a function
$\boldsymbol{\theta}_{\text{model}}$	Parameter vector of a model
$\boldsymbol{f}$	Function evaluation vector
$\mathscr{N}(\boldsymbol{m}, bmK)$	Multivariate Gaussian distribution
$\boldsymbol{m}$	Mean vector of a multivariate Gaussian distribution
$\boldsymbol{K}$	Covariance matrix of a multivariate Gaussian distribution
$p(x)$	Probability density of the random variable x
$\mathscr{N}(\boldsymbol{f} \mid \boldsymbol{m}, \boldsymbol{K})$	Probability density of the random vector $\boldsymbol{f}$ which follows a multivariate Gaussian distribution
$\lvert\boldsymbol{A}\rvert$	Determinant of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	Inverse of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{\top}$	Transpose of matrix $\boldsymbol{A}$
$A \mid C_1, C_2, \ldots$	Random variable representing conditioning of A given $C_1, C_2, \ldots$
D	Dimension of the inputs
$\mathscr{GP}(\mu, \kappa)$	Gaussian process

μ	Mean function
κ	Covariance function
$\boldsymbol{x}$	Input vector
$\boldsymbol{X}$	Input matrix
$\boldsymbol{f}\|\boldsymbol{X}$	Random vector of the function evalutions $\boldsymbol{f}$ given their corresponding inputs $\boldsymbol{X}$
$\boldsymbol{m}_{\boldsymbol{X}}$	Mean vector obtained by evaluating the mean function μ at all elements of the input vector $\boldsymbol{X}$
$\boldsymbol{K}_{\boldsymbol{X}_1\boldsymbol{X}_2}$	Covariance matrix obtained by evaluating the covariance function κ at all element pairs from the input vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$
$\boldsymbol{\phi}$	Parameter vector of the mean function μ
$\mathbb{E}_{p(x)}\left[F(x)\right]$	Expectation of some expression $F(x)$ depending on some random variable x which has probability density $p(x)$
$\mathrm{Cov}(x_1,x_2)$	Covariance between two random variables x_1 and x_2
κ_{SE}	SE covariance function
$\boldsymbol{\theta}$	Hyperparameter vector of the covariance function
σ_f^2	Variance hyperparameter of the SE covariance function
l_i	i-th input dimension length scale hyperparameter of the SE covariance function
y	Scalar output
$\boldsymbol{x} \rightarrow y$	Input-output relationship between input vector $\boldsymbol{x}$ and scalar output y
N	Number of observations
$\mathscr{P}(y\|\boldsymbol{\gamma})$	Probability density of output y, which follows a generic probability distribution with hyperparameter vector $\boldsymbol{\gamma}$
$\boldsymbol{\gamma}$	Hyperparameter vector of a probability distribution
$\mathscr{N}(y\|\mu_y,\sigma_y^2)$	Probability density of the output y, which follows a univariate Gaussian distribution with mean μ_y and variance σ_y^2.
μ_y	Mean (or location) hyperparameter of univariate probability distributions
σ_y^2	Variance hyperparameter of univariate probability distributions

σ_y	Standard deviation (or scale) hyperparameter of univariate probability distributions
$\mathscr{T}(y\|\mu_y,\sigma_y,\nu)$	Probability density of the output y, which follows a univariate Student-t distribution with location μ_y, scale σ_y^2 and degrees-of-freedom ν
$\boldsymbol{I}_N$	$N \times N$ Identity matrix
$\mathrm{tr}\boldsymbol{A}$	Trace of matrix $\boldsymbol{A}$
$\boldsymbol{x}'$	New input vector
N	Number of new observations
y'	New scalar output
$\boldsymbol{X}'$	New input matrix
$\boldsymbol{f}'$	New function evaluation random vector
$\boldsymbol{y}'$	New output vector
$\mathbb{D}_{\mathrm{KL}}[q(x)\|\|p(x)]$	KL divergence from the probability distribution $q(x)$ to the probability distribution $p(x)$
$q(x)$	The variational probability distribution used to approximate (in a KL divergence minimizing sense) some posterior probability distribution
$\boldsymbol{\lambda}$	Variational parameter vector
$\boldsymbol{z}$	Pseudo-input vector
$\boldsymbol{Z}$	Pseudo-input matrix
u	Scalar inducing variable
$\boldsymbol{u}$	Inducing variable vector
$\boldsymbol{m_u}$	Variational mean vector
$\boldsymbol{L_u}$	Cholesky factor of the variational covariance matrix
$\boldsymbol{m_f}$	Mean vector of the variational posterior $q(\boldsymbol{f})$
$\boldsymbol{K_f}$	Covariance matrix of the variational posterior $q(\boldsymbol{f})$
m_{f_i}	i-th element of the mean vector $\boldsymbol{m_f}$
$k_{f_{ii}}$	i-th element of the main diagonal of the covariance matrix $\boldsymbol{K_f}$
$\boldsymbol{m_{y'}}$	Mean vector of the variational posterior predictive $p(\boldsymbol{y}'\|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y})$
$\boldsymbol{K_{y'}}$	Covariance matrix of the variational posterior predictive $p(\boldsymbol{y}'\|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y})$

L	Number of latent GPs in a Chained GP model
F	Unknown multioutput function whose components f^j are modeled as GPs
$\boldsymbol{F}$	Multi-output function evaluation matrix
t, t'	Warping functions used to map their inputs to constraint-satisfying images
$\tilde{\boldsymbol{Z}}$	Pseudo-input set
$\boldsymbol{U}$	Inducing variable matrix
$\mathbb{E}[\boldsymbol{y}']$	Mean vector of the new predictions $\boldsymbol{y}'$
$\mathbb{V}[\boldsymbol{y}']$	Covariance matrix of the predictions $\boldsymbol{y}'$
$\boldsymbol{v}$	Whitened inducing variable vector
$\boldsymbol{m_v}$	Whitened variational mean vector
$\boldsymbol{L_v}$	Cholesky factor of the whitened variational covariance matrix
$\mathrm{chol}(\boldsymbol{A})$	Cholesky factor of matrix $\boldsymbol{A}$
$\mathbf{0}_{M\times N}$	$M \times N$ Matrix of zeros
α_i	Coefficients of the polynomial model
$\mathbf{W}(\lambda, k)$	Weibull distribution
k	Shape parameter of a Weibull distribution
λ	Scale parameter of a Weibull distribution
$\hat{y}$	Model predicted output
$\mu_{\hat{y}}$	Mean of the model prediction
$\sigma_{\hat{y}}$	Standard deviation of the model prediction

CONTENTS

1 INTRODUCTION . . . 20

1.1 The Wind Turbine Power Curve . . . 21

1.2 Motivation and Objectives . . . 22

1.3 Scientific Production . . . 23

1.4 Organization of the Dissertation . . . 24

2 WIND TURBINE POWER CURVE MODELING . . . 25

2.1 A typical WTPC . . . 25

2.1.1 Design Parameters . . . 26

2.1.2 Operating Ranges . . . 26

2.1.2.1 Region 1: No Generation, $v < v_{ci}$ . . . 27

2.1.2.2 Region 2: Maximum Power Point Tracking, $v_{ci} < v < v_{rated}$ . . . 27

2.1.2.3 Region 3: Rated Power Control, $v_{rated} < v < v_{co}$ . . . 27

2.1.2.4 Region 4: Safety Shutdown, $v > v_{co}$ . . . 27

2.2 The IEC 61400-12-2 Standard . . . 28

2.2.1 Data Sources . . . 28

2.2.2 Air Density Normalization . . . 29

2.2.3 Method of Bins . . . 29

2.3 Related Work . . . 30

2.4 The Logistic Function Model . . . 32

2.4.1 Asymptotic Behavior and Operational Ranges . . . 32

2.4.1.1 No generation, $v < v_{ci}$ . . . 32

2.4.1.2 Rated Power, $v_{rated} < v < v_{co}$ . . . 33

2.4.2 Logistic Models with 2 and 3 Parameters . . . 33

2.4.3 Linearizing Transformation and Parameter Initialization . . . 33

2.5 Discussion . . . 33

3 FUNDAMENTALS OF GAUSSIAN PROCESS MODELS . . . 35

3.1 Multivariate Gaussian Distribution . . . 35

3.2 GP as a Distribution over Functions . . . 36

3.2.1 Mean Function . . . 37

3.2.2 Covariance Function . . . 38

3.2.2.1 Squared Exponential Covariance Function . . . 39

3.3 Regression with Noisy Observations 39

3.3.1 Regression Model 39

3.3.2 Likelihoods 40

3.3.2.1 Gaussian Likelihood *40*

3.3.2.2 Student-t Likelihood *41*

3.3.3 Inference 42

3.3.3.1 Computational Complexity and Memory Requirements *42*

3.3.4 Prediction 43

3.4 Discussion 44

4 VARIATIONAL INFERENCE APPLIED TO GAUSSIAN PROCESS MODELS 46

4.1 Variational Inference on GP models 46

4.1.1 Kullback-Leiber Divergence 47

4.1.2 Evidence Lower Bound 47

4.1.3 Variational Distribution 48

4.1.4 Optimization Objective 48

4.1.5 Variational Posterior Approximation 49

4.2 Sparse Variational Gaussian Process Models 49

4.2.1 Augmented Regression Model 49

4.2.2 Variational Inference 50

4.2.2.1 Evidence Lower Bound *50*

4.2.2.2 Variational Distribution *51*

4.2.2.3 Computational Complexity and Memory Requirements *52*

4.2.3 Prediction 53

4.3 Chained Gaussian Process Models 55

4.3.1 Multiple Latent GP Regression Model 55

4.3.2 Likelihoods depending on Multiple Latent GPs 56

4.3.2.1 Heteroscedastic Gaussian Likelihood *56*

4.3.2.2 Heteroscedastic Student-t Likelihood *57*

4.3.2.3 Locally Robust Heteroscedastic Student-t Likelihood *57*

4.3.3 Variational Inference 58

4.3.3.1 Evidence Lower Bound *59*

4.3.3.2 *Variational Distribution* *59*

4.3.4 ***Prediction*** 61

4.4 **Discussion** 63

5 **PROPOSED MODELING FRAMEWORK** 64

5.1 **Modeling Framework Description** 64

5.1.1 ***Model Construction Rationale*** 64

5.1.2 ***Note on Model Implementation*** 66

5.1.3 ***Parameters Initialization*** 67

5.1.4 ***Optimization*** 67

5.2 **Benchmark Models** 67

5.2.1 ***Method of Bins (MoB)*** 68

5.2.2 ***Polynomial Regression (Poly-9)*** 68

5.2.3 ***Neural Networks - MLP(1,12,1)*** 68

5.2.4 ***Logistic Function (L3P)*** 69

5.2.5 ***Zero Mean GP (0-GP)*** 69

5.3 **Dataset Description** 70

5.3.1 ***Wind Turbine Operational Parameters*** 70

5.3.2 ***Data Sources*** 70

5.3.3 ***Data Cleaning*** 71

5.3.4 ***Data Seasonality*** 72

5.4 **Experiments and Results** 74

5.4.1 ***Comparison Criteria*** 74

5.4.1.1 *Root Mean-Squared Error (RMSE)* *74*

5.4.1.2 *Mean Negative Log Predictive Density (MNLPD)* *75*

5.4.1.3 *Fitting and Evaluating* *75*

5.4.2 ***Data-Fitting Capabilities*** 75

5.4.3 ***Robustness to Outliers*** 77

5.4.4 ***Heteroscedasticity*** 79

5.4.5 ***Seasonal Variations*** 81

5.5 **Discussion** 82

6 **CONCLUSION** 83

6.1 **Further Work** 83

6.1.1 ***Evaluation of Other Likelihoods*** 84

6.1.2 ***More Inputs*** 84

6.1.3 ***Alternative Quadrature Methods*** 84

REFERENCES 85

1 INTRODUCTION

Wind energy plays a significant role in the world's quest for more sustainable energy sources. Spread out across the globe, wind turbines (WTs) harvest the kinetic energy from the wind to generate electric power. The energy conversion process is the result of the interaction of multiple external and internal phenomena, which are considered below.

The process starts at the WT's rotor, which rotate by virtue of the aerodynamic forces developed by the contact of the wind with its blades. The dynamic interaction depends on the air density, which is a function of the ambient temperature, pressure and humidity. Furthermore, having rotors with radii greater than 50 m, modern WTs are subjected to all kinds of variations of the wind flow, characterized not only by its speed but also by its direction, turbulence, shear and veer. The complexity is even more aggravated by the presence of multiple WTs, which may cast an aerodynamic shadow over each other, and terrain topographic features such as inclination and roughness. Finally, the rotor's blades can also deviate from their ideal aerodynamic shape due to aging and natural effects such as dirt accumulation and icing.

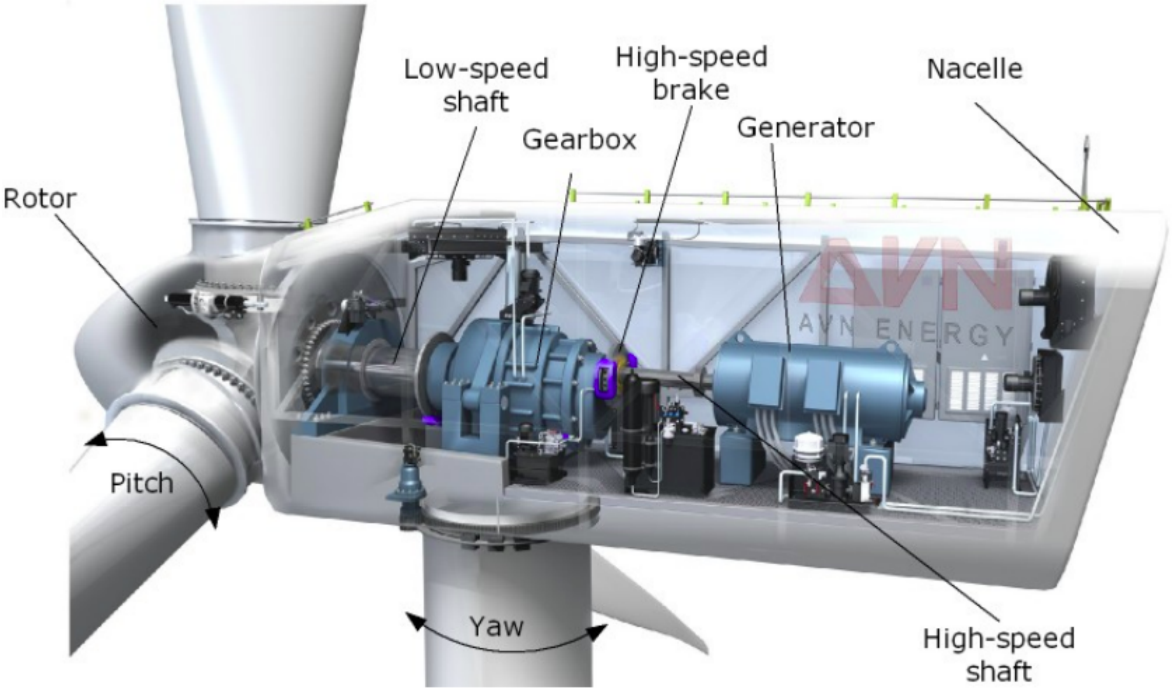
Inside the WT, as shown in fig. 1, the drive train, i.e., the shafts and gears connecting the rotor and the generator, transfers the mechanical torque from the former to the latter in a process involving hard-to-quantify and aging-dependent frictional energy losses, more evidently perceived in WT models using a gearbox. There are also energy losses in the generator as it cannot always operate on its optimal state due to the ever-changing mechanical torque provided by the wind speed despite the feedback control strategy employed by the WT control system.

Summarizing the phenomena discussed above, the energy conversion process can, in a simplified view, be described by the following equation:

$$P = \frac{1}{2}\eta_{\text{elec}}\eta_{\text{mech}}C_p(\lambda, \theta_{\text{pitch}})\pi R^2 v^3 \cos^3\theta_{\text{yaw}} \tag{1}$$

where P is the produced power in W, ρ is the air density kg/m^3, v is the horizontal wind speed in m/s, θ_{yaw} is the yaw misalignment angle, such that $v\cos\theta_{\text{yaw}}$ is the wind speed perpendicular to the rotor, R is the rotor radius in m, $C_p(\lambda, \theta_{\text{pitch}})$ is the power coefficient, which accounts for the efficiency of aerodynamic energy conversion in the blades and is a function of the tip speed ratio $\lambda = \omega R/v$, where ω is the rotational speed of the rotor in rad/s, and the pitch angle θ_{pitch}, η_{elec} is electrical energy conversion efficiency of the generator and η_{mech} is the mechanical energy conversion efficiency of the drive train. Accounting for all those factors would require very detailed models capable of interconnecting all of them, and collecting all the information

Figure 1 – Internal components of a modern WT.



Source - Adapted from Jepsen *et al.* (2010).

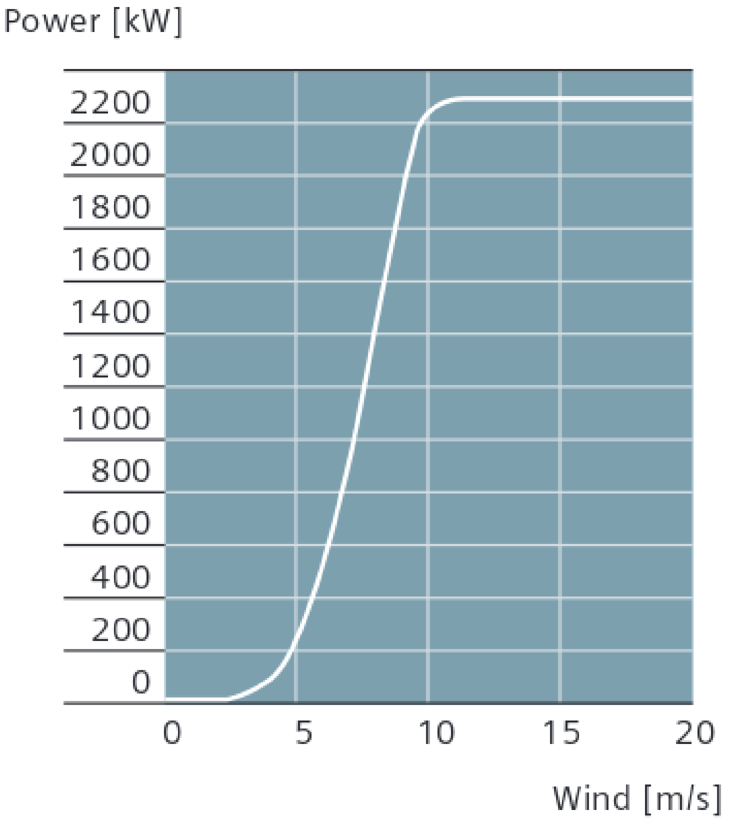
necessary to run them would be infeasible, especially for wind farms with hundreds of WTs. Usually, physics-based models of the WT energy generation process consist in the description of isolated components such as the aerodynamics of the rotor and the electromagnetism and control of the generator.

1.1 The Wind Turbine Power Curve

The wind turbine power curve (WTPC) abstracts those complex interactions as relationship connecting the wind speed v received by a WT's rotor to the power P it produces. It is of central importance for the wind energy industry as it can be used for numerous tasks such as power assessment and forecasting, capacity factor estimation, WT model selection, and performance monitoring (SOHONI *et al.*, 2016).

Due to its fundamental role in the aforementioned engineering calculations, the original equipment manufacturer (OEM) of a WT model is required to provide its WTPC as part of its technical documentation, as exemplified in fig. 2. Those models, the OEM-WTPCs, are often obtained from either idealized physical simulations or testing setups. It is usually given as set of wind speed and power pairs (v_i, P_i), that can be used to build a deterministic regression model in the form $P = f(v)$, resulting in a sigmoidal shaped function.

Figure 2 – OEM-WTPC for the model SWT-2.3-108, manufactured by Siemens.



Source - SWT-2.3-108 Datasheet.

In practice, however, the ideal conditions are seldom met, as actual WTs operating in the field are subjected to varying degrees of all the operating conditions presented above. This results in real observations of wind speed and power data (v_i, P_i) which keep the sigmoidal shape, but deviate from the deterministic model and exhibit noisy behavior around it, generating data sets with heteroscedasticity and outliers. This challenging modeling task is the main focus of this dissertation.

1.2 Motivation and Objectives

The main objective of this dissertation consists in proposing and evaluating a new semi-parametric, probabilistic and data-driven WTPC modeling framework. To accomplish this goal, the parametric and deterministic logistic function models (VILLANUEVA; FEIJÓO, 2018) are combined with non-parametric and probabilistic Gaussian processes (GPs) models (RASMUSSEN; WILLIAMS, 2006) and their heteroscedastic and robust extensions (LÁZARO-GREDILLA; TITSIAS, 2011; SAUL *et al.*, 2016).

The resulting models are expected to incorporate prior knowledge about the WTPC by means of the parametric part while being able to adapt to the data with the non-parametric one. More specifically, the proposed models should present the following characteristics.

1. Resemble the sigmoidal shape expected for the WTPC;
2. Output probabilistic predictions of the electric power P given a wind speed v, accounting for uncertainties both in the model and in the data;
3. Capable of modeling heteroscedastic behavior of the phenomenon;
4. Be robust to outliers.

To check if the objectives are accomplished, the proposed models are evaluated regarding multiple wind speed and power data sets (v_i, P_i) which are chosen to represent typical applications of WTPCs. They are also compared to the state of the art in WTPC modeling, such as polynomial regression, standard logistic functions models, neural networks and standard GP regression.

As a side objective, this dissertation also aims to discuss the fundamentals of GP models theory and probabilistic modeling in such a way to make it more understandable by readers which are not familiar with the topic by focusing on the intuition behind the involved equations. With this, it is expected that the reader will at least grasp the ideas behind applying those tools to a real engineering problem and perhaps become capable of transferring the knowledge to other problems. The author firmly believes that those techniques can successfully be applied to many technical problems, especially with the ever-increasing amount of data becoming available in multiple industries.

1.3 Scientific Production

The following journal paper is the result of the studies developed along the course of this research.

VIRGOLINO, G. C. de M.; MATTOS, C. L. C.; MAGALHÃES, J. A. F.; BARRETO, G. A. Gaussian processes with logistic mean function for modeling wind turbine power curves. **Renewable Energy**, [United Kingdom, v. 162, p. 458–465, 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0960148120309150. Acesso em: 05 dez. 2020.

1.4 Organization of the Dissertation

The remainder of this document is organized as follows:

In chapter 2, the WTPC modeling problem is further discussed, starting with a presentation of the main design parameters and operating ranges of a typical WTPC. The technical approach for WTPC Modeling, that is, the IEC 61400-12-2 International Standard (INTERNATIONAL ELECTROTECHNICAL COMMISSION, 2017) is covered, as well as a revision of related works in the renewable energy literature. It is finished with a more detailed presentation of the logistic function model as a preparation for the new framework to be proposed.

In chapter 3, the fundamentals of GP modeling are introduced in a general setting. After a review of the multivariate Gaussian distribution and definition the GP as an extension of it, the application of GPs to standard regression is considered, highlighting its merits and difficulties.

In chapter 4, variational inference is discussed as a tool to overcome the challenges of the standard GP regression. After a review of the variational inference procedure for general GP models, the sparse variational GP (SVGP) approximation is explored as a way to make the modeling more flexible and also require less computational and memory resources. It is finished with the presentation of the Chained GP, which extends the SVGP into a multiple latent GP regression model, capable of representing more features of the noisy distribution of the observations such as heteroscedasticity and localized heavy-tailedness.

In chapter 5, the proposed modeling framework of this dissertation is presented, convering both the theoretical rationale used to construct it and the implementation, initialization and optimization details of it. A set of benchmark models is selected fromfrom the literature, and their implementation details are given. After the detailed presentation of a 1-year of WT operation dataset and its distinct features, multiple experimental results are reported, comparing the proposed models with the benchmark ones.

In chapter 6, this dissertation is concluded with final remarks about the developed work and the achieved contributions to the WTPC modeling problem. It also gives a brief summary of possible future research topics.

2 WIND TURBINE POWER CURVE MODELING

The WTPC modeling problem can be stated as the construction of a mathematical model which describes the electric power P produced by a WT in terms of the wind speed v received by its rotor. Although it appears to be a simple regression problem with a single input and a single output, the wind speed and power samples (v_i, P_i) have its own peculiarities such as the usual sigmoidal shape, heteroscedastic behavior, i.e, the observed noise intensity on the power P depends on the wind speed v, and presence of outliers, which makes the WTPC modeling a challenging task.

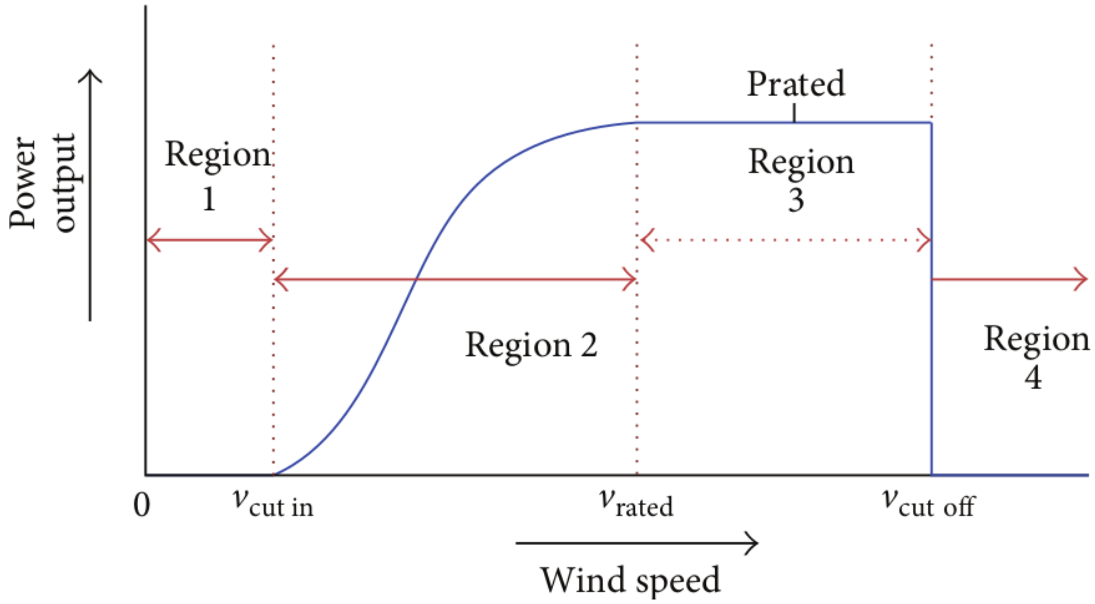
In this chapter, the peculiarities of the WTPC modeling problem are discussed. In section 2.1, the common characteristics of all pitch regulated WTs[1] are described in terms of their main design parameters and the associated operating ranges, emphasizing their impact on the wind speed and power data (v_i, P_i). In section 2.2, the IEC 61400-12-2 Standard (INTERNATIONAL ELECTROTECHNICAL COMMISSION, 2013) is visited, analyzing the technical approach to the WTPC modeling task. In section 2.3, the vast contributions to the topic present in the renewable energy literature are reviewed, drawing inspirations for the new modeling framework proposed in this dissertation and establishing comparison benchmarks based on the state of the art. In section 2.4, the logistic function model presented in Villanueva and Feijóo (2018) is revisited and correlated with the operating ranges of a WT. The chapter is concluded in section 2.5.

2.1 A typical WTPC

Before discussing the multiple approaches to WTPC modeling, it is important to analyze how a WTPC works and understand how this relates to the peculiarities of the wind speed and power data (v_i, P_i). Even though every WT has its specificities related to its model and operating conditions, the typical WTPC of pitch regulated WTs can be generically represented by a sigmoidal shaped function as in fig. 3 below.

[1] Pitch regulated WTs are the most widespread WT architechture in operation today, so this dissertation deliberately chooses to restrict the focus to them.

Figure 3 – Typical WTPC of a pitch regulated WT with its design parameters and operating ranges.



Source - Adapted from Sohoni *et al.* (2016).

2.1.1 Design Parameters

The WTPC in fig. 3 is characterized by four design parameters, namely, the cut-in wind speed v_{ci}, which is the minimum wind speed necessary to enable the WT to generate power, the rated wind speed v_{rated}, which is the minimum wind speed necessary for the WT to reach its rated power P_{rated}, i.e., its maximum generation, and the cut-off wind speed v_{co}, which is the maximum wind speed the WT can withstand before shutting down due to safety reasons.

The rated power P_{rated} is frequently used to define the normalized power p as

$$p = \frac{P}{P_{\text{rated}}}, \tag{2}$$

where P is the generated power for a given wind speed v. This definition allows the comparison of two WTs with different specificatons by considering the WTPC in terms of the normalized power p.

2.1.2 Operating Ranges

There are four well-defined regions in fig. 3, called the WTPC operating ranges. Each operating range has its own peculiarities which must be accurately represented by a WTPC model. They are described below, with emphasis on how their behaviors impact on the WTPC modeling problem.

2.1.2.1 *Region 1: No Generation, $v < v_{ci}$*

Wind speeds below the cut-in wind speed, $v < v_{\mathrm{ci}}$, are not enough to produce sufficient mechanical torque to start the generator. The WT will not generate any power, i.e., $P = 0$. The WT control system actively seeks to align the WT with the wind direction and regulate the pitch angle of the blades to favor the start-up process.

2.1.2.2 *Region 2: Maximum Power Point Tracking, $v_{ci} < v < v_{rated}$*

Wind speeds in the maximum power point tracking (MPPT) operating range are above the cut-in wind speed $v > v_{\mathrm{ci}}$ and are sufficiently strong to produce a mechanical torque capable of starting the power generation, hence $P > 0$. However, they are not enough to reach the maximum possible power generation $P = P_{\mathrm{rated}}$, as they are below the rated wind speed, $v < v_{\mathrm{rated}}$. This operating range and the transition between it and its neighbors is the main responsible for the typical sigmoidal shape presented by WTPCs.

During MPPT, the WT control system will do its best to maximize the power production by seeking optimal yaw alignment with the wind direction and optimal pitch angle of the rotor's blades to match the wind speed, which is a hard task due to the stochastic nature of the wind flow. The complex behavior of this operating range results in the deviation and scattering of the speed and power observations (v_i, P_i) around the idealized deterministic WTPC function $P = f(v)$, thus being of special interest for the modeling task.

2.1.2.3 *Region 3: Rated Power Control, $v_{rated} < v < v_{co}$*

Wind speeds above the rated wind speed, $v > v_{\mathrm{rated}}$, are strong enough to supply sufficient mechanical torque to reach a rated power generation: $P = P_{\mathrm{rated}}$. In fact, as the wind speed v grows, the mechanical torque provided by it becomes higher than needed, so the WT control system works to maintain the maximum generation while not overloading the WT by changing the pitch angle of the rotor's blades. This controlled behavior is maintained until the cut out wind speed is reached, $v < v_{\mathrm{co}}$, which marks the transition to the next operating range.

2.1.2.4 *Region 4: Safety Shutdown, $v > v_{co}$*

Wind speeds above the cut-out wind speed, $v > v_{\mathrm{co}}$ are deemed unsafe for the WT operation due to the high aerodynamic forces and mechanical torque they produce. The WT control system shuts down the power generation, hence $P = 0$, and take precautionary measures to minimize the aerodynamic forces applied to the rotor's blades. This operating range is seldom reached, thus including it in WTPC modeling would provide little gains and add unnecessary complexity, and is usually neglected by WTPC models.

2.2 The IEC 61400-12-2 Standard

Recognizing the importance of WTPC modeling, the International Electrotechnical Commision (IEC) proposed standardized methods which are widely accepted in the wind energy industry. The IEC-61400-12-2 (INTERNATIONAL ELECTROTECHNICAL COMMISSION, 2013) is an international standard (thereafter referred simply as IEC standard) describing the data acquisition and model fitting procedures to obtain the WTPCs of individual WTs operating in the field.

While this dissertation does not aspire to follow the full scope of the IEC standard, some parts of it are used as either foundations or comparison benchmarks for the WTPC modeling framework it aims to develop. They are reviewed below with appropriate remarks regarding their application in this dissertation.

2.2.1 Data Sources

The wind speed and electric power observations (v_i, P_i) is the main data needed for WTPC modeling, and the IEC standard requires them to be registered as 10-minute averages of continuous measurements. Almost all WT manufactures follow this data recording format, making it readily available in the WT's supervisory control and data acquisition (SCADA) system. As such, it is the main data source used by this dissertation.

The IEC standard also requires 10-minute averaged data for ambient temperature, pressure and relative humidity (T_i, B_i, ϕ_i) to account for varying air density effects. Although this data (or the air density data) is not guaranteed to be available in the SCADA, it is usual for many wind farms to have it recorded by one or more meteorological masts installed in the site. This dissertation makes use this data whenever it is available to apply the air density normalization methodology proposed by the IEC Standard, which is discussed below.

2.2.2 Air Density Normalization

IEC standard provides a simple yet effective way to account for the effects of varying air density in pitch regulated WTs. Given the measurements of the raw wind speed v_{raw} and air density ρ, the following wind speed normalization is applied:

$$v = v_{\text{raw}} \left(\frac{\rho}{\rho_{\text{ref}}} \right)^{1/3}, \tag{3}$$

which gives the normalized wind speed v as if the air density was fixed to a reference value ρ_{ref}. Hence, the air density effects can be incorporated into the WTPC by modeling it using the normalized wind speed v. The air density must ρ be either provided directly or computed with the following equation:

$$\rho = \frac{1}{T} \left[\frac{B}{R_0} - \phi B_w \left(\frac{1}{R_0} - \frac{1}{R_w} \right) \right], \tag{4}$$

where

- ρ is the air density, in kg/m^3;
- T is the ambient temperature, in K;
- B is the ambient pressure, in Pa;
- ϕ is the relative humidity, ranging from 0 to 1, adimensional;
- $R_0 = 287.05$ J/(kg·K) is the gas constant of dry air;
- $R_w = 461.50$ J/(kg·K) is the gas constant of water vapor;
- B_w is the water vapor pressure, in Pa, given by

  $$B_w = a \exp(bT), \tag{5}$$

  with constants $a = 2.05 \times 10^{-5}$ Pa and $b = 6.31846 \times 10^{-2}$ K^{-1}.

2.2.3 Method of Bins

The method of bins (MoB) is proposed by the IEC Standard to obtain a mathematical model for the WTPC. It is based on the wind speed and power observations (v_i, P_i), with the air density normalization given by eq. (3) already applied for some reference air density ρ_{ref}. The method can be implemented with the following steps:

1. Group the observations i in bins b_k with width $\Delta v = 0.5$ m/s with the following rule:

   $$b_k = \left\{ i : |v_i - k\Delta v| < \frac{1}{2}\Delta v \right\}, \quad k = 0, 1, \ldots, 50 \tag{6}$$

2. For each bin b_k, compute the mean of the wind speed v and power P:

$$\bar{v}_k = \frac{1}{N_k} \sum_{i \in b_k} v_i, \tag{7}$$

$$\bar{P}_k = \frac{1}{N_k} \sum_{i \in b_k} P_i, \tag{8}$$

where $N_k = |b_k|$ is the number of observations in the bin b_k;

3. Compose the WTPC $P = f(v)$ by interpolating the mean wind speed and mean power $(\bar{v}_k, \bar{P}_k)$ of each bin.

The main merit of the MoB is providing a simple way to build a WTPC model. However, it has some limitations related to the discretizing behavior induced by the data binning. The MoB is considered a comparison benchmark for the modeling framework proposed in this dissertation due to its technical importance for the WTPC modeling task.

2.3 Related Work

The WTPC modeling is a very active topic of research in the renewable energy literature. Many authors (SOHONI *et al.*, 2016; CARRILLO *et al.*, 2013; LYDIA *et al.*, 2014; EMINOGLU; TURKSOY, 2019; WANG *et al.*, 2019) provide detailed reviews of the subject, reflecting the richness of modeling paradigms, with techniques ranging from the well-established polynomial regression to the application of modern machine learning algorithms. More specifically, Sohoni *et al.* (2016) classifies the WTPC models as follows.

Discrete: models that inspired by method of bins discussed in section 2.2.3 and discretize the wind speed measurements v_i into intervals and analyze the mean of the power measurments P_i on them to characterize the WTPC by interpolation.

Deterministic vs. Probabilistic: A deterministic model considers that given a wind speed v, the power P is uniquely defined, whereas a probabilistic one accounts for uncertainties in the power P even for a exactly known wind speed v;

Parametric vs. Nonparametric: A parametric model has a defined functional form to model the WTPC, whereas a nonparametric one does not. According to this definition, Sohoni *et al.* (2016) classify neural networks (NNs) as nonparametric models. Parametric models offer interpretability in exchange of adaptability, while their nonparametric counterparts do the reverse.

Presumed Shape vs. Curve Fitting vs. Actual Data: Presumed shape models only use the design parameters to establish the WTPC, while curve fiting models uses WTPC data supplied by the WT's original equipment manufacturer (OEM). Actual data models are built directly from WT operational data.

Stochastic: Models which consider the temporal dependency between wind speed and power observations (v_i, P_i).

Some examples from the recent WTPC modeling literature are now discussed and classified based on the criteria proposed above. The objective is not to entirely cover that vast research topic, but rather to survey it and set similarities, differences and benchmarks for comparison with the modeling framework proposed by this dissertation.

Comparison Benchmarks: Polynomial regression and neural networks (NNs) appear in all the considered reviews (CARRILLO *et al.*, 2013; LYDIA *et al.*, 2014; SOHONI *et al.*, 2016; EMINOGLU; TURKSOY, 2019; WANG *et al.*, 2019), which show how prevalent they are in the WTPC modeling literature, making them appropriate benchmarks for comparison with new models. The polynomial regression can be classified as a deterministic, parametric and actual data model, and is covered in Li *et al.* (2001), Shokrzadeh *et al.* (2014), Guo and Infield (2018), Yan *et al.* (2019). NNs, in their turn, are deterministic, nonparametric and actual data models, and are covered in Li *et al.* (2001), Lydia *et al.* (2013), Manobel *et al.* (2018), Bai *et al.* (2019), Yan *et al.* (2019).

Gaussian Processes (GPs): In Pandit and Infield (2019), GPs are applied to WTPC modeling. The resulting model can be classified as probabilistic, nonparametric and actual data models. The main focus of that paper is analyzing the many possible stationary covariance functions that can be used with a GP to model a WTPC, concluding that the squared-exponential (SE) covariance function is one of the best options available for the task. In Pandit *et al.* (2019), GPs are compared do Support Vector Machine models, which in turn can be classifed as deterministic, nonparametric and actual data. The GP was preferred due to its probabilistic nature, which is also the option made by this dissertation.

Logistic Functions: Logistic functions are strong parametric model options for the WTPC modeling task as they generate sigmoidal shaped functions in agreement with the usual shape of a WTPC, as shown in fig. 3 and discussed in section 2.1.2. Multiple logistic function models are compared in Villanueva and Feijóo (2018) using data provided by the OEM of seven different WT models. The logistic function model is very important for this dissertation development and is discussed in detail in section 2.4.

2.4 The Logistic Function Model

The logistic function model is one of the base components of the WTPC modeling framework this dissertation aims to propose and evaluate. As such, it is now analyzed starting with the best results from Villanueva and Feijóo (2018), which concluded that exponential-based logistic functions constitute the best option for WTPC modeling. The 6-parameter logistic function (6PLE) is the most general of them and is given by

$$P(v) = \delta + \frac{(\alpha - \delta)}{\left[\varepsilon + \exp\left(-\beta\left(v - v_0\right)\right)\right]^{1/\gamma}}, \tag{9}$$

where α is the lower asymptote, δ is the upper asymptote, β is the growth rate, v_0 controls the location shift, γ controls the asymmetry and ε is usually close to 1 and has no clear interpretation. However, the parameter ε in eq. (9) is redundant and can be eliminated with the following reparametrization:

- $\alpha \to \delta + \varepsilon^{1/\gamma}(\alpha - \delta)$,
- $v_0 \to v_0 + \beta^{-1}\log\varepsilon$,
- $\beta = 1/s$,

which gives

$$P(v) = \delta + (\alpha - \delta)\left[1 + \exp\left(-\left(\frac{v - v_0}{s}\right)\right)\right]^{-1/\gamma}. \tag{10}$$

The upcoming analysis will further reduce the number of parameters to three by confronting them with the WTPC operating ranges described in section 2.1.2.

2.4.1 Asymptotic Behavior and Operational Ranges

The two asymptotes of eq. (10) are given by

$$P_{-\infty} = \lim_{v \to -\infty} P(v) = \delta, \tag{11}$$

$$P_{+\infty} = \lim_{v \to +\infty} P(v) = \alpha. \tag{12}$$

To interpret the results eq. (11) and eq. (12), recall the following operating ranges.

2.4.1.1 No generation, $v < v_{ci}$

When the wind speed v is less than the cut-in wind speed v_{ci}, the wind turbine will not generate any power: $P = 0$. Hence, eq. (11) gives

$$P(v) = 0,\ \forall v < v_{\mathrm{ci}} \quad \Longrightarrow \quad P_{-\infty} = \delta = 0. \tag{13}$$

2.4.1.2 Rated Power, $v_{rated} < v < v_{co}$

When the wind speed v is above the rated wind speed v_{rated}, the wind turbine will produce the rated power $P = P_{\text{rated}}$. As stated in section 2.1.2.4, the safety shutdown operating range is neglected. With those considerations, eq. (12) gives

$$P(v) = P_{\text{rated}},\ \forall v > v_{\text{rated}} \quad \implies \quad P_{-\infty} = \alpha = P_{\text{rated}}. \tag{14}$$

2.4.2 Logistic Models with 2 and 3 Parameters

By combining the results eq. (13) and eq. (14) with eq. (10), and recalling the definition of normalized power as in eq. (2), the initial logistic model can be reduced to three free parameters:

$$p(v) = \left[1 + \exp\left(-\left(\frac{v - v_0}{s}\right)\right)\right]^{-1/\gamma}, \tag{15}$$

which defines the 3-parameter logistic model (L3P, for short). Applying the restriction $\gamma = 1$ keeps the sigmoidal shape of the curve and defines the 2-parameter logistic model (L2P, for short):

$$p(v) = \left[1 + \exp\left(-\left(\frac{v - v_0}{s}\right)\right)\right]^{-1}. \tag{16}$$

2.4.3 Linearizing Transformation and Parameter Initialization

Equation (16) can be re-written as

$$\log\left(p(v)^{-1} - 1\right) = (v_0/s) + (-s^{-1})v. \tag{17}$$

The coefficients $a = v_0/s$ and $b = -s^{-1}$ in the right-hand side of eq. (17) can be estimated by the ordinary least squares (OLS) method. This gives reasonable values for the parameters v_0 and s when $\gamma = 1$, providing a way to initialize them closer to their optimal value when fitting the wind speed and normalized power data (v_i, p_i).

2.5 Discussion

This chapter discussed multiple details of the WTPC modeling task, starting with the examination of a typical WTPC of a pitch regulated WT, which was described in terms of four design parameters and the four operating ranges associated with them. The peculiarities of each operating range and their effects on the wind speed and power (v_i, P_i) were analyzed, showcasing the important characteristics that a good WTPC model must be able to express.

The IEC standard (INTERNATIONAL ELECTROTECHNICAL COMMISSION, 2013) was discussed to provide an overview of the technical approach to the WTPC task. The WT's SCADA was established as the main data source for the problem, while the wind farm's meteorological mast data was considered as an option to account for the effects of varying air density through normalization. The implementation of the method of the bins, the mathematical model proposed by the IEC standard, was described to serve as a comparison benchmark in this dissertation.

A survey of the scientific literature has shown the high interest of the academic community on the WTPC modeling topic. A methodological classification for the WTPC models by Sohoni *et al.* (2016) as reviewed to highlight the vast possibilities to approach the problem. All the considered reviews showcased polynomial regressions and neural networks, which indicated them as representative comparison benchmarks. Previous works involving GPs (PANDIT; INFIELD, 2019; PANDIT *et al.*, 2019) and logistic functions (VILLANUEVA; FEIJÓO, 2018) were also discussed, as they are the constituting parts of the WTPC modeling framework this dissertation aims to propose.

Finally, a deep analysis of the logistic function model from Villanueva and Feijóo (2018) was conducted as a preparation for its usage in the new WTPC modeling framework. It was firstly shown that one of the 6-parameter logistic function (6PLE) can be eliminated by reparametrization. Then, the parameters corresponding to their asymptotes were shown to be constrained by the operating ranges of a WTPC, eliminating two other parameters. It resulted in the L3P model eq. (15) for the normalized power p. A further simplification led to the L2P model, whose linearizing transformation allowed for reasonable parameter initialization by means of ordinary least squares.

3 FUNDAMENTALS OF GAUSSIAN PROCESS MODELS

The usual strategy to deal the with uncertainties about an unknown function $f : \mathscr{X} \rightarrow \mathbb{R}$ is to assume it follows a parametrical equation depending on a parameter vector $\boldsymbol{\theta}_{\text{model}}$ which behaves according to a probability distribution. In this setting, the uncertainty comes from the randomness in the parameter vector $\boldsymbol{\theta}_{\text{model}}$ and is pushed into the function f. This construction constrains the function f to the hypothesized functional form.

GPs offer a different approach to deal with uncertainties in unknown functions. Instead of a parametric form, the function f is understood as an “infinitely-long vector” of function evaluations $\boldsymbol{f}$, which is assumed to behave as an “infinite-variate Gaussian distribution”. This informally defines a probability distribution over the entire function space $f : \mathscr{X} \rightarrow \mathbb{R}$ that is not constrained by a parametric form.

In this chapter, the GP is defined and used to model regression problems with a presentation largely based on Rasmussen and Williams (2006), aiming to support the application of it to WTPC modeling, which is mentioned whenever convenient. In section 3.1, the definition and basic properties of the multivariate Gaussian distribution are reviewed. In section 3.2, those properties are extended to define the GPs as a distribution over functions. In Section 3.3, GPs are applied to the regression with noisy observations task. The chapter is finished with a summarization of the discussed subject in section 3.4.

3.1 Multivariate Gaussian Distribution

GPs can be understood as an infinite-variate Gaussian distribution, and most of its properties are inherited from it. Hence, it is important to present a brief review of it. The notation is chosen as a preparation for the transition to GPs.

Let $\boldsymbol{f} \in \mathbb{R}^N$ be a random vector which follows a multivariate Gaussian distribution, i.e.:

$$\boldsymbol{f} \sim \mathscr{N}(\boldsymbol{m}, \boldsymbol{K}), \tag{18}$$

where $\boldsymbol{m} \in \mathbb{R}^N$ is the mean vector and $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ is the covariance matrix, which must be positive semi-definite (PSD), i.e.,

$$\boldsymbol{x}^\top \boldsymbol{K} \boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{19}$$

The probability density of $\boldsymbol{f}$ is given by:

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m},\boldsymbol{K}) = \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{K}|}} \exp\left[(\boldsymbol{f}-\boldsymbol{m})^\top \boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{m})\right], \tag{20}$$

where $|\boldsymbol{K}|$ is the determinant of the matrix $\boldsymbol{K}$.

The multivariate Gaussian distribution has two fundamental properties, namely, **marginalization** and **conditioning**. To explore them, consider two disjoint subsets $\boldsymbol{f}_1 \in \mathbb{R}^{N_1}$ and $\boldsymbol{f}_2 \in \mathbb{R}^{N_2}$, $N_1 + N_2 = N$, of the random vector $\boldsymbol{f}$:

$$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{m},\boldsymbol{K}), \quad \boldsymbol{m} = \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \quad \boldsymbol{K} = \begin{bmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{bmatrix}, \tag{21}$$

with $\boldsymbol{K}_{12} = \boldsymbol{K}_{21}^\top$. The properties are stated as follows.

- **Marginalization:** The subsets $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ can be analyzed separately, and each of them follows a multivariate Gaussian distribution:

$$\boldsymbol{f}_1 \sim \mathcal{N}(\boldsymbol{m}_1,\boldsymbol{K}_{11}), \quad \boldsymbol{f}_2 \sim \mathcal{N}(\boldsymbol{m}_2,\boldsymbol{K}_{22}) \tag{22}$$

- **Conditioning:** The analysis of the subset $\boldsymbol{f}_2$, given the observation of the subset $\boldsymbol{f}_1$, follows a multivariate Gaussian distribution:

$$\boldsymbol{f}_2|\boldsymbol{f}_1 \sim \mathcal{N}(\boldsymbol{m}_2 + \boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}(\boldsymbol{f}_1 - \boldsymbol{m}_1), \boldsymbol{K}_{22} - \boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{K}_{21}^\top). \tag{23}$$

These properties permit a simple yet effective way to use the multivariate Gaussian distribution to learn from data. Using the marginalization, one can first observe the subset $\boldsymbol{f}_1$. Then, by conditioning, one can analyze the subset $\boldsymbol{f}_2$ and incorporate information obtained from $\boldsymbol{f}_1$. To perform this task, the mean vectors $\boldsymbol{m}_i$ and the covariance matrices $\boldsymbol{K}_{ij}$ must be computed, which will be done by using GPs.

3.2 GP as a Distribution over Functions

The GP definition can now be constructed using the multivariate Gaussian distribution from section 3.1. The formal mathematical justifications are skipped in favor of a more intuitive approach which covers what is necessary for the objectives of this dissertation.

Let $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^D$, be a random function[1] which follows a GP distribution, i.e.:

$$f \sim \mathcal{GP}(\mu, \kappa), \tag{24}$$

[1] f maps inputs $\boldsymbol{x} \in \mathcal{X}$ to random output variables $f(\boldsymbol{x}) \in \mathbb{R}$.

where $\mu : \mathscr{X} \to \mathbb{R}$ and $\kappa : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ are the mean and covariance functions, respectively. It is not possible to express the probability density of a random function directly. Instead, consider a finite subset of N inputs, $\{\boldsymbol{x}_i\}_{i=1,\ldots,N} \subset \mathscr{X}$, represented as the input matrix $\boldsymbol{X} = \left[\boldsymbol{x}_i\right]_{N\times D} \in \mathbb{R}^{N\times D}$ and the corresponding random output vector $\boldsymbol{f} = f(\boldsymbol{X}) = \left[f(\boldsymbol{x}_i)\right]_{N\times 1} \in \mathbb{R}^N$. The defining property of a GP is that for any given input matrix $\boldsymbol{X}$, the corresponding random output vector $\boldsymbol{f}$ conditionally follows a multivariate Gaussian distribution, i.e.,

$$\boldsymbol{f}|\boldsymbol{X} \sim \mathscr{N}(\boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{K}_{\boldsymbol{XX}}), \tag{25}$$

where

$$\boldsymbol{m}_{\boldsymbol{X}} = \mu(\boldsymbol{X}) = \left[\mu(\boldsymbol{x}_i)\right]_{N\times 1} \in \mathbb{R}^N, \tag{26}$$

$$\boldsymbol{K}_{\boldsymbol{XX}} = \kappa(\boldsymbol{X}, \boldsymbol{X}) = \left[\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{N\times N} \in \mathbb{R}^{N\times N}, \tag{27}$$

are respectively the mean vector and covariance matrix, with the probability distribution of $\boldsymbol{f}|\boldsymbol{X}$ given by eq. (20). The matrix $\boldsymbol{K}_{\boldsymbol{XX}}$ generated by the covariance function κ must be PSD for any input matrix $\boldsymbol{X}$. A function with this property is called a PSD kernel function.

The marginalization and conditioning properties are naturally inherited by subsets of input-output pairs $(\boldsymbol{X}_1, \boldsymbol{f}_1)$ and $(\boldsymbol{X}_2, \boldsymbol{f}_2)$ of the GP f. By substituting $\boldsymbol{f}_i$ by $\boldsymbol{f_i}|\boldsymbol{X}_i$ and $\boldsymbol{f_2}|\boldsymbol{f}_1$ by $\boldsymbol{f_2}|\boldsymbol{X}_2, \boldsymbol{f}_1, \boldsymbol{X}_1$.

The GP definition completes the problem of learning, now in the context of input-output pairs $(\boldsymbol{X}_1, \boldsymbol{f}_1)$ and $(\boldsymbol{X}_2, \boldsymbol{f}_2)$, by providing a way to construct the mean vectors $\boldsymbol{m}_i$ and covariance matrices $\boldsymbol{K}_{ij}$ in eqs. (22) and (23). The marginalization is applied to observations of $\boldsymbol{f}_1$, given $\boldsymbol{X}_1$. Then, given a new input set $\boldsymbol{X}_2$, the conditioning can be used to incorporate information obtained from $(\boldsymbol{X}_1, \boldsymbol{f}_1)$ to analyze $\boldsymbol{f}_2$.

3.2.1 Mean Function

The mean function μ of a GP f is responsible for generating the mean vector $\boldsymbol{m}_{\boldsymbol{X}}$ of its random outputs $\boldsymbol{f}$ in terms of the inputs $\boldsymbol{X}$, which prescribes how the GP f behaves in the absence of information (no conditioning on previous observations). In the context of learning, the mean function μ can be used to directly incorporate prior knowledge into the GP f. For example, when modeling WTPCs, as is the focus of this dissertation, the L3P logistic function given by eq. (15), with parameter vector $\boldsymbol{\phi} = [v_0,\ s,\ \gamma]$. The contribution of the mean function μ can be analyzed in two steps.

First, consider how a multivariate Gaussian random variable is distributed around its mean. The mean is also the mode of the distribution, i.e., the more frequently observed value of it. Hence, when plotted, the observations of input-output pairs $(\boldsymbol{x}_i, f(\boldsymbol{x}_i))$ will be scattered around $(\boldsymbol{x}_i, \mu(\boldsymbol{x}_i))$.

Second, consider the process of learning with a GP f by conditioning in a set of previous observations $(\boldsymbol{X}_1, \boldsymbol{f}_1)$, $\boldsymbol{f}_1 = f(\boldsymbol{X}_1)$. As the GP inherits the conditioning property of the multivariate Gaussian, the expected value of a set of new observations $(\boldsymbol{X}_2, \boldsymbol{f}_2)$, $\boldsymbol{f}_2 = f(\boldsymbol{X}_2)$ conditioned on the previous ones will follow (23), which, in GP notation, can be expressed as

$$\mathbb{E}_{p(\boldsymbol{f}_2|\boldsymbol{X}_2,\boldsymbol{f}_1,\boldsymbol{X}_1)}[\boldsymbol{f}_2] = \boldsymbol{m}_{\boldsymbol{X}_2} + \boldsymbol{K}_{\boldsymbol{X}_2\boldsymbol{X}_1}\boldsymbol{K}_{\boldsymbol{X}_1\boldsymbol{X}_1}^{-1}(\boldsymbol{f}_1 - \boldsymbol{m}_{\boldsymbol{X}_1}). \tag{28}$$

The first term of the sum, $\boldsymbol{m}_{\boldsymbol{X}_2} = \mu(\boldsymbol{X}_2)$ is the same as the unconditioned mean, and hence accounts for the prior knowledge provided by the mean function μ to the GP f. The second term of the sum, $\boldsymbol{K}_{\boldsymbol{X}_2\boldsymbol{X}_1}\boldsymbol{K}_{\boldsymbol{X}_1\boldsymbol{X}_1}^{-1}(\boldsymbol{f}_1 - \boldsymbol{m}_{\boldsymbol{X}_1})$, is the effect of the learned information into the conditioned mean. It is interesting to note that this term is proportional to $\boldsymbol{f}_1 - \boldsymbol{m}_{\boldsymbol{X}_1} = f(\boldsymbol{X}_1) - \mu(\boldsymbol{X}_1)$, which can be understood as the *deviation* from the prior mean providing information to the new observations' mean.

3.2.2 Covariance Function

The covariance function κ of a GP f is responsible for generating the covariance matrix $\boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}}$ of its random outputs $\boldsymbol{f}$ in terms of the inputs $\boldsymbol{X}$, which plays a very important role in the learning problem.

To simplify the analysis, consider two input-output pairs $(\boldsymbol{x}_1, f(\boldsymbol{x}_1))$ and $(\boldsymbol{x}_2, f(\boldsymbol{x}_2))$. The covariance between the random outputs, $\mathrm{Cov}(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) = \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$, express how they jointly vary, i.e., how, for example, $f(\boldsymbol{x}_2)$ should change given that $f(\boldsymbol{x}_1)$ changed. As it is computed in terms of the inputs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, this joint variation is specified by (**a**) the corresponding input $\boldsymbol{x}_i$ of each observation, which is given by the data, and (**b**) the functional form of the covariance function κ. Hence, the covariance function κ manages *how* the GP f transfer what was learned to new observations.

3.2.2.1 Squared Exponential Covariance Function

The SE covariance function $\kappa_{\text{SE}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{+}^{*}$, whose expression is given by

$$\kappa_{\text{SE}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{i=1}^{D}\left(\frac{x_{1i} - x_{2i}}{l_i}\right)^2\right], \tag{29}$$

was studied and elected as one of the best options regarding WTPC modeling in Pandit and Infield (2019), and hence is chosen to be used in this dissertation. It generates a "similarity by proximity" behavior by increasing the covariance between the random outputs $\text{Cov}(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2))$ as the inputs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ becomes closer each other, and decreasing it otherwise.

The hyper-parameters of the SE covariance function are the elements of the $\boldsymbol{\theta} = [\sigma_f^2,\ l_1,\ \ldots,\ l_D]^\top$ vector. The variance σ_f^2 controls the scale of f, and the length scales l_i, $i = 1, \ldots, D$ measures how much two inputs need to move away from each other to become uncorrelated. When dealing with normalized dimensions, large values of l_i indicate that the i-th dimension of the input $\boldsymbol{x}$ is less relevant than the others, as its contribution to the sum is smaller – this is the automatic relevance determination (ARD) property.

3.3 Regression with Noisy Observations

Consider the problem of learning a relationship $\boldsymbol{x} \to y$ with a data set in the form of N observations $(\boldsymbol{x}_i, y_i) \in \mathscr{X} \times \mathscr{Y} \subseteq \mathbb{R}^D \times \mathbb{R}$, where the relationship is not exact, but rather stochastic: given the value of an input $\boldsymbol{x}_i$, the output y_i is not exactly defined. The inability to determine the output y_i exactly may derive either from incomplete knowledge of *how* it is related to $\boldsymbol{x}_i$ or from the intrinsic randomness of the process.

3.3.1 Regression Model

To tackle the noisy regression problem, assume there is a function f which describes how noisy observations y_i depend on inputs $\boldsymbol{x}_i$. The function f is unknown and is modeled as a GP, i.e., $f \sim \mathscr{GP}(\mu, \kappa)$, which is the model's *prior*. Furthermore, the effects of $f_i = f(\boldsymbol{x}_i)$ on y_i as well as the intrinsic randomness of the process is modeled as a probability distribution $p(y_i|f_i)$, the model's *likelihood*. This distribution is the same for each noisy observations y_i. This gives the following regression model:

$$p(\boldsymbol{f}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m_X}, \boldsymbol{K_{XX}}), \tag{30}$$

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i), \tag{31}$$

where $\boldsymbol{y} = \left[y_i\right]_{N\times 1}$ is the noisy-output vector, $\boldsymbol{f} = \left[f(\boldsymbol{x}_i)\right]_{N\times 1}$ is the latent function vector, $\boldsymbol{m_X}$ and $\boldsymbol{K_{XX}}$ are defined as in eqs. (26) and (27). The ability to write the model's likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ as the product of each observation's likelihood $p(y_i|f_i)$ defines what is called a *factorizing* likelihood, meaning that the noisy observation y_i is conditionally independent of *any* other random variable given the corresponding latent function evaluation f_i.

3.3.2 Likelihoods

Any probability distribution $\mathscr{P}(y|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = \left[\gamma_j\right]_{n\times 1} \in \mathbb{R}^n$ is its hyperparameter vector, can be used as the likelihood, and its choice is a way of making assumptions which can incorporate domain knowledge to the model. The notation

$$p(y_i|f_i) = \mathscr{P}(y_i|\gamma_{j^*} = f_i, \boldsymbol{\gamma'}), \quad \boldsymbol{\gamma'} = \left[\gamma_{j', j' \neq j^*}\right]_{(n-1)\times 1} \in \mathbb{R}^{n-1} \tag{32}$$

is used to emphasize how the noisy observations of y_i depend on the latent function evaluation f_i through the functional form of the probability distribution $\mathscr{P}(y_i|\boldsymbol{\gamma})$. The function $f_i = f(\boldsymbol{x}_i)$ takes the place of the hyperparameter γ_{j*}, thus making it dependent on the input $\boldsymbol{x}_i$, whereas the likelihood hyperparameter vector $\boldsymbol{\gamma'}$ has a global effect.

Two examples of likelihoods are now considered.

3.3.2.1 Gaussian Likelihood

A very common example in the general and GP specific literature is the Gaussian likelihood. It is built from the univariate Gaussian distribution as follows.

$$\mathcal{N}(y|\mu_y, \sigma_y^2) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right], \tag{33}$$

$$p(y_i|f_i) = \mathcal{N}(y_i|\mu_y = f_i, \sigma_y^2), \tag{34}$$

where

- The likelihood's mean μ, modeled by the latent function f_i, gives the expected value of the noisy observation y_i. It is also referred as the likelihood's location;

- The likelihood's variance $\sigma_y^2 > 0$ is a hyperparameter that globally controls the scattering of each noisy observation y_i around its $\mu_y = f_i$. Its positive square root $\sigma_y > 0$ is also referred as the likelihood's standard deviation or the likelihood's scale.

Using the Gaussian likelihood assumes a symmetric, homoscedastic and lightly-tailed noise behavior of y_i around $\mu = f_i$. It is especially attractive because, as will be shown in sections 3.3.3 and 3.3.4, it generates tractable expressions[2] for inference and predictions.

3.3.2.2 Student-t Likelihood

In some applications such as WTPC modeling, the lightly-tailed hypotesis will not be observed as there can be outliers in the data, which would cause a over-estimation of the likelihood's variance σ_y^2. For those cases, one can use the Student-*t* likelihood, which is built from the localized and scaled Student-*t* distribution as follows.

$$\mathscr{T}(y|\mu_y, \sigma_y, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma_y^2}} \left[1 + \frac{1}{\nu}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]^{-\frac{\nu+1}{2}}, \tag{35}$$

$$p(y_i|f_i) = \mathscr{T}(y_i|\mu_y = f_i, \sigma_y, \nu), \tag{36}$$

where

- The likelihood's location μ_y, modeled by the latent function f_i, gives the expected value of the noisy observation y_i;
- The likelihood's scale $\sigma_y > 0$ is a hyperparameter that globally controls the scattering of each noisy observation y_i around its $\mu_y = f_i$. Its square root σ_y, can also be referred as the standard deviation of scale;
- The likelihood's degrees-of-freedom $\nu > 2$ is a hyperparameter that globally controls how heavily-tailed the distribution of each noisy observation y_i is. As the degrees-of-freedom ν increases, the Student-*t* likelihood approaches the Gaussian likelihood i.e.,

$$\lim_{\nu\to\infty} \mathscr{T}(y|\mu_y, \sigma_y, \nu) = \mathscr{N}(y|\mu_y, \sigma_y^2). \tag{37}$$

Using the Student-*t* likelihood assumes a symmetric, homoscedastic and possibly heavily-tailed noise behavior of y_i around $\mu_y` = f_i$. The possibility of heavy tails makes it robust, i.e., able to deal with outliers in the data, at the expense of dealing with intractable expressions[3] for the model's evidence and posterior distributions, as will be shown in sections 3.3.3 and 3.3.4.

[2] Expressions whose computation is readily available, not requiring methods such as numerical integration.

[3] Expressions whose computation require numerical methods, mainly numerical integrations.

3.3.3 Inference

By combining eqs. (30) and (31) and marginalizing the latent function vector, one gets the *evidence* of the model:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}) &= \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})\mathrm{d}\boldsymbol{f} \\ &= \int p(\boldsymbol{y}|\boldsymbol{f})\mathcal{N}(\boldsymbol{f}|\boldsymbol{m_X},\boldsymbol{K_{XX}})\mathrm{d}\boldsymbol{f} \end{aligned} \tag{38}$$

which can be understood as the model's capacity to explain the data $\boldsymbol{X},\boldsymbol{y}$. Although not explicitly represented, the model's evidence depends on the parameter vector $\boldsymbol{\phi}$ of the mean function μ, the hyperparameter vector $\boldsymbol{\theta}$ of the covariance function κ and the hyper-parameters of the likelihood distribution $p(y_i|f_i)$.

The inference process is done by finding values for the parameters and hyper-parameters such that they properly explain the available data. Different approaches for learning the parameters and hyper-parameters, such as gradient-based optimization or cross-validation criteria are discussed in Rasmussen and Williams (2006). This dissertation follows the optimization approach by minimizing the negative log-evidence of the model (which is equivalent to maximizing the model's evidence).

When using the Gaussian likelihood, the evidence is tractable[4]. Substituting eq. (34) into eq. (38) and integrating, the evidence of the model with Gaussian likelihood is given by

$$p(\boldsymbol{y}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{m_X},\boldsymbol{K_y}), \quad \boldsymbol{K_y} = \boldsymbol{K_{XX}} + \sigma_y^2\boldsymbol{I}_N. \tag{39}$$

3.3.3.1 Computational Complexity and Memory Requirements

It is important to investigate the gradient-based optimization to verify the computational complexity and memory requirements of computing the derivatives involved in it. Taking the negative-log of (20) and substituting the parameters according to eq. (39), the negative log-evidence is

$$-\log p(\boldsymbol{y}|\boldsymbol{X}) = \frac{1}{2}(\boldsymbol{y}-\boldsymbol{m_X})^\top\boldsymbol{K_y}^{-1}(\boldsymbol{y}-\boldsymbol{m_X}) + \frac{1}{2}\log|\boldsymbol{K_y}| + \frac{n}{2}\log 2\pi. \tag{40}$$

[4] See footnote 2 in page 41.

The derivatives w.r.t. the mean function parameter vector $\boldsymbol{\phi}$, the covariance function hyperparameter vector $\boldsymbol{\theta}$ and the noise variance σ_y^2 are given by

$$\frac{\partial}{\partial \phi_j}\left(-\log p\left(\boldsymbol{y}|\boldsymbol{X}\right)\right) = \boldsymbol{\alpha}\frac{\partial \boldsymbol{m_X}}{\partial \phi_j}, \tag{41}$$

$$\frac{\partial}{\partial \theta_i}\left(-\log p\left(\boldsymbol{y}|\boldsymbol{X}\right)\right) = \frac{1}{2}\operatorname{tr}\left[\left(\boldsymbol{K_y}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right)\frac{\partial \boldsymbol{K_y}}{\partial \theta_i}\right], \tag{42}$$

$$\frac{\partial}{\partial \sigma_y^2}\left(-\log p\left(\boldsymbol{y}|\boldsymbol{X}\right)\right) = \frac{1}{2}\operatorname{tr}\left(\boldsymbol{K_y}^{-1} - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right), \tag{43}$$

where $\boldsymbol{\alpha} = \boldsymbol{K_y}^{-1}(\boldsymbol{y} - \boldsymbol{m_X})$ and $\operatorname{tr}\boldsymbol{A}$ is the trace of the matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$. The matrix inversion operation $\boldsymbol{K_y}^{-1}$ present in eqs. (42) and (43) requires a computation with $\mathscr{O}(N^3)$ complexity which is performed at each optimization step and demands a $\mathscr{O}(N^2)$ memory. For large data sets, the computational complexity and memory requirements can become prohibitively large, motivating the usage of approximate inference methods.

When using a non-Gaussian likelihood (e.g. Student-*t* likelihood), eq. (38) must be numerically integrated, which introduces additional computational complexity. Furthermore, computing multivariate Gaussian expectations as in eq. (38) requires numerical integration methods such as N-dimensional Gauss-Hermite quadrature or Monte Carlo quadrature. These methods usually involve the Cholesky factor of $\boldsymbol{K_{XX}}$, whose computation also has $\mathscr{O}(N^3)$ complexity and $\mathscr{O}(N^2)$ memory requirements.

3.3.4 Prediction

Once the parameters and hyper-parameters are learned from the data, it is possible to compute the distribution of the latent function vector $\boldsymbol{f}$ given the data $\boldsymbol{X}, \boldsymbol{y}$, i.e., the model's *posterior*, by using the Bayes' rule:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}, \quad p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})}. \tag{44}$$

For the Gaussian likelihood, eq. (44) is tractable and is given by

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}) = \mathscr{N}(\boldsymbol{f}|\boldsymbol{m_X} + \boldsymbol{K_{XX}}\boldsymbol{\alpha}, (\boldsymbol{K_{XX}^{-1}} + \sigma_y^{-2}\boldsymbol{I}_N)^{-1}). \tag{45}$$

Given a set of new inputs $\boldsymbol{x}'_j \in \mathscr{X}$, $j = 1, \ldots, N'$, grouped as the new input matrix $\boldsymbol{X}' = \left[\boldsymbol{x}'_j\right]_{N'\times D}$, the posterior $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})$ can be used to obtain the distribution of the new latent function evalution vector $\boldsymbol{f}' = \left[f(\boldsymbol{x}'_j)\right]_{N'\times 1}$ conditioned on the new inputs $\boldsymbol{X}'$ and on the data $\boldsymbol{X}, \boldsymbol{y}$. In this regard, the following expression is obtained:

$$p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{y}) = \int p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{f},\boldsymbol{X})p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})\mathrm{d}\boldsymbol{f}, \tag{46}$$

which is called the model's *posterior predictive*. The terms in eq. (46) describe how the information is transferred from the data to the new predictions. More especifically, one can make the following observations.

- The posterior $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})$ (see eq. (44)) describes what was learned about latent function evaluation vector $\boldsymbol{f}$ from the data $\boldsymbol{X},\boldsymbol{y}$;
- The GP-conditional distribution $p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{f},\boldsymbol{X})$ (see eq. (23)) transfers the information from the latent function evaluation vector $\boldsymbol{f}$ to the new latent function evaluation vector $\boldsymbol{f}'$ as both are part of the same GP f. It is important to remember that $\boldsymbol{f}$ was initially studied without regard to $\boldsymbol{f}'$, which is possible due to the marginalization property;
- The marginalization over $\boldsymbol{f}$ pushes the uncertainties about $\boldsymbol{f}$, which is not directly observable, to $\boldsymbol{f}'$.

For the Gaussian likelihood, eq. (46) is tractable and given by

$$p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{y}) = \mathscr{N}(\boldsymbol{f}'|\boldsymbol{m}_{\boldsymbol{X}'} + \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{\alpha}, \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}'} - \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}}^{-1}\boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}'}) \tag{47}$$

Finally, predictions about the noisy observations vector $\boldsymbol{y}' = \left[y'_j\right]_{N'\times 1}$ can be done with a further step. By hypothesis, the likelihood is the same for each observation (even for those yet to be seen). Hence, the model's predictions for the noisy observation vector $\boldsymbol{y}'$ relative to the new inputs $\boldsymbol{X}'$ and conditioned on the data $\boldsymbol{X},\boldsymbol{y}$ is given by

$$p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{y}) = \int p(\boldsymbol{y}'|\boldsymbol{f}')p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{y})\mathrm{d}\boldsymbol{f}', \tag{48}$$

where the marginalization over $\boldsymbol{f}'$ pushes the uncertainties about $\boldsymbol{f}'$ to $\boldsymbol{y}'$. For the Gaussian likelihood, eq. (48) is tractable and given by

$$p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{y}) = \mathscr{N}(\boldsymbol{f}'|\boldsymbol{m}_{\boldsymbol{X}'} + \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{\alpha}, \boldsymbol{K}_{\boldsymbol{y}'} - \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}}^{-1}\boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}'}), \tag{49}$$

where $\boldsymbol{K}_{\boldsymbol{y}'} = \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}'} + \sigma_y^2\boldsymbol{I}_{N'}$.

3.4 Discussion

In this chapter, the GP was introduced as way to deal with uncertainties about functions f by looking at it as a 'infinitely-long vector" of function evaluations $\boldsymbol{f}$. This enabled the construction of the GP as a distribution over the space of functions in the form $f : \mathscr{X} \to \mathbb{R}$ through the extension of the N-variate Gaussian distribution over the space of vectors $\boldsymbol{f} \in \mathbb{R}^N$. The properties of marginalization and conditioning of the multivariate Gaussian distribution were also extended to the GP, which enables it to be used for learning. Given some inputs $\boldsymbol{X}$ and their function evaluations $\boldsymbol{f}$, with $f_i = f(\boldsymbol{x}_i)$, one can use the GP properties to make more accurate predictions about new function evaluations $\boldsymbol{f}'$ at new inputs $\boldsymbol{X}'$.

The GP was defined in terms of its mean function $\mu : \mathscr{X} \to \mathbb{R}$ and covariance function $\kappa : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ and their parallel with the N-variate Gaussian distribution mean vector $\boldsymbol{m} \in \mathbb{R}^N$ and covariance matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ was estabilished. The mean function μ was presented as a way to incorporate prior knowledge about the GP-modeled function f into the model, whereas the covariance function κ was shown to define how the information obtained in some function evalutions $\boldsymbol{f}_1$ is transferred to the predictions of $\boldsymbol{f}_2$.

The regression with noisy observations problem, i.e., learning the stochastic relationship $\boldsymbol{x} \to y$ given N noisy observations pairs $(\boldsymbol{x}_i, y_i) \in \mathscr{X} \times \mathscr{Y}$, was approached within the GP framework. It was modeled by a latent function f, supposed to be a GP, and the likelihood, a probability distribution connecting $f_i = f(x_i)$ to y_i. The inference process on the model was studied, showing that the computational complexity and memory requirements can make it direct usage infeasible for large N, even for the Gaussian likelihood which gives tractable expressions. The predictive distribution was deduced and each element of it was analyzed, showcasing how the information is transferred from the initial data $\boldsymbol{X}, \boldsymbol{y}$ to new observations $\boldsymbol{y}'$ at inputs $\boldsymbol{X}'$.

4 VARIATIONAL INFERENCE APPLIED TO GAUSSIAN PROCESS MODELS

The Bayesian inference process consists in applying Bayes' rule to find the model's posterior distribution of latent variables given the data, which, apart from very specific cases, leads to intractable[1] expressions, as pointed out in section 3.3.3 regarding GP models. One option to deal with this problem is to use variational inference (BLEI *et al.*, 2017), a technique which transforms the intractable computations in an constrained optimization problem by proposing a parametric probability distribution, the so-called variational distribution, as an approximation to the model's posterior and minimizing the Kullback-Leiber (KL) divergence between them.

In this chapter, variational inference is applied to GP models. As will be shown, it not only solves the problem of intractable expressions but can also be used to overcome the computational complexity and memory requirements associated with standard inference. In section 4.1, the variational inference procedure is discussed specifically for GP models, begining with a brief review of the definition and properties of the KL divergence. In section 4.2, the sparse variational GP (SVGP) (TITSIAS, 2009; HENSMAN *et al.*, 2015) model is presented as an option to make GPs more computationally feasible. In section 4.3, the Chained GP (SAUL *et al.*, 2016), a model with multiple latent SVGPs, is analyzed, opening the possibilities of more complex likelihoods able to express more features of the noisy observations such as heteroscedasticity and localized heavy-tailedness. The chapter is finished with a summarization of the discussed subject in section 4.4.

4.1 Variational Inference on GP models

In this section, the variational inference procedure (BLEI *et al.*, 2017) is discussed in the context of the GP models described in chapter 3. It consists in avoiding intractable integrals needed to compute a model's evidence such as in eq. (38) and instead expressing it in terms of its posterior as in eq. (44), which is also unknown. To solve this, the model's evidence is written in terms of the KL divergence from a variational distribution to the model's posterior and is lower bounded using the Gibbs' inequality.

The resulting evidence lower bound (ELBO) is then used as an optimization objective, which gives an approximate inference scheme on the original model's evidence and also makes the variational distribution an approximation, in a KL divergence minimizing sense, to the

[1] See footnote 3 in page 41.

original model's posterior. The variational distribution is chosen in such a way that the resulting expressions become tractable.

4.1.1 Kullback-Leiber Divergence

Before describing the application of variational inference methods to GP models, it is useful to record the definition of the KL divergence. Given two probability distributions $p(x)$ and $q(x)$, the KL divergence[2] from $q(x)$ to $p(x)$ is given by

$$\mathbb{D}_{\mathrm{KL}}[q(x)\|p(x)] = \mathbb{E}_{q(x)}\left[\log\frac{q(x)}{p(x)}\right] = \int\left(\log\frac{q(x)}{p(x)}\right)q(x)\mathrm{d}x. \tag{50}$$

The Gibbs' inequality ensures its non-negativity, i.e.,

$$\mathbb{D}_{\mathrm{KL}}[q(x)\|p(x)] \geq 0, \tag{51}$$

with the equality happening only when $q(x) = p(x)$. Hence, the KL Divergence $\mathbb{D}_{\mathrm{KL}}[q(x)\|p(x)]$ can be interpreted as an asymmetric dissimilarity measure between the probability distributions $q(x)$ and $p(x)$.

Furthermore, given two M-variate Gaussian distributions $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_p, \boldsymbol{K}_p)$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_q, \boldsymbol{K}_q)$, the KL divergence from $q(\boldsymbol{x})$ to $p(\boldsymbol{x})$ is given by

$$\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{x})\|p(\boldsymbol{x})] = \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{K}_p^{-1}\boldsymbol{K}_q) + (\boldsymbol{m}_p - \boldsymbol{m}_q)^\top\boldsymbol{K}_p^{-1}(\boldsymbol{m}_p - \boldsymbol{m}_q) - M + \log\frac{|\boldsymbol{K}_p|}{|\boldsymbol{K}_q|}\right], \tag{52}$$

whose computation has $\mathcal{O}(M^3)$ complexity and $\mathcal{O}(M^2)$ memory requirement.

Finally, it can be shown that the KL divergence between the factorizing probability distributions $q(a,b) = q(a)q(b)$ and $p(a,b) = p(a)p(b)$ is the sum of the KL divergence between their factors, i.e,

$$\mathbb{D}_{\mathrm{KL}}[q(a,b)\|p(a,b)] = \mathbb{D}_{\mathrm{KL}}[q(a)\|p(a)] + \mathbb{D}_{\mathrm{KL}}[q(b)\|p(b)]. \tag{53}$$

4.1.2 Evidence Lower Bound

The deduction of the ELBO starts by isolating the model's evidence in eq. (44), which gives

$$p(\boldsymbol{y}|\boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})}. \tag{54}$$

[2] Also referred as the relative entropy of $q(x)$ w.r.t. $p(x)$.

Let $q(\boldsymbol{f})$ be a generic probability distribution, which aims to approximate the model's posterior, i.e., the variational posterior. Multiplying and dividing eq. (54) by the variational distribution $q(\boldsymbol{f})$ and taking the logarithm of both sides, one arrives at the following expression:

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = \log p(\boldsymbol{y}|\boldsymbol{f}) + \log\left(\frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})}\right) - \log\left(\frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{X})}\right). \tag{55}$$

The left-hand side of eq. (55) does not depend on $q(\boldsymbol{f})$. Taking the expectation w.r.t. $q(\boldsymbol{f})$ on both sides and noting the definition of the KL divergence as in eq. (50):

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] + \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f})\|p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f})\|p(\boldsymbol{f}|\boldsymbol{X})]. \tag{56}$$

The second term in the right-hand side of eq. (56) is still dependent on the posterior $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})$. Applying eq. (51) in eq. (56) gives the following inequality:

$$\log p(\boldsymbol{y}|\boldsymbol{X}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f})\|p(\boldsymbol{f}|\boldsymbol{X})], \tag{57}$$

whose left-hand side is called the model's evidence lower bound, as it is a lower bound for the model's log-evidence $\log p(\boldsymbol{y}|\boldsymbol{X})$.

4.1.3 Variational Distribution

The variational distribution $q(\boldsymbol{f})$ was purposely left undefined in section 4.1.2, intending to highlight the flexibility of choosing any valid probability distribution for it. In the upcoming sections, the variational inference methodology will be applied to extensions of the GP models of chapter 3 which aim to tackle the problems discussed in section 3.4. In this context, a clever choice of the variational distribution $q(\boldsymbol{f})$ will be fundamental in the deduction of tractable[3] expressions.

4.1.4 Optimization Objective

As discussed in section 3.3.3, the inference process aims to find values for the parameters and hyperparameters of the model such that they properly explain the data. Instead of directly optimizing the model's log-evidence, as was done previously, the ELBO in eq. (57) is used as the optimization objective, which indirectly forces the model's log-evidence to increase.

In general, the ELBO in eq. (57) depends not only on the parameter vector $\boldsymbol{\phi}$ of the mean function μ, the hyperparameter vector $\boldsymbol{\theta}$ of the covariance function κ and the

[3] See footnote 2 in page 41.

hyperparameters of the likelihood distribution $p(y_i|f_i)$ but also on any eventual variational parameters $\boldsymbol{\lambda}$ used to define the variational distribution $q(\boldsymbol{f})$. The gradient based optimization is performed w.r.t. all of these variables.

4.1.5 Variational Posterior Approximation

As the optimization of the ELBO proceeds, the slack of the inequality in eq. (57) gets tighter. This slack is precisely the KL divergence from the variational posterior to the true posterior $\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f})\|p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})]$ which is then compressed towards 0, although constrained by the proposed functional form of $q(\boldsymbol{f})$. This grantees that the obtained variational posterior $q(\boldsymbol{f})$ is, under the given restrictions, a good approximation (in a KL divergence sense) to the true and unknown posterior $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})$, i.e.,

$$q(\boldsymbol{f}) \approx p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}). \tag{58}$$

As will be shown, this approximation can be used to obtain approximate predictions.

4.2 Sparse Variational Gaussian Process Models

The sparse variational Gaussian process (SVGP) model was proposed in Titsias (2009) to tackle the computational complexity and memory requirements challenges imposed by standard GP regression inference, as discussed in the end of section 3.3.3. It was further developed in Hensman *et al.* (2015) to also deal with non-Gaussian likelihoods, which is how this section presents the subject (although using a slightly different notation and development). Readers looking for the more rigorous theoretical basis of the SVGP are referred to Matthews *et al.* (2016), Matthews (2017).

4.2.1 Augmented Regression Model

The SVGP approach can be described as the augmentation of the standard GP regression model from section 3.3.1 with M *pseudo-inputs* $\boldsymbol{z}_i \in \mathscr{X}$, $M \ll N$, grouped in the pseudo-input matrix $\boldsymbol{Z} = \left[\boldsymbol{z}_i\right]_{M\times D}$, and the corresponding *inducing variables* vector $\boldsymbol{u} = \left[f(\boldsymbol{z}_i)\right]_{M\times 1}$, which is connected to the latent function vector $\boldsymbol{f}$ through the same GP f. Due to the marginalization property of the GP, the pseudo-inputs $\boldsymbol{Z}$ and the inducing variables $\boldsymbol{u}$ do not change the latent function vector $\boldsymbol{f}$ behavior. Also, due to the assumption of conditional independence of noisy observations $\boldsymbol{y}$ given their corresponding latent function evaluations $\boldsymbol{f}$

from other variables, the model likelihood can be expressed as

$$p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{f}). \tag{59}$$

4.2.2 Variational Inference

Inference on the SVGP model is done by adapting the variational inference procedure from section 4.1 to account for the the model augmentation, resulting in a tractable ELBO for the model's log-evidence and an approximate model's posterior. The chosen approximating variational distribution enables the reduction of the computational complexity and memory requirements to $\mathscr{O}(NM^2)$ and $\mathscr{O}(M^2)$ respectively, which is the main merit of the SVGP.

4.2.2.1 Evidence Lower Bound

The SVGP ELBO is obtained by accounting for the pseudo-inputs $\boldsymbol{Z}$ and inducing variables $\boldsymbol{u}$. This is done by swapping $\boldsymbol{X}$ for $\boldsymbol{X},\boldsymbol{Z}$ and $\boldsymbol{f}$ for $\boldsymbol{f},\boldsymbol{u}$ in eq. (57), which gives

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}) \geq \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{u})\right] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f},\boldsymbol{u})\|p(\boldsymbol{f},\boldsymbol{u}|\boldsymbol{X},\boldsymbol{Z})], \tag{60}$$

where $q(\boldsymbol{f},\boldsymbol{u})$ is the variational posterior, which approximates the model's posterior $p(\boldsymbol{f},\boldsymbol{u}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y})$.

The expectation term $\mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{u})\right]$ in eq. (60) can be simplified using eq. (59) and noting that it is independent of the inducing variables $\boldsymbol{u}$, which gives

$$\mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{u})\right] = \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right]. \tag{61}$$

Substituting eq. (61) in eq. (60) gives

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}) \geq \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{u})\right] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f},\boldsymbol{u})\|p(\boldsymbol{f},\boldsymbol{u}|\boldsymbol{X},\boldsymbol{Z})]. \tag{62}$$

The right-hand side of eq. (62) is the SVGP ELBO. For factorizing likelihoods as the ones in eq. (31), it can be expressed as a summation over the data:

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}) \geq \left(\sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\log p(y_i|f_i)\right]\right) - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f},\boldsymbol{u})\|p(\boldsymbol{f},\boldsymbol{u}|\boldsymbol{X},\boldsymbol{Z})], \tag{63}$$

where $q(f_i)$ is the marginal distribution of each element f_i of the latent function evaluation vector $\boldsymbol{f}$.

4.2.2.2 Variational Distribution

As discussed in section 4.1.3, the variational distribution $q(\boldsymbol{f},\boldsymbol{u})$ needs to be cleverly chosen to produce tractable expressions for optimization. Aiming to simplify the KL divergence in eq. (63), the factorization

$$q(\boldsymbol{f},\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{u}) \tag{64}$$

is imposed, with the inducing variables variational distribution $q(\boldsymbol{u})$ independent of the latent function evalutions $\boldsymbol{f}$. This makes the following simplification possible:

$$\begin{aligned}
\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f},\boldsymbol{u})\|p(\boldsymbol{f},\boldsymbol{u}|\boldsymbol{X},\boldsymbol{Z})] &\overset{1}{=} \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log\frac{q(\boldsymbol{f},\boldsymbol{u})}{p(\boldsymbol{f},\boldsymbol{u}|\boldsymbol{X},\boldsymbol{Z})}\right] \\
&\overset{2}{=} \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log\frac{p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{u})}{p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})p(\boldsymbol{u}|\boldsymbol{Z})}\right] \\
&\overset{3}{=} \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{u})}\left[\log\frac{q(\boldsymbol{u})}{p(\boldsymbol{u}|\boldsymbol{Z})}\right] \\
&\overset{4}{=} \mathbb{E}_{q(\boldsymbol{u})}\left[\log\frac{q(\boldsymbol{u})}{p(\boldsymbol{u}|\boldsymbol{Z})}\right] \\
&\overset{5}{=} \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{u})\|p(\boldsymbol{u}|\boldsymbol{Z})].
\end{aligned} \tag{65}$$

The steps 1-5 of eq. (65) are justified as follows:

1. KL divergence definition, eq. (50);
2. Variational distribution factorization, eq. (64), and conditioning for the GP f, eq. (23);
3. Equal terms cancelation;
4. Marginalization over the latent function evaluations $\boldsymbol{f}$ since the expression inside the expectation does not depend on it;
5. KL divergence definition, eq. (50).

Substituing eq. (65) into eq. (63) gives

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z}) \geq \left(\sum_{i=1}^{N}\mathbb{E}_{q(f_i)}[\log p(y_i|f_i)]\right) - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{u})\|p(\boldsymbol{u}|\boldsymbol{Z})]. \tag{66}$$

For mathematical tractability of eq. (66), $q(\boldsymbol{u})$ is chosen to be a M-variate normal distribution:

$$q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{m_u},\boldsymbol{L_u}\boldsymbol{L_u}^\top), \tag{67}$$

where $\boldsymbol{m_u} \in \mathbb{R}^M$ is the variational mean vector and the matrix $\boldsymbol{L_u} \in \mathbb{R}^{M\times M}$ is lower-triangular, which ensures the variational covariance matrix $\boldsymbol{L_u}\boldsymbol{L_u}^\top$ is PSD, as required by eq. (20). With this choice, both the variational posterior $q(\boldsymbol{u})$ and the prior $p(\boldsymbol{u}|\boldsymbol{Z})$ of the inducing variables $\boldsymbol{u}$ are M-variate Gaussian distributions, which enables the KL divergence from the former to the latter $\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{u})\|p(\boldsymbol{u}|\boldsymbol{Z})]$ to be computed with eq. (52).

The expectation in eq. (66) depends on the marginal distributions $q(f_i)$ of the elements f_i from the latent function evaluation vector $\boldsymbol{f}$, whose joint distribution $q(\boldsymbol{f})$ can be obtained by marginalizing $q(\boldsymbol{u})$ in eq. (64), giving

$$q(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{u})\mathrm{d}\boldsymbol{u}. \tag{68}$$

Since both the latent function evaluations $\boldsymbol{f}$ and the inducing variables $\boldsymbol{u}$ are function evaluations of the same GP f at inputs $\boldsymbol{X}$ and pseudo-inputs $\boldsymbol{Z}$, eq. (23) gives

$$\begin{aligned} p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z}) &= \mathcal{N}(\boldsymbol{f}|\boldsymbol{m}_{\boldsymbol{f}|\boldsymbol{X}\boldsymbol{r},\boldsymbol{u},\boldsymbol{Z}},\boldsymbol{K}_{\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z}}), \\ \boldsymbol{m}_{\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z}} &= \boldsymbol{m}_{\boldsymbol{X}} + \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{Z}}^{-1}(\boldsymbol{u}-\boldsymbol{m}_{\boldsymbol{Z}}), \\ \boldsymbol{K}_{\boldsymbol{f}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z}} &= \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}} - \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{Z}}^{-1}\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{X}}. \end{aligned} \tag{69}$$

Substituting eq. (69) into eq. (68) gives

$$\begin{aligned} q(\boldsymbol{f}) &= \mathcal{N}(\boldsymbol{f}|\boldsymbol{m}_{\boldsymbol{f}},\boldsymbol{K}_{\boldsymbol{f}}), \\ \boldsymbol{m}_{\boldsymbol{f}} &= \boldsymbol{m}_{\boldsymbol{X}} + \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{Z}}^{-1}(\boldsymbol{m}_{\boldsymbol{u}}-\boldsymbol{m}_{\boldsymbol{Z}}) \\ \boldsymbol{K}_{\boldsymbol{f}} &= \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}} + \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{Z}}^{-1}\left(\boldsymbol{L}_{\boldsymbol{u}}\boldsymbol{L}_{\boldsymbol{u}}^{\top} - \boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{Z}}\right)\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{Z}}^{-1}\boldsymbol{K}_{\boldsymbol{Z}\boldsymbol{X}}. \end{aligned} \tag{70}$$

As stated by eq. (70), the joint distribution $q(\boldsymbol{f})$ of the latent function evaluation vector $\boldsymbol{f}$ is a N-variate Gaussian distribution. Hence, the marginal distribution of each of its elements f_i is given by

$$q(f_i) = \mathcal{N}\left(f_i|m_{f_i},k_{f_{ii}}\right), \tag{71}$$

where the distribution's mean m_{f_i} is the i-th element of the mean vector $\boldsymbol{m}_{\boldsymbol{f}}$, and distribution's variance $k_{f_{ii}}$ is the i-th element of the main diagonal of the covariance matrix $\boldsymbol{K}_{\boldsymbol{f}}$, respectively.

4.2.2.3 Computational Complexity and Memory Requirements

It is now possible to analyze the computational complexity and memory requirements for the SVGP ELBO given by eq. (66), which requires the evaluation of the following items:

- The KL divergence of two M-variate Gaussians, as in eq. (52). As stated in section 4.1.1, it has $\mathcal{O}(M^3)$ computational complexity and $\mathcal{O}(M^2)$ memory requirement;

- The distribution described by eq. (70). The matrix inversion only involves $M \times M$ matrices, and hence has $\mathscr{O}(M^3)$ computational complexity and $\mathscr{O}(M^2)$ memory requirement. The matrix products, however, involve $N \times M$ and $M \times M$ matrices, which has $\mathscr{O}(NM^2)$ complexity complexity and $\mathscr{O}(NM)$ additional memory requirement. Under the assumption that $N \gg M$, the computational complexity is $\mathscr{O}(NM^2)$ and the memory requirement is $\mathscr{O}(NM)$;
- The summation of N uni-dimensional Gaussian expectations following the marginal distributions of each f_i. For a general likelihood, this results in the evaluation of N independent quadrature algorithms[4]. If the likelihood is Gaussian, a closed form expression is, as usual, available:

$$‘\mathbb{E}_{q(f_i)}\left[\log p(y_i|f_i)\right] = -\frac{1}{2}\left[\log(2\pi\sigma_y^2) + \frac{(y_i - m_{f_i})^2 + k_{f_{ii}}}{\sigma_y^2}\right]. \tag{72}$$

This leads to a computational complexity of order $\mathscr{O}(NM^2)$ and a memory requirement of order $\mathscr{O}(NM)$. Selecting a number M of pseudo inputs $\boldsymbol{Z}$ and inducing variables $\boldsymbol{u}$ such that $M \ll N$ results in computations with much smaller computational and memory requirements than the standard GP regression inference.

4.2.3 Prediction

As discussed in section 4.1.5, the learned variational posterior $q(\boldsymbol{f}, \boldsymbol{u})$ can be used to perform predictions by taking advantage of the equivalent of eq. (58) for the SVGP, i.e,

$$q(\boldsymbol{f}, \boldsymbol{u}) \approx p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}). \tag{73}$$

[4] Each summand is a Gaussian expectation with mean m_{f_i} and variance $k_{f_{ii}}$depending on i as in eq. (71). The same quadrature algorithm, e.g. Gauss-Hermite quadrature, can be used to evaluate it for each i, but it need to be run with different parameters N times.

Consider, as in section 3.3.4, a set of new inputs grouped in the input matrix $\boldsymbol{X}'$ and their corresponding function evalutions $\boldsymbol{f}'$. The SVGP's (variational) posterior predictives is:

$$\begin{aligned}
p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}) &\overset{1}{=} \int p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{f},\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})p(\boldsymbol{f},\boldsymbol{u},|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y})\mathrm{d}\boldsymbol{f}\mathrm{d}\boldsymbol{u} \\
&\overset{2}{\approx} \int p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{f},\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{f},\boldsymbol{u})\mathrm{d}\boldsymbol{f}\mathrm{d}\boldsymbol{u} \\
&\overset{3}{=} \int p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{f},\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})p(\boldsymbol{f}|\boldsymbol{f},\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{u})\mathrm{d}\boldsymbol{f}\mathrm{d}\boldsymbol{u} \\
&\overset{4}{=} \int p(\boldsymbol{f}',\boldsymbol{f}|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{u})\mathrm{d}\boldsymbol{f}\mathrm{d}\boldsymbol{u} \\
&\overset{5}{=} \int p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{u},\boldsymbol{Z})q(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\
&\overset{6}{=} q(\boldsymbol{f}'). \qquad (74)
\end{aligned}$$

The steps 1-6 of eq. (74) are explained as follows.

1. Posterior predictive distribition definition, as in eq. (46);
2. Variational posterior approximation, as in eq. (73);
3. Variational posterior definition, as in eq. (64);
4. Conditional distribution definition;
5. Marginalization property of the GP f. Note that the latent function evaluations $\boldsymbol{f}$, the new function evaluations $\boldsymbol{f}'$ and the inducing variables $\boldsymbol{u}$ are all evalutions of from the same latent GP f;
6. Similarity with the variational posterior definition, as in eq. (64), swapping $\boldsymbol{f}$ and $\boldsymbol{X}$ for $\boldsymbol{f}'$ and $\boldsymbol{X}'$, respectively.

The result in eq. (74) allows the computation of $q(\boldsymbol{f'})$ from eq. (70) by swapping $\boldsymbol{X}$ for $\boldsymbol{X}'$. It is also worth noting that it does not depend directly on the data $\boldsymbol{X},\boldsymbol{y}$, which can be interpreted as the information obtained through the SVGP framework being concentrated only in $\boldsymbol{Z},\boldsymbol{u}$.

The prediction of the new noisy observations $\boldsymbol{y}'$ can be approximately computed by substituing the variational posterior predictive from eq. (74) into an adaptation of eq. (48) for the SVGP, given by

$$\begin{aligned}
p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}) &= \int p(\boldsymbol{y}'|\boldsymbol{f}')p(\boldsymbol{f}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y})\mathrm{d}\boldsymbol{f}' \\
&\approx \int p(\boldsymbol{y}'|\boldsymbol{f}')q(\boldsymbol{f}')\mathrm{d}\boldsymbol{f}'. \qquad (75)
\end{aligned}$$

For a general likelihood, eq. (75) requires numerical integration. In the Gaussian likelihood case, the integral is tractable and gives the following analytical expression:

$$p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}'|\boldsymbol{m}_{\boldsymbol{y}'},\boldsymbol{K}_{\boldsymbol{y}'}), \tag{76}$$

$$\boldsymbol{m}_{\boldsymbol{y}'} = \boldsymbol{m}_{\boldsymbol{X}'} + \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}(\boldsymbol{m}_{\boldsymbol{u}} - \boldsymbol{m}_{Z})$$

$$\boldsymbol{K}_{\boldsymbol{y}'} = \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{X}'} + \boldsymbol{K}_{\boldsymbol{X}'\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\left(\boldsymbol{L}_{\boldsymbol{u}}\boldsymbol{L}_{\boldsymbol{u}}^{\top} - \boldsymbol{K}_{\boldsymbol{ZZ}}\right)\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}'} + \boldsymbol{I}_{N'}\sigma_{\varepsilon}^{2}.$$

4.3 Chained Gaussian Process Models

The variational heteroscedastic Gaussian process (VHGP) model was proposed in Lázaro-Gredilla and Titsias (2011) to deal with heteroscedastic behavior in the data by adding of another latent GP to model it. Although successfully being able to represent the dependence of the scattering of the output y on the input $\boldsymbol{x}$, the VHGP has the same computational complexity and memory requirements of the standard GP regression from section 3.3.

Expanding on the idea of multiple latent GPs to model more complex features of the data, the chained Gaussian process was then proposed in Saul *et al.* (2016). It incorporates the scalability improvements of the SVGP on each latent GP in a integrated variational inference scheme, allowing the usage of more detailed likelihoods to model the regression problem in hand.

4.3.1 Multiple Latent GP Regression Model

Consider L unknown functions $f^j : \mathscr{X} \to \mathbb{R}$, each of them modeled as independent GPs, i.e., $f^j \sim \mathscr{GP}(\mu^j, \kappa^j)$, grouped in the multi-output function $F : \mathscr{X} \to \mathbb{R}^L$, which describes how noisy observations y_i depend on inputs $\boldsymbol{x}_i$. The evaluation of the j-th function f^j on the i-th input $\boldsymbol{x}_i$ is denoted $f_i^j = f^j(\boldsymbol{x}_i)$, and hence the evaluation of the multi-output function F on it can be expressed as the (row) vector $\boldsymbol{f}_i = F(\boldsymbol{x}_i) = \left[f^j(\boldsymbol{x}_i)\right]_{1\times L} = \left[f_i^j\right]_{1\times L}$, the likelihood is given by the probability distribution $p(y_i|\boldsymbol{f}_i) = p\big(y_i|f_i^1,\ldots,f_i^L\big)$. As in section 3.3.1, the likelihood probability distribution is the same for each noisy observation y_i. This gives the following regression model:

$$p(\boldsymbol{f}^j|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}^j|\boldsymbol{m}_{\boldsymbol{X}}^j,\boldsymbol{K}_{\boldsymbol{XX}}^j), \quad j = 1,\ldots,L \tag{77}$$

$$p(\boldsymbol{F}|\boldsymbol{X}) = \prod_{j=1}^{L} p(\boldsymbol{f}^j|\boldsymbol{X}), \tag{78}$$

$$p(\boldsymbol{y}|\boldsymbol{F}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{f}_i) \tag{79}$$

where $\boldsymbol{y} = [y_i]_{N\times 1}$ is the noisy-output (column) vector, $\boldsymbol{f}^j = [f^j(\boldsymbol{x}_i)]_{N\times 1} = [f_i^j]_{N\times 1}$ is the j-th latent function (column) vector, $\boldsymbol{F} = [f^j(\boldsymbol{x}_i)]_{N\times L} = [f_i^j]_{N\times L}$ is the multi-output latent function matrix, which is built using the latent function vectors $\boldsymbol{f}^j$ as columns or, equivalently, by using the multi-output function evaluation vectors $\boldsymbol{f}_i$ as its rows, $\boldsymbol{m}_{\boldsymbol{X}}^j$ and $\boldsymbol{K}_{\boldsymbol{XX}}^j$ are defined as in eqs. (26) and (27), for each $j = 1, \dots, L$. This model is a detailed statement of the one proposed by Saul *et al.* (2016), with the extension that the GPs which model the latent functions f^j are allowed to have non-zero mean functions μ^j.

In this dissertation, models with $L = 2$ and $L = 3$ latent functions f^j are considered. For notational convenience, they will be denoted as $f^1 = f$, $f^2 = g$ and $f^3 = h$, respectively. In the more general case with $L = 3$, the regression model eqs. (77) to (79) can be written as

$$p(\boldsymbol{f}|\boldsymbol{X}) = \mathscr{N}(\boldsymbol{f}|\boldsymbol{m}_{\boldsymbol{X}}^f, \boldsymbol{K}_{\boldsymbol{XX}}^f), \tag{80}$$

$$p(\boldsymbol{g}|\boldsymbol{X}) = \mathscr{N}(\boldsymbol{g}|\boldsymbol{m}_{\boldsymbol{X}}^g, \boldsymbol{K}_{\boldsymbol{XX}}^g), \tag{81}$$

$$p(\boldsymbol{h}|\boldsymbol{X}) = \mathscr{N}(\boldsymbol{h}|\boldsymbol{m}_{\boldsymbol{X}}^h, \boldsymbol{K}_{\boldsymbol{XX}}^h), \tag{82}$$

$$p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}). \tag{83}$$

4.3.2 Likelihoods depending on Multiple Latent GPs

In this disseration, the focus is restricted to likelihoods $p(y_i|\boldsymbol{f}_i)$ dealing with heteroscedasticity and localized outliers, which are the main features of the WTPC modeling problem. They can be constructed as an extension of the ones considered in section 3.3.2 by using additional latent functions evaluations f_i^j to model other hyperparameters of the probability distribution $\mathscr{P}(y_i|\boldsymbol{\gamma})$ in addition to the usual approach of modeling the location hyperparamter.

Three likelihoods are now considered.

4.3.2.1 Heteroscedastic Gaussian Likelihood

The heteroscedastic Gaussian likelihood was initially proposed by in Lázaro-Gredilla and Titsias (2011) and is built by considering two latent functions f and g to respectively model the location and scale hyperparameters μ and σ_y of a univariate Gaussian distribution as in eq. (33). As the likelihood's scale needs to be positive, i.e., $\sigma_y > 0$, it cannot be directly modeled by the latent function g, which can assume negative values. To overcome this constrained, a positive transforming function $t : \mathbb{R} \to \mathbb{R}_+$ is used to warp g, which gives the following likelihood[5]:

$$p(y_i|f_i,g_i) = \mathscr{N}(y_i|\mu_y = f_i, \sigma_y^2 = t(g_i)^2), \tag{84}$$

where the likelihood's location μ_y and scale σ_y have the same interpretation as in eq. (33).

The heteroscedastic Gaussian likelihood still assumes symetric and lightly-tailed noise behavior for y_i around $\mu_y = f_i$, but now but presents heteroscedasticity as the likelihood's scale σ_y is not global and instead depends on $\boldsymbol{x}_i$ through $g_i = g(\boldsymbol{x}_i)$. Regarding the positive-transforming function t, the choice $t(\cdot) = \exp(\cdot)$ has computational benefits as it leads to some tractable expressions[6] for inference, as will be shown in section 4.3.3.

4.3.2.2 Heteroscedastic Student-t Likelihood

The heteroscedastic Student-t likelihood was initially proposed in Saul *et al.* (2016) and is built in the same way as the heteroscedastic Gaussian, only swapping the univariate Gaussian distribution from eq. (33) for the localized and scaled Student-t distribution from eq. (35). This gives the following likelihood [7]:

$$p(y_i|f_i,g_i) = \mathscr{T}(y_i|\mu_y = f_i, \sigma_y = t(g_i), \nu), \tag{85}$$

where the likelihood's location μ_y, scale σ_y and degrees of freedom ν have the same interpretation a‘s in eq. (35), and $t : \mathbb{R} \to \mathbb{R}_+$ is some positive-transforming function.

The heteroscedastic Student-t likelihood assumes symmetric and heteroscedastic behavior for y_i. It is robust to outliers due to the possibility of heavy-tailed noise behavior controlled by the likelihood's degrees of freedom $\nu > 2$, a hyperparameter with global effect. Regarding the positive-transforming function t, the choice $t(g) = \exp(g)$ does not give tractable expressions, but is still a valid and useful choice.

4.3.2.3 Locally Robust Heteroscedastic Student-t Likelihood

The locally robust heteroscedastic Student-t likelihood is built as an extension of the heteroscedastic Student-t likelihood by including one more latent function h to model the degrees of freedom hyperparameter ν of the Student-t distribution as in eq. (35). Due to the

[5] In Lázaro-Gredilla and Titsias (2011), the latent function g is used to model the *variance* σ_y^2 hyperparameter, i.e., $\sigma_y^2 = t(g_i)$.

[6] See footnote 2 in page 41.

[7] In Saul *et al.* (2016), the latent function g is used to model the *squared scale* σ_y^2 hyperparameter, i.e., $\sigma_y^2 = t(g_i)$.

restriction $\nu > 2$, another transforming function $t' : \mathbb{R} \to [2, \infty)$ is used to warp h, resulting in the following likelihood:

$$p(y_i | f_i, g_i, h_i) = \mathscr{T}(y_i | \mu_y = f_i, \sigma_y = t(g_i), \nu = t'(h)), \tag{86}$$

where the likelihood's location μ_y and scale σ_y have the same interpretation as in eq. (35), and $t : \mathbb{R} \to \mathbb{R}_+$ is some positive-transforming function.

The locally robust heteroscedastic Student-t likelihood assumes symmetric and heteroscedastic behavior for y_i. It is robust to outliers due to the possibility of localized heavy-tailed noise behavior controlled by the likelihood's degrees of freedom $\nu > 2$, which depends on $\boldsymbol{x}_i$ through $h_i = h(\boldsymbol{x}_i)$. The transforming functions t and t' can be choosen as $t(g) = \exp(g)$ and $t'(h) = 2 + \exp(h)$.

4.3.3 Variational Inference

As was pointed out in section 3.3.3, exact inference on the GP regression model is only possible for the Gaussian likelihood and even in this case can have prohibitive computational complexity and memory requirements. The Chained GP regression model not only deals with intractable likelihoods but also with multiple GPs, which aggravates the difficulties of exact inference from the single GP case. Hence, the SVGP framework discussed in section 4.2 is applied to each GP.

The variational inference methodology is very similar to the one discussed in section 4.2.2. For each latent function f^j, the model is augmented with M^j pseudo-inputs $\boldsymbol{z}_i^j \in \mathscr{X}$, grouped in the pseudo-input matrix $\boldsymbol{Z}_i^j = \left[\boldsymbol{z}_i^j\right]_{M^j \times D}$, and the corresponding inducing variables vector $\boldsymbol{u}^j = \left[f^j(\boldsymbol{z}_i^j)\right]_{M^j \times 1}$. For notational convenience, define the pseudo-input set $\tilde{\boldsymbol{Z}} = \{\boldsymbol{Z}^j\}_{l=1}^L$ and the inducing variable set $\boldsymbol{U} = \{\boldsymbol{u}^j\}_{l=1}^L$. In the case where $M_l = M, \forall l = 1, \dots, L$, those sets can be expressed as the pseudo-input third-rank tensor $\tilde{\boldsymbol{Z}} = \left[\boldsymbol{Z}^j\right]_{M \times D \times L} = \left[\boldsymbol{z}_i^j\right]_{M \times D \times L}$ and the inducing variable matrix $\boldsymbol{U} = \left[\boldsymbol{u}^j\right]_{M \times L} = \left[f^j(\boldsymbol{z}_i^j)\right]_{M \times L}$. As was the case with the SVGP, the conditional independence of the observations $\boldsymbol{y}$ given their corresponding function evaluations $\boldsymbol{F}$ form other variables allows the model likelihood to be expressed as

$$p(\boldsymbol{y} | \boldsymbol{F}, \boldsymbol{U}) = p(\boldsymbol{y} | \boldsymbol{F}). \tag{87}$$

This variational inference methodology is very similar to the one proposed in Saul *et al.* (2016), but relaxing the original hypothesis of all inducing variables vector $\boldsymbol{u}^j$ having the same size M and sharing the same pseudo-input matrix $\boldsymbol{Z} \in \mathbb{R}^{M\times D}$, i.e, $\boldsymbol{Z}^j = \boldsymbol{Z}$, $\forall j = 1,\dots,L$. Below is a detailed deduction of the Chained ELBO, leading to results in accordance with the original reference.

4.3.3.1 Evidence Lower Bound

The Chained GP ELBO is obtained by accounting for all the latent functions evaluations $\boldsymbol{F}$, the pseudo-inputs $\tilde{\boldsymbol{Z}}$ and inducing variables $\boldsymbol{U}$. This is done by swapping $\boldsymbol{f}$ for $\boldsymbol{F}$, $\boldsymbol{X}$ for $\boldsymbol{X},\boldsymbol{Z}$ and $\boldsymbol{f}$ for $\boldsymbol{F},\boldsymbol{U}$ in eq. (57), which gives

$$\log p(\boldsymbol{y}|\boldsymbol{X},\tilde{\boldsymbol{Z}}) \geq \mathbb{E}_{q(\boldsymbol{f},\boldsymbol{U})}\left[\log p(\boldsymbol{y}|\boldsymbol{F},\boldsymbol{U})\right] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{F},\boldsymbol{U})\|p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}})], \tag{88}$$

where $q(\boldsymbol{F},\boldsymbol{U})$ is the variational posterior, which approximates the model's posterior $p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y})$.

As was the case with the SVGP eqs. (59) and (61), the chained GP conditional independence eq. (87) substituted in eq. (88) gives

$$\log p(\boldsymbol{y}|\boldsymbol{X},\tilde{\boldsymbol{Z}}) \geq \mathbb{E}_{q(\boldsymbol{F})}\left[\log p(\boldsymbol{y}|\boldsymbol{F})\right] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{F},\boldsymbol{U})\|p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}})]. \tag{89}$$

The right-hand side of eq. (89) is the Chained GP ELBO. For factorizing likelihoods as the ones in eq. (79), it can be expressed as a summation over the data:

$$\log p(\boldsymbol{y}|\boldsymbol{X},\tilde{\boldsymbol{Z}}) \geq \left(\sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{f}_i)}\left[\log p(y_i|\boldsymbol{f}_i)\right]\right) - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{F},\boldsymbol{U})\|p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}})], \tag{90}$$

where $q(\boldsymbol{F},\boldsymbol{U})$ is the variational posterior, which approximates the model's posterior $p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y})$, and $q(\boldsymbol{f}_i)$ is the marginal distribution of each row $\boldsymbol{f}_i$ of the latent function evaluation matrix $\boldsymbol{F}$.

4.3.3.2 Variational Distribution

The form of the variational posterior is again chosen to simplify the ELBO in eq. (89), starting by the KL divergence term. Since the model's prior distribution $p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}})$ factorizes as in eq. (78), it is possible to take advantage of the KL divergence factorization of eq. (53) by choosing variational distribution which also factorizes over the latent functions evaluations $\boldsymbol{f}^j$, $\boldsymbol{u}^j$, i.e.,

$$q(\boldsymbol{F},\boldsymbol{U}) = \prod_{l=1}^{L} q(\boldsymbol{f}^j,\boldsymbol{u}^j), \tag{91}$$

which gives the simplification

$$\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{F},\boldsymbol{U})\|p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}})] = \sum_{j=1}^{L} \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f}^j,\boldsymbol{u}^j)\|p(\boldsymbol{f}^j,\boldsymbol{u}^j|\boldsymbol{X},\boldsymbol{Z}^l)]. \tag{92}$$

Furthermore, imposing a factorization like the one from eq. (64) for each factor $q(\boldsymbol{f}^j, \boldsymbol{u}^j)$, i.e,

$$q(\boldsymbol{f}^j, \boldsymbol{u}^j) = p(\boldsymbol{f}^j | \boldsymbol{X}, \boldsymbol{u}^j, \boldsymbol{Z}^j) q(\boldsymbol{u}^j), \quad j = 1, \ldots, L, \tag{93}$$

permits a simplification similar to eq. (65), giving

$$\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{f}^j, \boldsymbol{u}^j) \| p(\boldsymbol{f}^j, \boldsymbol{u}^j | \boldsymbol{X}, \boldsymbol{Z}^l)] = \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{u}^j) \| p(, \boldsymbol{u}^j | \boldsymbol{Z}^l)] \tag{94}$$

which can be applied to eq. (92), resulting in

$$\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{F}, \boldsymbol{U}) \| p(\boldsymbol{F}, \boldsymbol{U} | \boldsymbol{X}, \tilde{\boldsymbol{Z}})] = \sum_{j=1}^{L} \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{u}^j) \| p(, \boldsymbol{u}^j | \boldsymbol{Z}^l)]. \tag{95}$$

Substituting eq. (95) into eq. (90) gives

$$\log p(\boldsymbol{y} | \boldsymbol{X}, \tilde{\boldsymbol{Z}}) \geq \left(\sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{f}_i)} [\log p(y_i | \boldsymbol{f}_i)] \right) - \left(\sum_{j=1}^{L} \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{u}^j) \| p(\boldsymbol{u}^j | \boldsymbol{Z}^l)] \right). \tag{96}$$

For mathematical tractability of eq. (96), the factors $q(\boldsymbol{u}^j)$ are chosen to be M^j-variate Gaussian distributions similarly to eq. (67), i.e.,

$$q(\boldsymbol{u}^j) = \mathcal{N}(\boldsymbol{u}^j | \boldsymbol{m}_{\boldsymbol{u}}^j, \boldsymbol{L}_{\boldsymbol{u}}^j \boldsymbol{L}_{\boldsymbol{u}}^{j^\top}), \quad j = 1, \ldots, L. \tag{97}$$

This choice enables eq. (95) to be computed as the sum of KL divergences between M^j-variate Gaussian distributions, $j = 1, \ldots, L$, using the expression of eq. (52). Also, similarly to eq. (70), the expression for the variational distributions $q(\boldsymbol{f}^j)$ are given by

$$\begin{aligned} q(\boldsymbol{f}^j) &= \mathcal{N}(\boldsymbol{f}^j | \boldsymbol{m}_{\boldsymbol{f}}^j, \boldsymbol{K}_{\boldsymbol{f}}^j), \quad j = 1, \ldots, L \\ \boldsymbol{m}_{\boldsymbol{f}}^j &= \boldsymbol{m}_{\boldsymbol{X}}^j + \boldsymbol{K}_{\boldsymbol{X}\mathbf{Z}^j}^j \boldsymbol{K}_{\mathbf{Z}^j\mathbf{Z}^j}^{j^{-1}} (\boldsymbol{m}_{\boldsymbol{u}}^j - \boldsymbol{m}_{\mathbf{Z}^j}^j) \\ \boldsymbol{K}_{\boldsymbol{f}}^j &= \boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}}^j + \boldsymbol{K}_{\boldsymbol{X}\mathbf{Z}^j}^j \boldsymbol{K}_{\mathbf{Z}^j\mathbf{Z}^j}^{j^{-1}} (\boldsymbol{L}_{\boldsymbol{u}}^j \boldsymbol{L}_{\boldsymbol{u}}^{j^\top} - \boldsymbol{K}_{\mathbf{Z}^j\mathbf{Z}^j}^j) \boldsymbol{K}_{\mathbf{Z}^j\mathbf{Z}^j}^{j^{-1}} \boldsymbol{K}_{\mathbf{Z}^j\boldsymbol{X}}^j. \end{aligned} \tag{98}$$

To compute expectation terms of eq. (96), one also needs the marginal distributions $q(\boldsymbol{f}^i)$. Combining eqs. (91) and (94) and marginalizing the inducing variables $\boldsymbol{u}^j$ gives

$$q(\boldsymbol{F}) = \prod_{j=1}^{L} q(\boldsymbol{f}^j). \tag{99}$$

Marginalizing all but the i-th row of the latent function evaluation matrix $\boldsymbol{F}$ from eq. (99) gives

$$q(\boldsymbol{f}_i) = \prod_{j=1}^{L} q(f_i^j), \tag{100}$$

where $q(f_i^j)$ is the distribution of the i-th element f_i^j of the latent function evaluation vector $\boldsymbol{f}^j$, which is given by

$$q(f_i^j) = \mathcal{N}(f_i^j | m_{f_i}^j, k_{f_{ii}}^j), \quad j = 1, \ldots, L \tag{101}$$

where the distribution's mean $m_{f_i}^j$ is the i-th element of the mean vector $\boldsymbol{m}_{\boldsymbol{f}}^j$, and distribution's variance $k_{f_{ii}}^j$ is the i-th element of the main diagonal of the covariance matrix $\boldsymbol{K}_{\boldsymbol{f}}^j$, respectively. As stated in section 4.3.2, the heteroscedastic Gaussian likelihood with the exponential as the positive-transforming function has an analytical expression for the expectations, given by

$$\begin{aligned}\mathbb{E}_{q(f_i,g_i)}\left[\log p(y_i|f_i,g_i)\right] = &-\frac{1}{2}\log(2\pi) - m_{g_i} \\ &-\frac{1}{2}\left[(y - m_{f_i})^2 + k_{f_{ii}}\right]\exp\left(2k_{g_{ii}} - 2m_{g_i}\right).\end{aligned} \tag{102}$$

4.3.4 Prediction

As discussed in section 4.1.5, the learned variational posterior $q(\boldsymbol{F},\boldsymbol{U})$ can be used to perform predictions by taking advantage of the equivalent of eq. (58) for the Chained GP, i.e,

$$q(\boldsymbol{F},\boldsymbol{U}) \approx p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}). \tag{103}$$

Consider, again, as in section 3.3.4, a set of new inputs grouped in the input matrix $\boldsymbol{X}'$ and their corresponding function evalution vectors $\boldsymbol{f}^{j'}$, grouped in the function evaluation matrix $\boldsymbol{F}'$. Like in eq. (74), the model's (variational) posterior predictive is given by

$$p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}) = \int p(\boldsymbol{F}'|\boldsymbol{X}',Chained\boldsymbol{F},\boldsymbol{X},\boldsymbol{U},\tilde{\boldsymbol{Z}})p(\boldsymbol{F},\boldsymbol{U}|\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y})\mathrm{d}\boldsymbol{F}\mathrm{d}\boldsymbol{U}. \tag{104}$$

Due to the independence between the GPs f^j, the term $p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{F},\boldsymbol{X},\boldsymbol{U},\tilde{\boldsymbol{Z}})$ from eq. (104) factorize as follows:

$$p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{F},\boldsymbol{X},\boldsymbol{U},\tilde{\boldsymbol{Z}}) = \prod_{j=1}^{L} p(\boldsymbol{f}^{j'}|\boldsymbol{X}',\boldsymbol{f}^j,\boldsymbol{X},\boldsymbol{u}^j,\boldsymbol{Z}^j). \tag{105}$$

By definition,

$$\mathrm{d}\boldsymbol{F} = \prod_{j=1}^{L}\mathrm{d}\boldsymbol{f}^j, \quad \mathrm{d}\boldsymbol{U} = \prod_{j=1}^{L}\mathrm{d}\boldsymbol{u}^j. \tag{106}$$

Substituting eqs. (105) and (106) and into eq. (104) and using the variational posterior eq. (91) to approximate the model's true posterior gives

$$p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}) \approx \int \prod_{j=1}^{L} p(\boldsymbol{f}^{j'}|\boldsymbol{X}',\boldsymbol{f}^j,\boldsymbol{X},\boldsymbol{u}^j,\boldsymbol{Z}^j)q(\boldsymbol{f}^j,\boldsymbol{u}^j)\mathrm{d}\boldsymbol{f}^j\mathrm{d}\boldsymbol{u}^j. \tag{107}$$

As the product terms inside the integrals of eq. (107) factorizes, it can be simplified as

$$p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}) \approx \prod_{j=1}^{L} \int p(\boldsymbol{f}^{j'}|\boldsymbol{X}',\boldsymbol{f}^{j},\boldsymbol{X},\boldsymbol{u}^{j},\boldsymbol{Z}^{j})q(\boldsymbol{f}^{j},\boldsymbol{u}^{j})\mathrm{d}\boldsymbol{f}^{j}\mathrm{d}\boldsymbol{u}^{j}. \tag{108}$$

Simplifying the integrals using the steps used to obtain eq. (74) gives

$$p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}) \approx \prod_{j=1}^{L} q(\boldsymbol{f}^{j'}) = q(\boldsymbol{F}'), \tag{109}$$

which means that the variational posterior predictive $q(\boldsymbol{F}')$ has the same form of the variational distribution $q(\boldsymbol{F})$, only swapping $\boldsymbol{X}$ for $\boldsymbol{X}'$.

Similarly to eq. (75), the prediction of the new noisy observations $\boldsymbol{y}'$ can be approximated as

$$\begin{aligned} p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}) &= \int p(\boldsymbol{y}'|\boldsymbol{F}')p(\boldsymbol{F}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y})\mathrm{d}\boldsymbol{F}' \\ &\approx \int p(\boldsymbol{y}'|\boldsymbol{F}')q(\boldsymbol{F}')\mathrm{d}\boldsymbol{F}'. \end{aligned} \tag{110}$$

For a general likelihood, eq. (110) requires numerical integration. In the heteroscedastic Gaussian likelihood case, part of the integral can be evaluated analytically. Substituting eq. (84) into eq. (79) for the new noisy observations $\boldsymbol{y}'$ gives

$$p(\boldsymbol{y}'|\boldsymbol{f}',\boldsymbol{g}') = \prod_{i=1}^{N'} \mathcal{N}(y_i'|f_i',\exp(g_i')^2) = \mathcal{N}(\boldsymbol{y}'|\boldsymbol{f}',\boldsymbol{S}'), \tag{111}$$

where $\boldsymbol{S}'$ is a diagonal matrix whose i-th diagonal element is given by $s_{ii}' = \exp(g_i')^2$. Substituting eqs. (98) and (111) into eq. (109) gives

$$\begin{aligned} p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y}) &= \int \mathcal{N}(\boldsymbol{y}'|\boldsymbol{f}',\boldsymbol{S}')\mathcal{N}(\boldsymbol{f}'|\boldsymbol{m}_{\boldsymbol{f}'},\boldsymbol{K}_{\boldsymbol{f}'})q(\boldsymbol{g}')\mathrm{d}\boldsymbol{f}'\mathrm{d}\boldsymbol{g}' \\ &= \int \mathcal{N}(\boldsymbol{y}'|\boldsymbol{m}_{\boldsymbol{f}'},\boldsymbol{K}_{\boldsymbol{f}'}+\boldsymbol{S}')q(\boldsymbol{g}')\mathrm{d}\boldsymbol{g}'. \end{aligned} \tag{112}$$

Even though eq. (112) does not give an analytical expression, it reduces the numerical integration problem to only one latent function vector $\boldsymbol{g}'$, as $\boldsymbol{f}'$ was integrated analytically. This shows the computational benefits of the heteroscedastic Gaussian likelihood. Furthermore, the mean and variance of the predictive distribution of the new observations $\boldsymbol{y}'$ are available in closed form as

$$\mathbb{E}[\boldsymbol{y}'] = \boldsymbol{m}_{\boldsymbol{f}'}, \tag{113}$$

$$\mathbb{V}[\boldsymbol{y}'] = \boldsymbol{K}_{\boldsymbol{f}'} + \boldsymbol{C}', \tag{114}$$

where $\boldsymbol{C}'$ is a diagonal matrix whose i-th diagonal element is given by $c_{ii}' = \exp(2m_{g_i'} + 2k_{g_{ii}'})$, with $m_{g_i'}$ being the i-th element of the vector $\boldsymbol{m}_{\boldsymbol{g}'}$ and $k_{g_{ii}'}$ the i-th element of the main diagonal of the covariance matrix $\boldsymbol{K}_{\boldsymbol{g}'}$, respectively.

4.4 Discussion

In this chapter, variational inference was presented in the context of GP models as an option to avoid the intractable expressions needed for Bayesian inference, as described in section 3.3.3. After a review of the KL divergence definition and properties, the ELBO was derived for the standard GP regression model from section 3.3.1 with a generic variational distribution $q(\boldsymbol{f})$. The ELBO was proposed as an optimization objective by optimizing the model parameters and hyperparameters togheter with the variatioal parameters in a generic variational inference setting. Finally, it was shown that the variational posterior $q(\boldsymbol{f})$ is a good approximation (in a KL divergence sense) of the true model's posterior $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})$.

The SVGP was introduced as a way to deal with the large complexity and memory requirements of the standard GP regression. It was built by augmenting the standard GP regression model from section 3.3.1 with M pseudo-inputs $\boldsymbol{Z} \in \mathscr{X}^M$ and their corresponding inducing variables $\boldsymbol{u} \in \mathbb{R}^M$, which formed input-output pairs of the latent GP f, i.e., $\boldsymbol{u} = f(\boldsymbol{Z})$. The variational inference process was carried on using a factorized variational distribution $q(\boldsymbol{f},\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{u})q(\boldsymbol{u})$, with $q(\boldsymbol{u}) = \mathscr{N}(\boldsymbol{u}|\boldsymbol{m}_{\boldsymbol{u}},\boldsymbol{L}_{\boldsymbol{u}}\boldsymbol{L}_{\boldsymbol{u}}^{\top})$, resulting in a tractable ELBO with much smaller complexity and memory requirements than the traditional GP regression from section 3.3.3. The predictions $p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y})$ were performed with the variational posterior predictive $q(\boldsymbol{f}')$, which was obtained as a by-product of the variational inference process.

Finally, the Chained GP was discussed as an extension of the GP regression model from section 3.3.1 which enabled the modeling of more noisy observations' features by employing multiple latent GPs f^j, $j = 1,\ldots,L$. After the introduction of three multi-latent likelihoods built by extending the likelihoods from section 3.3.2, the SVGP augmentation was applied to each latent GP f^j and variational inference was once again used to derive a tractable ELBO, which enjoyed the scalability benefits of the SVGP. Predictions $p(\boldsymbol{y}'|\boldsymbol{X}',\boldsymbol{X},\tilde{\boldsymbol{Z}},\boldsymbol{y})$ for the Chained GP were also performed with the variational posterior predictive $q(\boldsymbol{F}')$.

5 PROPOSED MODELING FRAMEWORK

The previous chapters 3 and 4 have presented multiple theoretical properties of Chained GP models such as incorporation of prior knowledge, adaptability to the data, representation of heteroscedastic noise and (local) robustness to outliers, which motivates their usage in the WTPC modeling problem analyzed in chapter 2. However, there is an already large pool of modeling options available in the literature, and every new candidate must be compared to the state-of-art models by evaluating how they perform when applied to real WT operational data.

In this chapter, the modeling framework proposed by this dissertation and benchmark models are evaluated on computational experiments where all models are fitted in multiple scenarios related to a rich 1-year of WT operational data. In section 5.1, the proposed modeling framework is described in depth, covering both the rationale behind its construction and its computational implementation details. In section 5.2, the benchmark models are formalized and their computational implementation is presented. In section 5.3, the dataset which will be used to compare the models is explored emphasizing the multiple features present in it which are of interest regarding the WTPC modeling problem. In section 5.4, the experimental scenarios are described and the obtained results are discussed showing how the proposed models compare to the benchmark ones. The chapter is finished with a summarization of the discussed subject in section 5.5

5.1 Modeling Framework Description

Given N observations of wind speed and normalized power (v_i, p_i), the WTPC modeling can be analyzed as a GP regression problem (see section 3.3) with the wind speed as the input, $\boldsymbol{X} = [v_i]_{N\times 1} \in \mathbb{R}$, with input dimensions $D = 1$, and the normalized power as the output, $\boldsymbol{y} = [p_i]_{N\times 1}$. Aiming to model the heteroscedastic behavior of the noise and also to be robust to outliers, the Chained GP framework is chosen due to the possibility of using the more complex likelihoods from section 4.3.2.

5.1.1 Model Construction Rationale

As discussed in sections 2.1 and 2.4, the logistic function can represent many properties of a typical WTPC. This information can be assimilated in the GP regression by setting the location hyperparameter μ_y of the likelihood as a latent GP f with the L3P logistic mean function from eq. (15). To complete the definition of the latent GP f the SE covariance function from eq. (29) was chosen to be its covariance function, which gives

$$\begin{aligned}
\mu_y(\boldsymbol{x}) &= f(\boldsymbol{x}) \\
f &\sim \mathscr{GP}(\mu_f, \kappa_f) \\
\mu_f(\boldsymbol{x}) &= \left[1 + \exp\left(-\left(\frac{\nu - \nu_0}{s}\right)\right)\right]^{-1/\gamma} \\
\kappa_f(\boldsymbol{x}, \boldsymbol{x}') &= \sigma_f^2 \exp\left[-\frac{1}{2}\left(\frac{\nu - \nu'}{l_f}\right)^2\right],
\end{aligned} \tag{115}$$

where $\boldsymbol{x} = [\nu]$ and $\boldsymbol{x}' = [\nu']$.

The likelihoods from section 4.3.2 also need a latent GP g to model their scale hyperparameter σ_y. Since there's no strong prior knowledge about them, the constant function was chosen as its mean functions, and the SE covariance function from eq. (29) as its covariance functions, which gives

$$\begin{aligned}
\sigma_y(\boldsymbol{x}) &= \exp(g(\boldsymbol{x})) \\
g &\sim \mathscr{GP}(\mu_g, \kappa_g) \\
\mu_g(\boldsymbol{x}) &\equiv c_g \\
\kappa_g(\boldsymbol{x}, \boldsymbol{x}') &= \sigma_g^2 \exp\left[-\frac{1}{2}\left(\frac{\nu - \nu'}{l_g}\right)^2\right].
\end{aligned} \tag{116}$$

The exponential function was chosen as the positive transforming function due to its useful analytical results for the heteroscedastic Gaussian likelihood. It is also a valid and easy-to-implement option for the other two likelihoods.

Specifically for the locally robust heteroscedastic Student-t likelihood, the degrees-of-freedom hyperparameter ν is modeled by another latent GP h, which was built similarly to the latent GP g as there's no prior information to incorporate on it, which gives

$$\begin{aligned}
\nu(\boldsymbol{x}) &= 3 + \exp(h(\boldsymbol{x})) \\
h &\sim \mathscr{GP}(\mu_h, \kappa_h) \\
\mu_h(\boldsymbol{x}) &\equiv c_h \\
\kappa_h(\boldsymbol{x}, \boldsymbol{x}') &= \sigma_h^2 \exp\left[-\frac{1}{2}\left(\frac{\nu - \nu'}{l_h}\right)^2\right].
\end{aligned} \tag{117}$$

The warping function $t'(a) = 3 + \exp(a)$ was chosen to make sure that the degrees of freedom hyperparameter ν satisfies $\nu > 3$. This constraint is also enforced for the heteroscedastic Student-t likelihood.

The model follows the assumption of factorizing likelihoods, as in eq. (31), but given all latent GPs, i.e.,

$$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{g},\boldsymbol{h}) = \prod_{i=1}^{N} p(y_i|f_i,g_i,h_i), \tag{118}$$

where $y_i = p_i$.

Inspired by the method of bins from section 2.2.3, The pseudo-inputs $\boldsymbol{Z}^f$, $\boldsymbol{Z}^g$, $\boldsymbol{Z}^h$ are chosen to be evenly-spaced from 0 m/s to 25 m/s with increment $\Delta v = 0.5$ m/s, resulting in $M = 51$ pseudo-inputs for each latent GP. It is worth noting the presence of the variational parameters $(\boldsymbol{m}^f_{\boldsymbol{u}}, \boldsymbol{L}^f_{\boldsymbol{u}})$, $(\boldsymbol{m}^g_{\boldsymbol{u}}, \boldsymbol{L}^g_{\boldsymbol{u}})$ and $(\boldsymbol{m}^h_{\boldsymbol{u}}, \boldsymbol{L}^h_{\boldsymbol{u}})$.

The construction outlined in this section defined three models, namely:

- **L3P-HG-GP**, using the heteroscedastic Gaussian likelihood;
- **L3P-HS-GP**, using the heteroscedastic Student-t likelihood;
- **L3P-LRHS-GP**, using the locally robust Student-t likelihood.

5.1.2 Note on Model Implementation

The models were implemented in GPflow (MATTHEWS *et al.*, 2017), using the `models.SVGP` model class and `kernels.SeparateIndependent` kernel class. The heteroscedastic Gaussian likelihood had its analytic expressions implemented as custom code by the author, while the (locally robust) heteroscedastic Student-t likelihoods were implement with Gauss-Hermite quadratures and Tensorflow Probability's `distributions.StudentT` class.

During this dissertation development, the author has actively contributed with the GPflow community to develop multi-latent likelihoods and refactored the implementation of Gauss-Hermite quadrature, which culminated in the `quadraute.NDiagGHQuad` and `likelihoods.Quadrat` classes. The code has been fully integrated in the GPflow package and has been used to implement the proposed models.

It is worth noting that GPflow utilizes the whitened representation of the inducing variables u, which is exemplified below for the latent GP f:

$$\boldsymbol{u}^f = \boldsymbol{m}^f_{\boldsymbol{Z}^f} + \mathrm{chol}(\boldsymbol{K}^f_{\boldsymbol{Z}^f\boldsymbol{Z}^f})\boldsymbol{v}^f, \quad \boldsymbol{v}^f \sim \mathcal{N}(\boldsymbol{m}^f_{\boldsymbol{v}}, \boldsymbol{L}^f_{\boldsymbol{v}}), \tag{119}$$

where $\boldsymbol{L} = \mathrm{chol}(\boldsymbol{A})$ is the Cholesky factor of matrix $\boldsymbol{A}$, i.e., $\boldsymbol{L}$ is a lower-triangular matrix which satisfies $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{A}$. This representation effectively reparametrizes the variational parameters $(\boldsymbol{m}^f_{\boldsymbol{u}}, \boldsymbol{L}^f_{\boldsymbol{u}})$ in terms of the *withened* variational parameters $(\boldsymbol{m}^f_{\boldsymbol{v}}, \boldsymbol{L}^f_{\boldsymbol{v}})$ through the following transformation equation:

$$\begin{aligned} &\boldsymbol{m}_{\boldsymbol{u}}^{f} = \boldsymbol{m}_{\mathbf{Z}^f}^{f} + \boldsymbol{m}_{\boldsymbol{v}}^{f} \\ &\boldsymbol{L}_{\boldsymbol{u}}^{f} = \mathrm{chol}(\boldsymbol{K}_{\mathbf{Z}^f\mathbf{Z}^f}^{f})\boldsymbol{L}_{\boldsymbol{v}}^{f}. \end{aligned} \tag{120}$$

The same is true for latent GPs g and h.

5.1.3 Parameters Initialization

To initialize the L3P mean function parameters, the coefficients $a = v_0/s$ and $b = -s^{-1}$ from the linearized L2P model, eq. (17), were estimated with OLS, which enabled the initialization $(x_0, s, \gamma) = (-ab^{-1}, -b^{-1}, 1)$. The OLS used a subset of the data satisfying $0.05 < p < .95$ and was implemented with Scikit-Learn's (PEDREGOSA *et al.*, 2011) `linear_model.LinearRegre` class. The constant mean functions were initialized with their constants set to zero, i.e., $c_g = c_h = 0$.

All covariance functions were initialized with unitary length scale and variance, i.e., $l_f = l_g = l_h = 1$ and $\sigma_f = \sigma_g = \sigma_h = 1$. Those are GPflow's defaults, and they work well when the data does not have extreme values, which is the case as the wind speed is mostly bounded between 0 m/s and 20 m/s; and the normalized power, between 0 an 1.

Specifically for the heteroscedastic Student-t Likelihood, the degrees of freedom hyperparameter was initialized as $\nu = 4$, satisfying $\nu > 3$.

The whitened variational parameters are initialized to zero mean and identity covariance, i.e., $\boldsymbol{m}_{\boldsymbol{v}}^{f} = \boldsymbol{m}_{\boldsymbol{v}}^{g} = \boldsymbol{m}_{\boldsymbol{v}}^{h} = \mathbf{0}_{M\times 1}$ and $\boldsymbol{L}_{\boldsymbol{v}}^{f} = \boldsymbol{L}_{\boldsymbol{v}}^{g} = \boldsymbol{L}_{\boldsymbol{v}}^{h} = \boldsymbol{I}_{M}$. Those are GPflow's defaults, and they correspond to the variational posterior initialized as the prior, which does not add any new knowledge to the model, and hence is a good starting point.

5.1.4 Optimization

For model fitting, the factorized Chained GP ELBO, eq. (96), divided by the number of observations N, was optimized for 500 iterations of the L-BFGS optimization algorithm (NOCEDAL; WRIGHT, 2006), using the implementation from Tensorflow Probability's `optimizers.lbfgs` function.

5.2 Benchmark Models

5.2.1 Method of Bins (MoB)

The method of bins is included as a WTPC modeling benchmark due to its technical importance. It was initially presented in section 2.2.3, and were implemented as described there, but using the normalized power p instead of the raw power measures of P for easiness of comparison with the other models.

5.2.2 Polynomial Regression (Poly-9)

Polynomial regression is one of the main WTPC modeling benchmarks (LI *et al.*, 2001; SHOKRZADEH *et al.*, 2014; GUO; INFIELD, 2018; YAN *et al.*, 2019). In the upcoming experiments, a 9-th degree polynomial relating the wind speed v to the normalized power p is used, i.e.,

$$p(v) = \sum_{i=1}^{9} \alpha_i v^i = \alpha_0 + \alpha_1 v + \alpha_2 v^2 + \cdots + \alpha_8 v^8 + \alpha_9 v^9. \tag{121}$$

The parameter vector $\boldsymbol{\phi} = [\alpha_0,\ \alpha_1,\ \ldots,\ \alpha_8, \alpha_9]^T$ is estimated with the OLS method. For this purpose, the Moore-Penrose pseudoinverse matrix (GOLUB; Van Loan, 2012) is constructed via singular value decomposition, treating singular values that are smaller than a given tolerance as zero. The implementation from Scikit-Learn's (PEDREGOSA *et al.*, 2011) `LinearRegression` class was used.

5.2.3 Neural Networks - MLP(1,12,1)

Neural networks are the other main WTPC modeling benchmarks (LI *et al.*, 2001; LYDIA *et al.*, 2013; MANOBEL *et al.*, 2018; BAI *et al.*, 2019; YAN *et al.*, 2019). In the upcoming experiments, A multilayer perceptron (MLP) network with the wind speed v as the input, one hidden layer with 12 hidden neurons with hyperbolic tangent activation function, and a output layer with linear activation function outputing the normalized power p is used. As usual, input data to the MLP network was transformed to zero mean and unit variance.

The model was implemented with Tensorflow's (ABADI *et al.*, 2016) `keras` module, and the weights and bias were initialized with the Xavier uniform initializer (GLOROT; BENGIO, 2010) and zeroes, respectively, which are their defaults in Tensorflow. The root mean squared error (RMSE) was used as the loss function, and the fitting was by optimizing it for 500 iterations of the BFGS optimization algorithm (NOCEDAL; WRIGHT, 2006), using the implementation from Scipy's (JONES *et al.*, 2001–) `minimize` function indirectly by employing GPflow's (MATTHEWS *et al.*, 2017) `ScipyOptimizer` wrapper class.

5.2.4 Logistic Function (L3P)

Logistic functions are very import WTPC models due to their capabilities of reflecting the multiple operational ranges of a typical WTPC, as discussed in sections 2.1 and 2.4, and are one of the main constituting parts of the models proposed by this dissertation. As such, they are included as benchmark models using the L3P model, eq. (15).

The model was implemented with Numpy (OLIPHANT, 2006–) functions. The parameters x_0 and s are initialized with the OLS estimates obtained from eq. (17) using a subset of the data satisfying $0.05 < p < .95$, which was implemented with Scikit-Learn's (PEDREGOSA *et al.*, 2011) `LinearRegression` class, and the paramter γ was initialized as 1. Then, the three parameters were optimized to minimize the RMSE loss function for 500 iterations of the BFGS optimization algorithm (NOCEDAL; WRIGHT, 2006) implemented in Scipy's (JONES *et al.*, 2001–) `minimize` function.

5.2.5 Zero Mean GP (0-GP)

The GP with zero mean function is the most basic form of GP model that has been applied to WTPC modeling (PANDIT; INFIELD, 2019; PANDIT *et al.*, 2019), and is the other main constituting part of the models proposed in this dissertation. As such, it is considered as a benchmark model.

It is represented in the upcoming experiments by a SVGP (see section 4.2) with the wind speed as the input, $\boldsymbol{x} = [v]$, the normalized power as the output, $y = p$, and a Gaussian likelihood as in eq. (34). The mean and covariance functions were set as the constant zero and the squared exponential covariance functions, respectively, i.e., $\mu \equiv 0$ and $\kappa = \kappa_{\mathrm{SE}}$ - see eq. (29). Inspired by the MoB from section 2.2.3, The pseudo-inputs $\boldsymbol{Z} = [z_i]_{M\times 1}$ were evenly-spaced from 0 m/s to 25 m/s with increment $\Delta v = 0.5$ m/s, resulting in $M = 51$ pseudo-inputs.

The model was implemented in GPflow (MATTHEWS *et al.*, 2017). For initialization, the covariance function length scale and variance hyperpameters were set as $l_f = 1$ and $\sigma_f^2 = 1$ respectively; the likelihood variance hyperparameter, as $\sigma_y^2 = 1$; and the whitened variational parameters, eq. (120), as $\boldsymbol{m_v} = \boldsymbol{0}_{M\times 1}$ and $\boldsymbol{L_v} = \boldsymbol{I}_M$.

For model fitting, the factorized SVGP ELBO, given in eq. (63), divided by the number of observations N, was optimized for 500 iterations of the L-BFGS optimization algorithm (NOCEDAL; WRIGHT, 2006), using the implementation from Tensorflow Probability's `optimizers.lbfgs_minimize` function.

5.3 Dataset Description

The WTPC modeling problem has multiple features such as seasonal variability, heteroscedastic noise and the presence of outliers. As such, having an dataset capable of expressing all those peculiarities is of great importance for WTPC model comparison. In this section, a dataset of 1-year of operational data of a WT, kindly provided by Delfos IM (DELFOS INTELLIGENT MAINTENANCE, 2017–), is described, highlighting the features which will be used to analyze the WTPC models in the upcoming experiments.

5.3.1 Wind Turbine Operational Parameters

The studied WT has the following operational parameters:

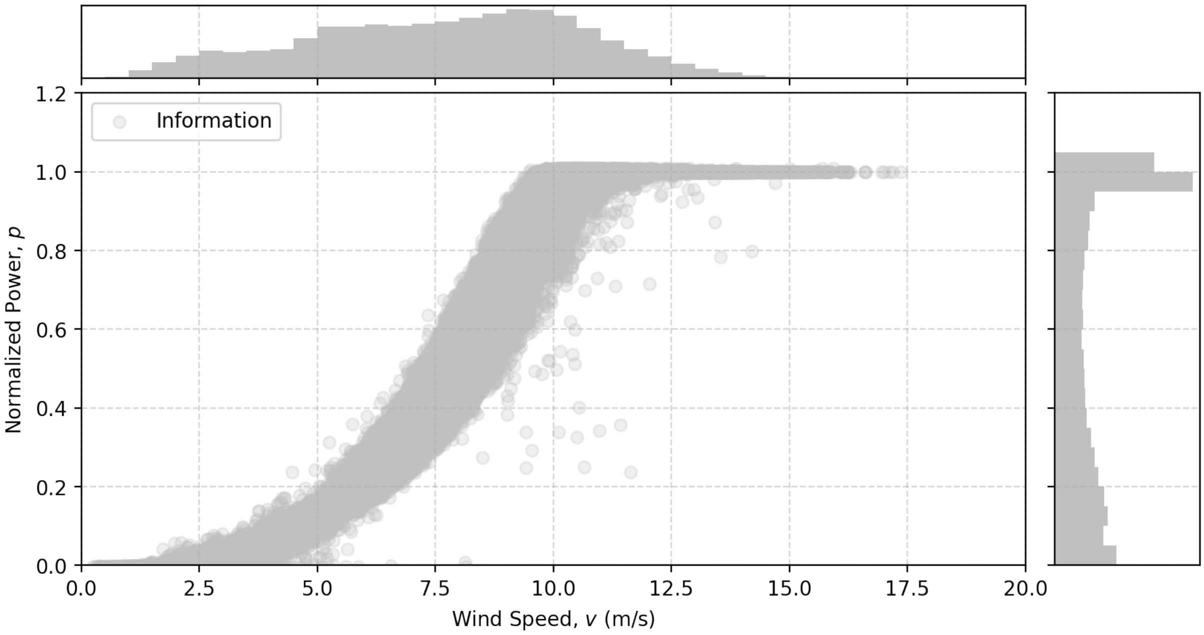
- Rated power: $P_{\text{rated}} = 2100$ kW;
- Cut-in wind speed: $v_{\text{ci}} = 3.0$ m/s;
- Rated wind speed: $v_{\text{rated}} = 11.0$ m/s;
- Cut-out wind speed: $v_{\text{co}} = 25.0$ m/s.

5.3.2 Data Sources

The dataset was built by following the recommendations of the IEC standard, as described in section 2.2.1. The wind speed and power data (v_i, P_i) was obtained directly from the WT SCADA system, and complementary measurements of ambient temperature, pressure and relative humidity (T_i, B_i, ϕ_i) were obtained from the meteorological mast closest to the studied WT. All data consists in 10-minute averages, for a total of $N_0 = 51994$ valid observations in the analyzed period of 1 year.

The air density normalization procedure from section 2.2.2 was applied to the wind speed v, with the reference air density set to $\rho_{\text{ref}} = 1.06$ kg/m^2, which is the usual set-up for WTPC analysis in this specific wind farm. Also, eq. (2) was used to obtain the normalized power p, resulting in observations of wind speed and normalized power (v_i, p_i), which are the main focus of the upcoming experiments.

Figure 4 – Dataset **A**, built using the event log to filter abnormal states. The effects of heteroscedasticity are clearly visible.



Source - The author.

In addition to the 10-minute data, the event log of the WT was provided. By crossing it with the data, each observation was classified accordingly to the most severe active alarm category during the its 10-minute interval. The alarm severity classes are detailed below, in ascending severity order.

1. **Information**: alarms which provide information about the WT current operational state;
2. **Warning**: alarms with warnings about some subsystem in the WT;
3. **Power Limitation**: alarms which cause the WT to not be able to produce its maximum power;
4. **Stop**: alarms which cause the WT to cease operation.

5.3.3 Data Cleaning

The event log brings very useful information about the operational state of the WT. Noting that observations with alarm severity greater than “Information” represent abnormal operational conditions which are not suitable for WTPC modeling, the dataset **A** was built by filtering them out. The resulting dataset ended up with $N_A = 47387$ observations and is shown in fig. 4.

However, it is important to acknowledge that event logs are not always available when modeling the WTPC. As such, the dataset **B** was built by considering a data-based filter

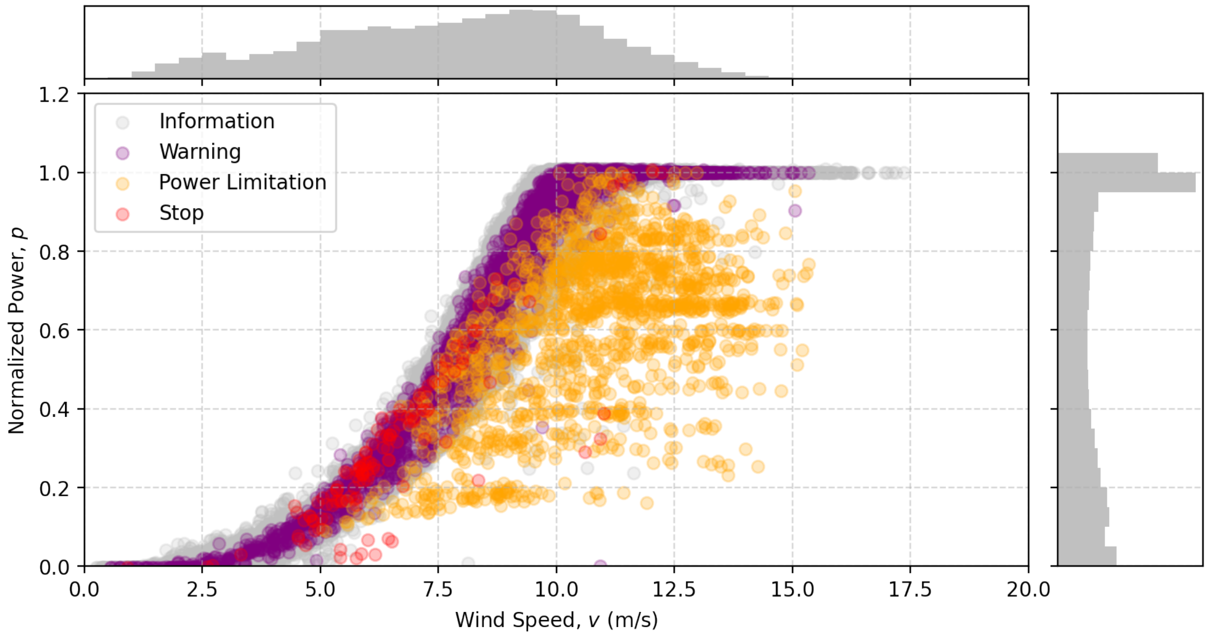


Figure 5 – Dataset **B**, built using the data-based filter. The presence of “Power Limitation” alarms generates many outliers.

by excluding observations (v_i, p_i) with normalized power less than zero for wind speeds above the cut in wind speed, i.e, $p_i < 0$ and $v_i > v_{\mathrm{ci}}$. The resulting dataset ended up with $N_B = 51465$ observations and is shown in fig. 5.

As can be seen in figs. 4 and 5, the dataset **A** shows more clearly the heteroscedasticity effect, while the dataset **B** has many more outliers, which are generated mainly because of “Power Limitation”. As such, the dataset **B** will be used for experiments concerning outliers and **A** for the other ones. It is important to note that the heteroscedasticity is present in both datasets, but it is easier to graphically analyze it in dataset **A**.

5.3.4 Data Seasonality

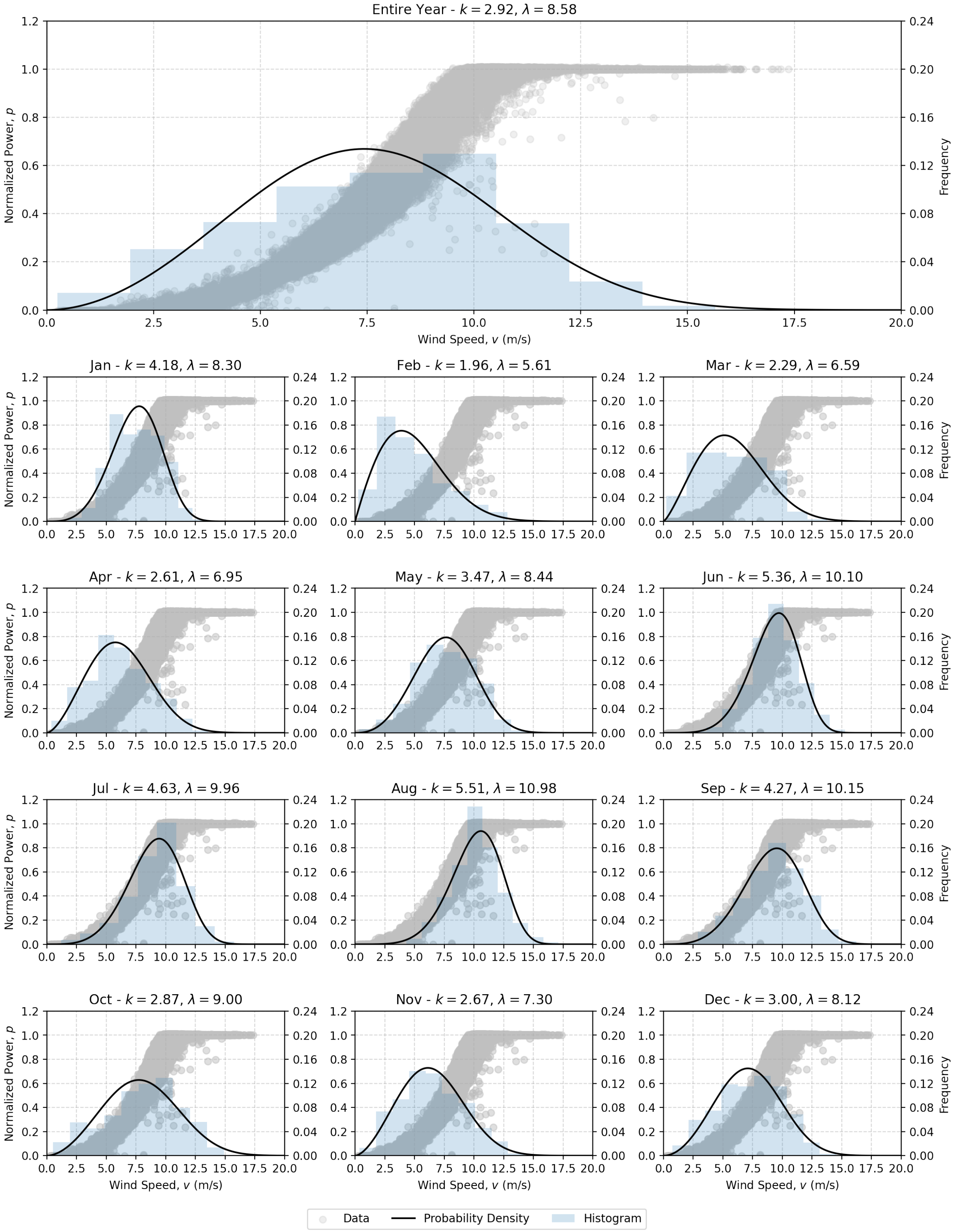
In WTPC modeling, the wind speed is usually modeled as a Weibull distribution, $v \sim \mathscr{W}(\lambda, k)$, whose probability density is given by

$$p(v) = \mathscr{W}(v|\lambda, k) = \frac{k}{\lambda}\left(\frac{v}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{v}{\lambda}\right)^{k}\right], \quad v \geq 0, \tag{122}$$

where $k > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively.

The dataset **A** was chosen for a in-depth seasonality analysis as it is, in a data visualization sense, cleaner than dataset **B**. To investigate how the wind speed v distribution varies through the year, a Weibull distribution was fitted for the entire year and also each month individually using the Scipy’s (JONES *et al.*, 2001–) `stats.weibull_min.fit` function.

Figure 6 – Monthly seasonality of the wind speed distribution for dataset **A**. The black lines evidentiate how the fitted Weibull distribution, whose parameters are annotated on top of each plot, varies for each month.



Source - The author

As can be seen in the resulting fig. 6, the wind speed distribution significantly changes through the year, which is evidentiated by the changes on both distribution parameters values and also by the graphical changes on the plotted histograms and the probability densities. This seasonal behavior affects WTPC modeling, as it is not desirable to always have to wait the acquisition of a full year of data before fitting a model. As such, this feature will be explored in seasonality experiments, comparing how models fitted for a given month perform in the others.

5.4 Experiments and Results

As pointed out in section 2.3, there are multiple WTPC modeling approaches available in the literature, which begs the question: how does the proposed modeling framework from section 5.1 compare to benchmark models such as the ones described in section 5.2? In this section, those models are evaluated in four experiments concerning their ability to fit the data and the main features of the datasets **A** and **B** described in section 5.3, namely, *heteroscedasticity*, *robustness to outliers* and *seasonality*. The results are analyzed in terms of quantitative comparison criteria and also qualitatively through graphical inspection.

5.4.1 Comparison Criteria

Two main criteria are considered to analyze the results of the experiments: the root mean squared error (RMSE), which can be applied to any model, and the mean negative log predictive density (MNLPD), which is applicable to probabilistic models only.

5.4.1.1 Root Mean-Squared Error (RMSE)

The RMSE is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{123}$$

where N is the number of observations, y_i is the value of the i-th output and $\hat{y}_i$ is the predicted value by the model for the i-th input $\boldsymbol{x}_i$. For deterministic models, the predicted value $\hat{y}_i$ is given by $\hat{y}_i = \mathrm{model}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\mathrm{model}})$, whereas for probabilistic ones, the expected value of the predictive distribution is used.

The RMSE is widely used in the WTPC modeling literature and is a easy-to-compute metric which measures how far the predictions $\hat{y}_i$ deviate from the observed outputs y_i, placing more emphasis on extreme deviations due to the quadratic exponent. As such, it is considered as the main comparison criteria.

5.4.1.2 Mean Negative Log Predictive Density (MNLPD)

Despite its widespread adoption, the RMSE is not able to fully evaluate probabilistic models as it focuses only on the expected value of its predictions. As such, it is not capable of analyzing if a model is able to properly represent the noise behavior of the predictions, and hence cannot be used to analyze features of the noise distribution such as heteroscedasticity. Aiming to quantify this very important aspect of WTPC modeling, the MNLPD comparison criteria is introduced, which is given by

$$\mathrm{MNLPD} = \frac{1}{N}\sum_{i=1}^{N} -\log p(y_i'|\boldsymbol{x}_i', \boldsymbol{X}, \boldsymbol{y}) \tag{124}$$

where N is the number of observations and $p(y_i'|\boldsymbol{x}_i', \boldsymbol{X}, \boldsymbol{y})$ is the posterior predictive density of the outputs y_i' given the inputs $\boldsymbol{x}_i'$ for a model fitted using the input-output data $(\boldsymbol{X}, \boldsymbol{y})$.

5.4.1.3 Fitting and Evaluating

In the upcoming experiments, the models will be fitted to a dataset and then compared by evaluating a proper comparison criteria on another dataset, which is not necessarily the same used for model fitting. To clearly distinguish them, following notation is used:

CRITERIA (E|F),

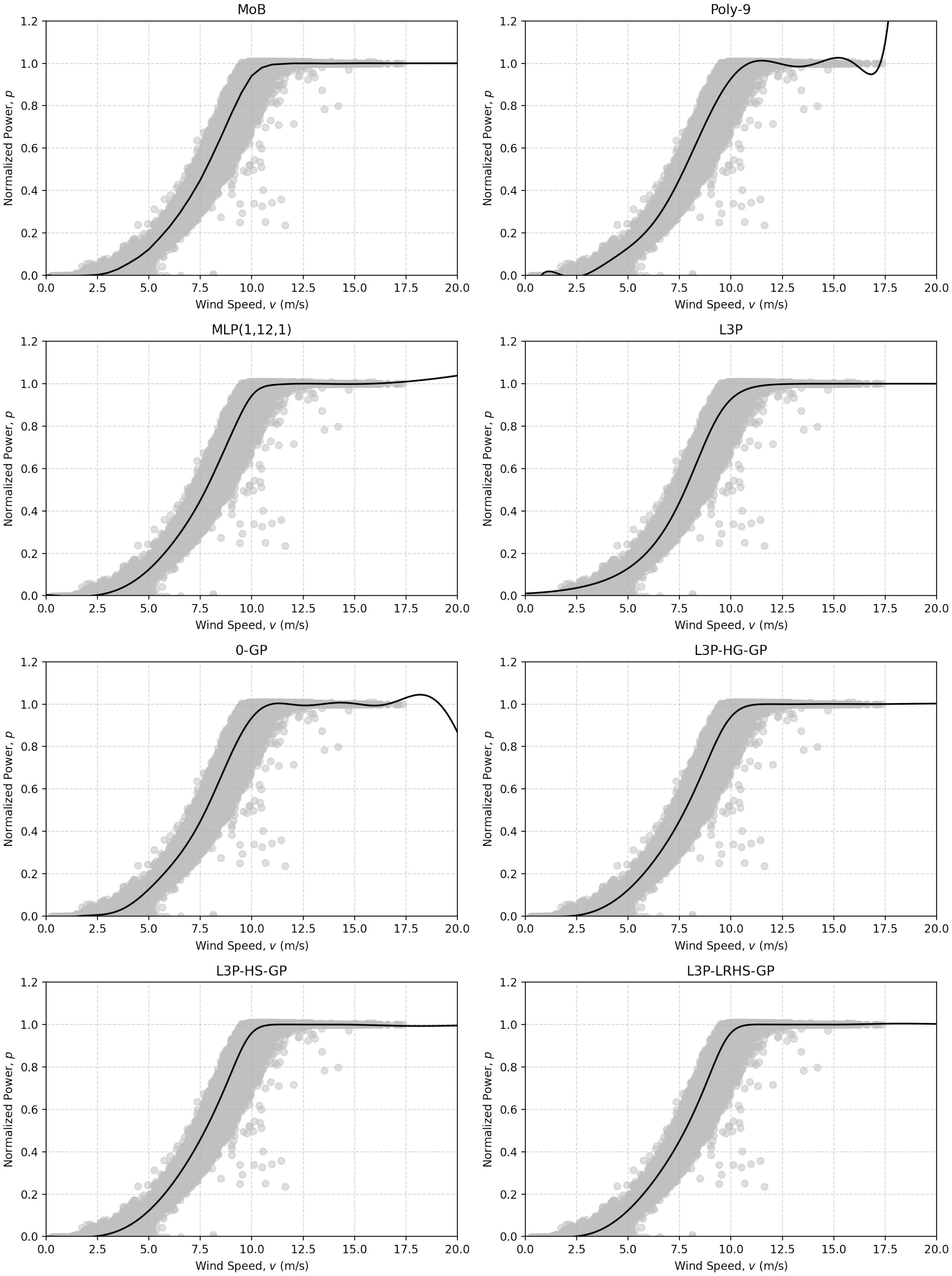
where **CRITERIA** is either the RMSE or the MNLPD, **E** is the dataset on which the criteria is evaluated and **F** is the dataset used to fit the model currently analyzed models.

5.4.2 Data-Fitting Capabilities

In this experiment, an initial comparison between the WTPC models regarding the ability to fit the available data is analyzed. All models from sections 5.1 and 5.2 were fitted for dataset **A** from section 5.3 and compared in terms of their RMSE evaluated on dataset **A**. The results are shown in fig. 7 and in table 1.

As can be seen in table 1, the proposed models (L3P-HG-GP, L3P-HS-GP and L3P-LRHS-GP) present competitive RMSE scores. All of them were better than the L3P and Poly-9 benchmark models, and the L3P-HG-GP was very close to the best result, which was

Figure 7 – Predicted values for all models fitted to dataset **A**.



Source - The author

Table 1 – RMSE of dataset **A** for all models fitted to dataset **A**.

Model	**RMSE** (A|A)
MoB	4.9305×10^{-2}
Poly-9	5.0119×10^{-2}
MLP(1,12,1)	4.9265×10^{-2}
L3P	5.2048×10^{-2}
0-GP	4.9440×10^{-2}
L3P-HG-GP	4.9295×10^{-2}
L3P-HS-GP	5.0013×10^{-2}
L3P-LRHS-GP	4.9959×10^{-2}

obtained by the MLP(1,12,1). However, it is important to remember that the proposed models are probabilistic ones and do not aim for the best RMSE exclusively, but to represent the complete noise behavior of the observations. Also, the proposed models are more complex and hence can be harder to optimize: it is possible that 500 iterations of the L-BFGS algorithm are not enough to reach their optima, an investigation which is beyond the scope of this experiment.

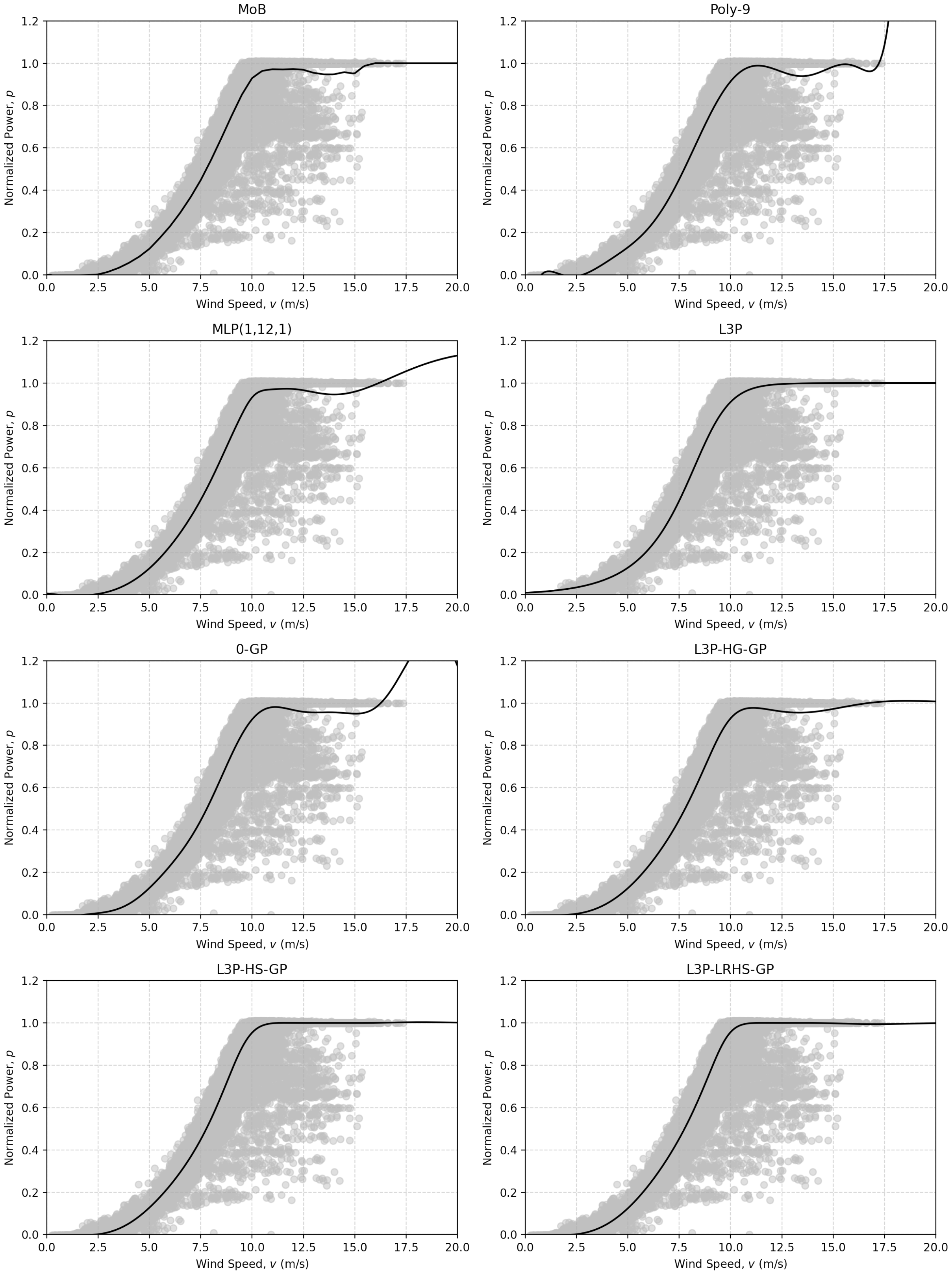
Furthermore, analyzing fig. 7, it is clear that the the models Poly-9, MLP(1,12,1) and 0-GP have conceptual problems as they fail to represent the usual WTPC behavior for high wind speeds (see fig. 3), whereas the MoB and all the L3P-related follow the expected shape. As such, the proposed models combine the theoretical benefits of the L3P model with the adaptability of GPs to produce models which can fit the data well while not deviating from the WTPC expected shape.

5.4.3 Robustness to Outliers

In this experiment, the task of obtaining a WTPC model which represents the normal operation of a WT without access to operational state labels such as the one obtained through the event log is analyzed. To simulate this scenario, all models from sections 5.1 and 5.2 were fitted to the dataset **B**, which contains more outliers, as depicted in fig. 5, and then compared in terms of the RMSE evaluated on dataset **A**, which has way less outliers (see fig. 4). The results are shown in fig. 8 and in table 2.

The results in table 2 shows that the two of the proposed models, namely, L3P-HS-GP and L3P-LRHS-GP, have the best RMSE on dataset **A**, which is the objective of this experiment. Comparing it to table 1 evidentiate how impactful the presence of outliers can be to the WTPC modeling problem, which highlights the importance of using robust models such as the proposed ones.

Figure 8 – Predicted values for all models fitted to dataset **B**.



Source - The author

Table 2 – RMSE of datasets **A** and **B** for all models fitted to dataset **B**.

Model	**RMSE** (**A**|**B**)	**RMSE** (**B**|**B**)
MoB	5.1326×10^{-2}	7.3952×10^{-2}
Poly-9	5.1889×10^{-2}	7.4556×10^{-2}
MLP(1,12,1)	5.1086×10^{-2}	7.3928×10^{-2}
L3P	5.2654×10^{-2}	7.6071×10^{-2}
0-GP	5.1199×10^{-2}	7.4093×10^{-2}
L3P-HG-GP	5.1033×10^{-2}	7.3982×10^{-2}
L3P-HS-GP	4.9820×10^{-2}	7.6261×10^{-2}
L3P-LRHS-GP	4.9694×10^{-2}	7.6143×10^{-2}

Table 3 – MNLPD of dataset **A** for all GP models fitted on datasets **A** and **B**.

Model	**MNLPD** (**A**|**A**)	**MNLPD** (**A**|**B**)
0-GP	-1.5879	-1.4444
L3P-HG-GP	-1.9878	-1.7263
L3P-HS-GP	-2.3292	-2.2973
L3P-LRHS-GP	-2.3365	-2.3047

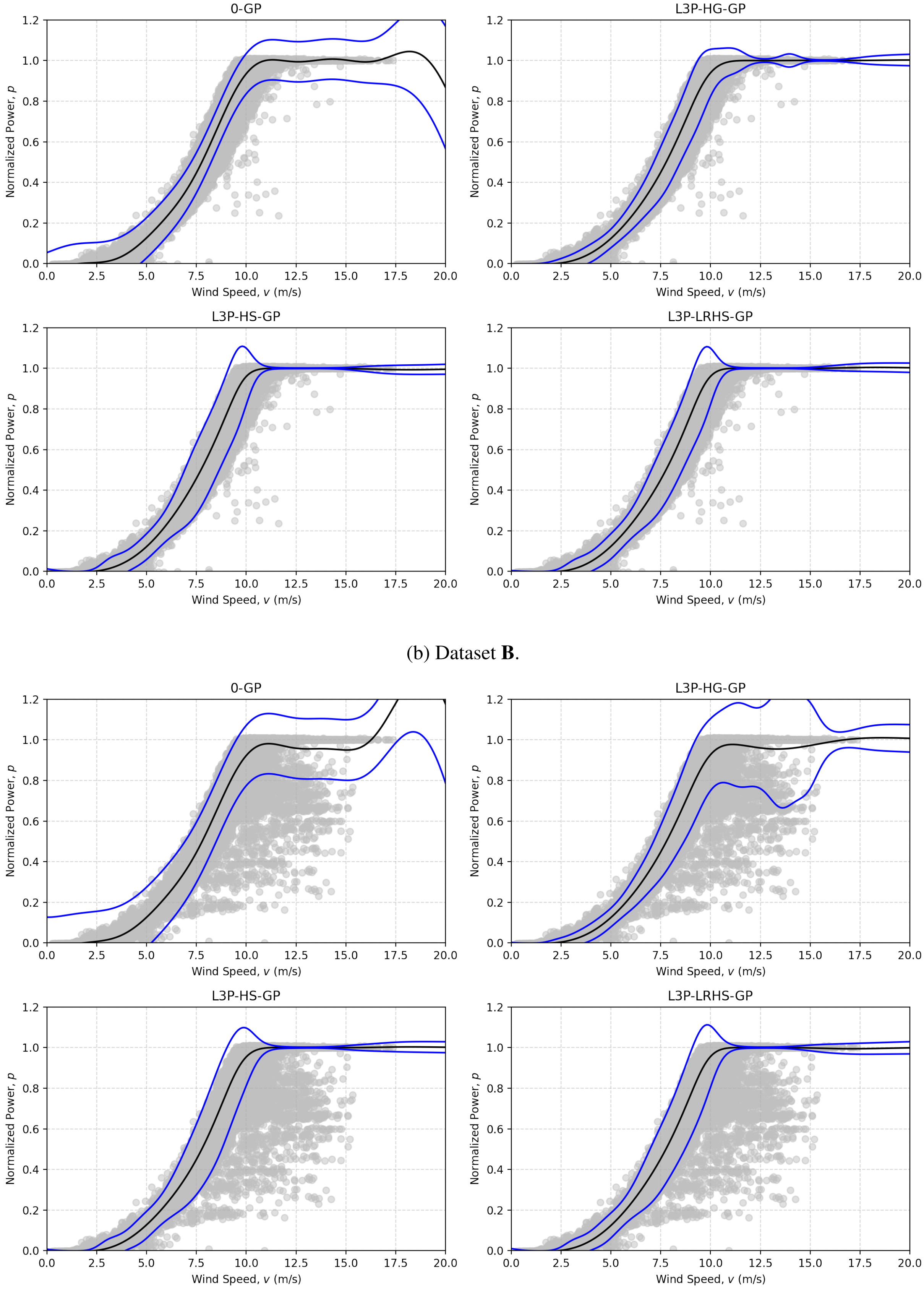
The issue can be further analyzed by comparing figs. 7 and 8, which shows that the presence of outliers in the high wind speed region is responsible for distortions on many of the considered models, with the exceptions being the L3P, L3P-HS-GP and L3P-LRHS-GP models. Whilst the L3P component and its strict functional form contributes to it, it is not the single responsible for this desired behavior as evidentiated the deviations present in the L3P-HG-GP model, which ends up incorrectly adapting itself to the outliers due to the GP component. In fact, only the proposed models with Student-t derived likelihoods are robust to outliers, which is in accordance to the theory.

5.4.4 Heteroscedasticity

In this experiment, the probabilistic WTPC models are analyzed regarding their ability to represent the heteroscedasticity observed in the data. All GP models from sections 5.1 and 5.2 were fitted to both datasets **A** and **B** and compared in terms of the MNLPD evaluated on dataset **A**. The other models are not considered as they are deterministic and as such do not quantify uncertainty. The results are shown in fig. 9 and in table 3.

The heteroscedasticity effect is caracterized by the variation of the separation between the blue lines in fig. 9 as the wind speed varies. As can be seen in fig. 9, all the proposed models (L3P-HG-GP, L3P-HS-GP and L3P-LRHS-GP) are able to properly express it, whereas the benchmark model 0-GP exhibits a constant separation between them as its Gaussian likelihood

Figure 9 – Predictive probability density for all GP models fitted to datasets **A** and **B**. The black lines are the mean of the predictions $\mu_{\hat{y}}$, and the blue lines delimit the symmetric 2 standard deviations interval of the predictions $\mu_{\hat{y}} \pm 2\sigma_{\hat{y}}$.



(a) Dataset **A**.

(b) Dataset **B**.

Source - The Author

Table 4 – Month-based cross-validation statistics for RMSE scores. $\mathbf{Q}_1$ and $\mathbf{Q}_3$ are the lower and upper quartiles of the data, respectively.

Model	**Mean**	**St.Dev.**	**Min.**	$\mathbf{Q}_1$	**Median**	$\mathbf{Q}_3$	**Max.**
MoB	0.0505	0.0064	0.0369	0.0468	0.0507	0.0533	0.0701
Poly-9	1.0187	3.3139	0.0383	0.0509	0.0634	0.2715	23.1293
MLP(1,12,1)	0.0517	0.0084	0.0369	0.0467	0.0509	0.0538	0.0881
L3P	0.0542	0.0061	0.0417	0.0502	0.0536	0.0575	0.0734
0-GP	0.0530	0.0116	0.0371	0.0472	0.0513	0.0541	0.1209
L3P-HG-GP	0.0512	0.0069	0.0371	0.0469	0.0508	0.0544	0.0684
L3P-HS-GP	0.0532	0.0077	0.0375	0.0477	0.0523	0.0594	0.0738
L3P-LRHS-GP	0.0526	0.0078	0.0374	0.0472	0.0518	0.0580	0.0737

is inherently homoscedastic. Those graphical observations are in accordance with values of MNLPD ($\mathbf{A}|\mathbf{A}$) in table 3, where the 0-GP benchmark model has much worse results than the proposed models.

Comparing figs. 9a and 9b, it is possible to see that the presence of outliers strongly impacts the L3P-HG-GP model, which exhibits a very large uncertainty for higher wind speeds when fitted to dataset **B**. Oppositely, the L3P-HS-GP and L3P-LRHS-GP models are more robust to them, presenting very similar graphical results independently of the dataset used for fitting. Those qualitative graphical observations are quantitatively reflected in table 3, where the MNLPD increases much more for L3P-HG-GP than for L3P-HS-GP and L3P-LRHS-GP when the fitting dataset is changed.

5.4.5 Seasonal Variations

In this experiments, the impact of seasonal variations of the wind speed distribution on the WTPC modeling task is analyzed. A cross-validation scheme was adopted by splitting the dataset **A** into monthly datasets $\mathbf{A}_1$, $\mathbf{A}_2$, …, $\mathbf{A}_{12}$ and all models from sections 5.1 and 5.2 were fitted for each them. Then, for each model-month combination, the RMSE of the other months, i.e., RMSE ($\mathbf{A}_i|\mathbf{A}_j$), $i \neq j$, was evaluated, totalizing $12 \cdot 11 = 132$ evaluations of the RMSE score per model. The statistics of this cross-validation procedure are shown in table 4.

The analysis of table 4 shows that the proposed models have very competitive mean and median results, as they are close to the results of the MoB and MLP(1,12,1), which are the best performing models in this respect. Moreover, the proposed models also have a smaller standard deviation than the MLP(1,12,1), which means they behave more consistently across the months. Furthermore, the maximum cross-validated RMSE of the proposed models is in general small (in fact, the L3P-HG-GP is the best performing model in this aspect), which shows that even in the worst-case, the proposed models are still performing well. All those observations shows that the proposed modeling framework does not suffer performance impacts from seasonal variability and generalizes well for different wind speed distributions.

5.5 Discussion

In this chapter, the main objective of this dissertation was achieved with the proposition of a WTPC modeling framework based on Chained GPs and Logistic function models. The construction rationale of the models was presented highlighting how the domain knowledge was incorporated into the model, followed by a detailed explanation of the models implementation, initialization and optimization procedures.

The benchmark models were formally presented to build a set of state-of-art methodologies to be compared with the proposed one. The implementation, initialization and optimization procedures of those models were also briefly explained.

A rich 1-year of WT operation dataset was introduced to serve as a case-study for comparing the proposed modeling framework to the benchmark models. The WT parameters and data sources were described and the air density normalization procedure of the IEC Standard was applied to the data. The available event log allowed thr construction of two datasets **A** and **B**, which were used to exemplify the peculiarities of the WTPC modeling problem such as heteroscedasticity and the presence of outliers. Finally, an analysis of the wind speed data seasonality was conducted, showing the importance of considering this aspect when building WTPC models.

All those pieces were joined together in a series of experiments comparing the models in terms of data-fitting capabilities, robustness to outliers, heteroscedasticity representation and ability to deal with seasonal variations. The results of those experiments have shown that the proposed modeling framework is either on-par or better than the considered benchmark models, constituting a great resource for WTPC modeling.

6 CONCLUSION

The WTPC modeling problem is of great interest for both scientific and technical literature and the wind power industry. Motivated by its importance and also by its peculiarities such as heteroscedasticity, presence of outliers and data seasonality, this dissertation sought to add its contributions to this challenging problem by proposing a new probabilistic, semi-parametric, and data-based modeling framework based on Gaussian processes combined with logistic function models.

Due to its importance for the proposed modeling framework, the GP theory was revisited, always emphasizing the practical interpretation of the presented concepts. Aiming to cope with the large amount of data used WTPC modeling, The SVGP was then introduced to tackle the scalability issues inherent to the standard GP regression. Finally, the Chained GP was discussed in order to deal with more complex features in the noisy data distribution such as the ones found in WTPC modeling.

The proposed modeling framework was put to the test in a series of computational experiments using a dataset of one year of WT operational data. Those studies showed that the new models were not only able to give competitive results when compared to selected state-of-art models regarding deterministic metrics but also to successfully model the characteristics of the uncertainty intrinsic to the noisy observations.

Therefore, the objectives of this dissertation were successfully achieved. As such, it is expected that the proposed modeling methodology becomes part of the vast set of WTPC modeling tools, contributing to the development of the wind energy industry and, consequently, to the worldwide transition towards a more sustainable energy sources.

6.1 Further Work

The proposed modeling framework developed in this dissertation established a probabilistic approach to the WTPC modeling problem based on the adaptability of GP models combined with domain knowledge based choices such as the logistic function as the GP mean function and likelihoods tailored for the noise features of the WTPC data. Building on top of those features, some interesting directions of future research are now discussed.

6.1.1 Evaluation of Other Likelihoods

The likelihoods studied in this dissertation were chosen due to their ability to represent heteroscedasticity and, in the Student-t distribution based cases, be robust to outliers. However, they are not able to represent asymmetric distributions, which can be the case for at least part of the WTPC data. As such, likelihoods based on asymmetric distributions are good options for future research.

6.1.2 More Inputs

This dissertation followed the more usual approach to WTPC in both technical and scientific literature of modeling the normalized power p as a function of only one input, the wind speed v. However, it is also recognized in both literatures that many more variables such as wind turbulence intensity, wind veer and sheer, yaw misalignment angle and air density. Fortunately, the proposed modeling framework can be easily extended to include then as additional inputs of the latent GPs, which opens a very promising research option.

6.1.3 Alternative Quadrature Methods

Although not explicitly evaluated in the experiments of section 5.4, the computational time and storage requirements for fitting the L3P-LRHS-GP model are much more demanding than the L3P-HS-GP model. This happens because the former needs to evaluate three-dimensional Gaussian expectations whilst the latter only needs a two-dimensional one, and the chosen algorithm to compute them, the Gauss-Hermite quadrature, scales poorly for higher dimensions. Hence, exploring other options to compute expectations such as the Monte Carlo quadrature or the Unscented Transform is an important enhancement not only for WTPC modeling, but for Chained GPs in general.

REFERENCES

ABADI, M. *et al.* **TensorFlow**: A system for large-scale machine learning. In: USENIX SYMPOSIUM ON OPERATING SYSTEMS DESIGN AND IMPLEMENTATION, 12., 2016, Savannah, GA. **Proceedings** [...]. [*S. l.*: *s. n.*], 2016. v. 1, p. 265–283. Disponível em: http://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf. Acesso em: 05 dez. 2020.

BAI, L. *et al.* Wind turbine power curve estimation based on earth mover distance and artificial neural networks. **IET Renewable Power Generation**, v. 13, no. 15, p. 2939–2946, 2019.

BLEI, D. M.; KUCUKELBIR, A.; MCAULIFFE, J. D. Variational inference: A review for statisticians. **Journal of the American statistical Association**, [*s. l.*], v. 112, no. 518, p. 859–877, 2017.

CARRILLO, C. *et al.* Review of power curve modelling for wind turbines. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 21, p. 572–581, 2013.

DELFOS INTELLIGENT MAINTENANCE. **Delfos Intelligent Maintenance**. 2017–. Disponível em: "http://www.delfosim.com/". Acesso em: 05 dez. 2020.

EMINOGLU, U.; TURKSOY, O. Power curve modeling for wind turbine systems: a comparison study. **International Journal of Ambient Energy**, [*s. l.*], p. 1–10, 2019.

GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 13., 2010, Sardinia. **Proceedings** [...]. [*S. l.*: *s. n.*], 2010. v. 1, p. 249–256.

GOLUB, G. H.; Van Loan, C. F. **Matrix Computations**. 4th. ed. [*S. l.*]: Johns Hopkins University Press, 2012.

GUO, P.; INFIELD, D. Wind turbine power curve modeling and monitoring with gaussian process and sprt. **IEEE Transactions on Sustainable Energy**, [*s. l.*], v. 11, no. 1, p. 107–115, 2018.

HENSMAN, J.; MATTHEWS, A.; GHAHRAMANI, Z. Scalable variational gaussian process classification. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 18., 2015, [S. l.] **Online Proceedings** [...]. Cambridge, MA: JMLR, 2015. v. 38, p. 351–360. Disponível em: http://proceedings.mlr.press/v38/hensman15.pdf. Acesso em: 05 dez. 2020.

INTERNATIONAL ELECTROTECHNICAL COMMISSION. **Wind turbines - Part 12-2**: Power performance of electricity-producing wind turbines based on nacelle anemometry. Geneva: IEC, 2013. v. 2013.

INTERNATIONAL ELECTROTECHNICAL COMMISSION. **Wind energy generation systems - Part 12-1**: Power performance measurements of electricity producing wind turbines. Geneva: IEC, 2017. v. 2017.

JEPSEN, F.; SØBORG, A.; YANG, Z. Disturbance control of the hydraulic brake in a wind turbine. In: INTERNATIONAL ENERGY CONFERENCE, 2010, Manama. **Online Proceedings** [...]. New Jersey: IEEE, 2010. p. 530–535. Disponível em: http://ieeexplore.ieee.org/document/5771739. Acesso em: 05 dez. 2020.

JONES, E.; OLIPHANT, T.; PETERSON, P. *et al.* **SciPy:** Open source scientific tools for Python. 2001–. Disponível em: http://www.scipy.org/. Acesso em: 05 dez. 2020.

LÁZARO-GREDILLA, M.; TITSIAS, M. K. Variational heteroscedastic gaussian process regression. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 28., 2011, Bellevue, WA. **Online Proceedings** [...]. Bellevue, WA: ICML, 2011. p. 841–848. Disponível em: http://icml.cc/Conferences/2011/papers/456_icmlpaper.pdf. Acesso em: 05 dez. 2020.

LI, S.; WUNSCH, D. C.; O'HAIR, E.; GIESSELMANN, M. G. Comparative analysis of regression and artificial neural network models for wind turbine power curve estimation. **Journal Solar Energy Engineering**, [*S. l.*], v. 123, no. 4, p. 327–332, 2001.

LYDIA, M. *et al.* Advanced algorithms for wind turbine power curve modeling. **IEEE Transactions on sustainable energy**, [*s. l.*], v. 4, no. 3, p. 827–835, 2013.

LYDIA, M. *et al.* A comprehensive review on wind turbine power curve modeling techniques. **Renewable and Sustainable Energy Reviews**, [*S. l.*], v. 30, p. 452–460, 2014.

MANOBEL, B.; SEHNKE, F.; LAZZÚS, J. A.; SALFATE, I.; FELDER, M.; MONTECINOS, S. Wind turbine power curve modeling based on gaussian processes and artificial neural networks. **Renewable Energy**, [*S. l.*], v. 125, p. 1015–1020, 2018.

MATTHEWS, A. G. d. G. **Scalable Gaussian process inference using variational methods.** 2017. Thesis (Ph.D.) — University of Cambridge, 2017.

MATTHEWS, A. G. d. G. *et al.* On sparse variational methods and the kullback-leibler divergence between stochastic processes. In: ARTIFICIAL INTELLIGENCE AND STATISTICS, 19., 2016, Cadiz. **Online Proceedings** [...]. Cambridge, MA: JMLR, 2016. v. 51, p. 231–239. Disponível em: http://proceedings.mlr.press/v51/matthews16.html. Acesso em: 05 dez. 2020.

MATTHEWS, A. G. d. G. *et al.* GPflow: A Gaussian process library using TensorFlow. **Journal of Machine Learning Research**, [Cambridge, MA], v. 18, no. 40, p. 1–6, apr 2017. Disponível em: http://jmlr.org/papers/v18/16-537.html. Acesso em: 05 dez. 2020.

NOCEDAL, J.; WRIGHT, S. J. **Numerical Optimization**. 2nd. ed. New York, NY, USA: Springer, 2006.

OLIPHANT, T. **NumPy:** A guide to NumPy. 2006–. USA: Trelgol Publishing. Disponível em: http://www.numpy.org/. Acesso em: 05 dez. 2020.

PANDIT, R. K.; INFIELD, D. Comparative analysis of gaussian process power curve models based on different stationary covariance functions for the purpose of improving model accuracy. **Renewable energy**, [*S. l.*], v. 140, p. 190–202, 2019.

PANDIT, R. K.; INFIELD, D.; KOLIOS, A. Comparison of advanced non-parametric models for wind turbine power curves. **IET Renewable Power Generation**, v. 13, no. 9, p. 1503–1510, 2019.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RASMUSSEN, C. E.; WILLIAMS, C. K. I. **Gaussian processes for machine learning**. [*S. l.*]: The MIT Press, 2006. ISBN 026218253X.

SAUL, A. D. *et al.* Chained gaussian processes. In: ARTIFICIAL INTELLIGENCE AND STATISTICS, 19., 2016, Cadiz. **Online Proceedings** [...]. Cambridge, MA: JLMR, 2016. p. 1431–1440. Disponível em: http://proceedings.mlr.press/v51/saul16.pdf. Acesso em: 05 dez. 2020.

SHOKRZADEH, S.; JOZANI, M. J.; BIBEAU, E. Wind turbine power curve modeling using advanced parametric and nonparametric methods. **IEEE Transactions on Sustainable Energy**, [*S. l.*], v. 5, no. 4, p. 1262–1269, 2014.

SOHONI, V.; GUPTA, S.; NEMA, R. A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems. **Journal of Energy**, [London], v. 2016, 2016.

TITSIAS, M. Variational learning of inducing variables in sparse gaussian processes. In: **Artificial Intelligence and Statistics**. Cambridge, MA: JMLR, 2009. p. 567–574. Disponível em: http://jmlr.org/papers/v18/16-537.html. Acesso em: 05 dez. 2020.

VILLANUEVA, D.; FEIJÓO, A. Comparison of logistic functions for modeling wind turbine power curves. **Electric Power Systems Research**, [*S. l.*, v. 155, p. 281–288, 2018.

VIRGOLINO, G. C. de M.; MATTOS, C. L. C.; MAGALHÃES, J. A. F.; BARRETO, G. A. Gaussian processes with logistic mean function for modeling wind turbine power curves. **Renewable Energy**, [United Kingdom, v. 162, p. 458–465, 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0960148120309150. Acesso em: 05 dez. 2020.

WANG, Y.; HU, Q.; LI, L.; FOLEY, A. M.; SRINIVASAN, D. Approaches to wind power curve modeling: A review and discussion. **Renewable and Sustainable Energy Reviews**, [United Kingdom], v. 116, p. 109422, 2019.

YAN, J.; ZHANG, H.; LIU, Y.; HAN, S.; LI, L. Uncertainty estimation for wind energy conversion by probabilistic wind turbine power curve modelling. **Applied energy**, [United Kingdom, v. 239, p. 1356–1370, 2019.