Skew Gaussian Process for non Linear Regression

By

Ekhlas Yousef Soliman Al-Momani

Supervisor

Dr. Moh'd Taleb Alodat

Program: Statistics

May 5, 2011

Skew Gaussian Process for non Linear Regression

By

Ekhlas Yousef Soliman Al-Momani

B.Sc. Mathematics, Irbid National University, 2006

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of statistics, Yarmouk University, Irbid, Jordan.

Approved by:

Dr. Moh'd T. Alodat...........................Chairman.

Associate Professor of Statistics, Yarmouk University.

Prof. Zeyad R. Al-Rawi Z..R..Al..Rawi**....Member.**

Professor of Statistics, Yarmouk University.

Dr. Qutaibeh D. Katatbeh.....................Member.

Associate Professor of Mathematics and Statistics, Jordan University for Science and Technology.

بسم الله الرحمن الرحيم

" قالوا سبحانك لاعلم لنا إلا ما علمتنا انك أنت

العليم الحكيم "

البقرة 32

الإهداء

إلى من منحني العزم والإصرار والثبات

والدي العزيز

إلى من علمتني قيمة العلم و معناه

والدتي العزيزة

إلى من يفرح لفرحي وساندني في تحمل أعباء الدراسة

زوجي الفاضل

إلى من أمدوني بالدعم المادي والمعنوي طول مدة الدراسة

أخي العزيز المهندس أنس

أخواتي العزيزات: المدرسة الفاضلة يمامه، المهندسة وفاء، الدكتورة إنصاف والدكتورة هدى

إلى كل من كانوا قدوتي وعلموني كل ما بوسعهم

أساتذتي الأفاضل

إلى كل هؤلاء والى صديقاتي العزيزات، اهدي عملي المتواضع هذا

إخلاص المومني

Acknowledgment

At the beginning, I want to thank ALLAH, the most merciful and most gracious.

I wish to express my deepest gratitude to my supervisor, Dr. Mohammad Alodat for suggestion the problem and for her advice, patience, steadfast encouragement, support and continuous help during the course of my work on the subject of this thesis.

I would like to thank the members of the defense committee, Prof. Zeyad R, Al-Rawi and Dr. Qutaibeh D. Katatbeh for their valuable comments and remarks that that improve my thesis.

I want to say, I am very grateful to Yarmouk University and Irbid National University for providing the opportunity to do my graduate studies.

I would like to convey special thanks to my parents, husband, brother and sisters for their patience and encouragement throughout my study.

Finally, I would like to say, thank you my friends, my fellows at Yarmouk University.

ABSTRACT

Al-Momani, Ekhlas yousef. Skew Gaussian Process for non Linear Regression. Master of Science Thesis, Department of Statistics, Yarmouk University, 2011. (Supervisor: Dr. Moh'd Taleb Alodat).

In this thesis, we extend the Gaussian process for regression model by assuming a skew-Gaussian process prior on the input function and a skew-Gaussian white noise on the error term. Under these assumptions, the predictive density of the output function at a new fixed input is obtained in a closed form. Also, we study the Gaussian process predictor when the errors depart from Gaussianity to skew-Gaussian white noise. The bias is derived in a closed form and is studied for some special cases.

We conduct a simulation study to compare the empirical distribution function of the Gaussian process predictor under Gaussian white noise and skew-Gaussian white noise.

Keywords: Conditional distribution; Gaussian process (G); Likelihood approximation; Monte-Carlo approximation; Multivariate normal distribution; Predictive density function; Regression model; Regular polygon; Skew-Gaussian process (SG).

ملخص

المومني، إخلاص يوسف. العملية الجاوسية الملتوية للإنحدار غير الخطي، رسالة ماجستير في العلوم، قسم الإحصاء، جامة اليرموك 2011. (المشرف الرئيسي: الدكتور محمد طالب العودات).

في هذه الأطروحة نوسع العملية الجاوسية بنموذج عبر افتراض أن الاقتران المدخل يتبع التوزيع الجاوسي الملتوي وأيضا يكون هذا الفرض بالنسبة للخطأ مع وجود متغيرات مزعجة. تبعا لهذه الإفتراضات، فإن التوزيع المتوقع للإقتران الناتج عند ادخال قيمة جديدة يتم الحصول عليها بشكل مغلق. أيضا ندرس التوقع الجاوسي عندما يكون الخطأ بعيد عن التوزيع الجاوسي ويقترب من التوزيع الجاوسي الملتوي بمتغيرات مزعجة. تم اشتقاق صيغة محددة لمعرفة الفرق بين القيمة المتوقعة والقيمة الحقيقية ويدرس لبعض الحالات الخاصة. نجري التجربة عدة مرات لمقارنة الاقتران الحقيقي مع الإقتران الذي تم توقعه بالعملية الجاوسية خلال خضوعها للتوزيع الجاوسي الملتوي بمتغيرات مزعجة.

الكلمات المفتاحية: التوزيع الشرطي؛ العملية الجاوسية؛ احتمال التقريب؛ تقريب مونتي كارلو؛ التوزيع الطبيعي متعدد المتغيرات؛ اقتران الكثافة التنبؤي؛ نموذج الانحدار؛ مضلع منتظم؛ العملية الجاوسية الملتوية.

List of Abbreviations

1- $SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$: Skew-Normal distribution of n-dimension with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and skewness parameter $\boldsymbol{\lambda}$.

2- $CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{v}, \boldsymbol{\Delta})$: Closed Skew-Normal distribution of p-dimension with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{D}$, $\boldsymbol{v}, \boldsymbol{\Delta}$.

3- pdf : The probability density function.

4- CDF : The cumulative density function.

5- $N(\mu, \sigma^2)$: Normal distribution with mean μ and variance σ^2.

6- $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Normal distribution of n-dimension with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

7- $\Phi(.)$: CDF of the normal distribution.

8- $\phi(.)$: pdf of the normal distribution.

9- k(., .): The covariance function.

10- $\boldsymbol{Y}^T$: The transpose of $\boldsymbol{Y}$, i.e. if $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, then $\boldsymbol{Y}^T = (Y_1 \quad \dots \quad Y_n)$.

11- $p(X|Y)$: The distribution of X given Y.

12- $M_{X|Y}$: The moment generating function of X given Y.

13- $var(X)$: the variance of X.

14- $E(X)$: The expected value of X, i.e. the mean of X.

15- *mse*: mean square error.

16- GPR : Gaussian Process for Regression.

17- SGPR : Skew Gaussian Process for Regression.

18- iid : identically independent distribution.

TABLE OF CONTENTS

CHAPTER FOUR

CHAPTER FIVE

APPENDIX

CHAPTER ONE

INTRODUCTION

In statistical literature, the assumption of Gaussianity or normality has been made on statistical models for long time when analyzing spatial data. The popularity of using Gaussian assumption is due to the suitable properties that are possessed by the Gaussian or normal distribution such as closure under marginal and conditional distributions, as well as the closure under convolution.

Despite of such nice properties of Gaussian distribution, the assumption of Gaussianity is hardly fulfilled for various spatial processes due to lack of symmetry, unimodality, etc. Alodat et al. (2010) have given several examples. Another example is given by Anagreh et al. (2010) where they showed that the distribution of wind speeds of four stations in Jordan are well-fitted by a skew normal distribution. It is found that if a statistical model, which is relying on the Gaussianity assumption, is used to analyze a skewed data, then unrealistic or nonsensical results will be produced.

A number of methods have been proposed to treat skewed data. All these methods are relying on Gaussianizing the data, i.e., by transforming the data to near Gaussian data. Such transformation method is not recommended due to the following different reasons (Alodat et al., 2010):

1. It is not easy to find a suitable transformation to achieve normality for several data sets.

2. Since transformations are usually performed component-wise (where normality of marginal's does not guarantee the joint normality), then the statistical problem might be not invariant under these transformations, which leads to biased estimates.

3. A transformation on the data may reduce the amount of information in the original data unless the transformation is a sufficient statistic.

4. Despite of the difficulty in interpreting the transformed data, the data skewness has some interpretation and hence could not be ignored (Buccianti, 2005).

For these reasons, a non-Gaussian model is needed to capture the skewness in data. The challenge in using non-Gaussian distribution or assumption on statistical models is in finding the predictive density formula which is analytically intractable.

Recently, random processes, that possess a skewness parameter, have been defined by several researchers. Alodat and Aludaat (2007) employed the skew normal theory as presented in Genton (2004) to define a new random process, called the skew-Gaussian process. Also they gave an application real data. Relying on the multivariate closed-skew normal distribution of Gonzaalez-Farias et al. (2003), Allard and Navea (2007) defined what they called the closed skew-normal random field. Other skew random processes or fields are introduced and applied to real data in (Zhang and El-Sharaawi, 2009), and (Alodat and Al-Rawwash, 2009).

All these skew processes are defined based on the pioneer skew-normal distributions introduced in the sequence of papers: Azzalini (1985, 1986), Azzalini and Dalla valle (1996) and Azzalini and Capitanio, 1999). The skew-normal or skew-Gaussian distribution is defined as follows. A random vector $\boldsymbol{X}_{(n\times 1)}$ is said to have an n −dimentional multivariate skew-normal distribution if it has the pdf

$$f_X(\boldsymbol{x}) = 2\phi_n(\boldsymbol{x};\boldsymbol{0},\Sigma)\Phi(\boldsymbol{\alpha}^T\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^n \qquad (1.1)$$

where $\phi_n(.;\boldsymbol{0},\Sigma)$ is the pdf of $N_n(\boldsymbol{0},\Sigma)$, $\Phi(.)$ is the CDF of $N(0,1)$, and $\boldsymbol{\alpha}_{(n\times 1)}$ is a vector called the skewness parameter.

It has been shown that the family of skew normal distributions possesses properties that are close to or coincide with those of normal family, in addition it contains the normal family, i.e., when $\boldsymbol{\alpha} = \boldsymbol{0}$ (Alodat et al., 2010).

Such properties have attracted researchers to extend the well known statistical techniques under the skew-normality assumption. To see how this family was attractive to researchers, please (see Genton, 2004).

There are still a lot of statistical models which have not been yet investigated or not under skew-normality assumption. The Gaussian process for regression (GPR) model is a statistical technique introduced by Neil (1995) to treat a non-linear regression $Y(t) = f(t) + \epsilon(t)$ from a Bayesian point of view. Simply, the technique assumes a Gaussian process as a prior on $f(t)$ while $\epsilon(t)$ is assumed to have a white noise, i.e., an independent Gaussian process is used to define distribution over functions space, and $f(t)$ is a realization of that distribution.

Alodat et al. (2010) have extended the GPR model under skew normal assumptions. They considered two cases. In the first case they assumed a skew Gaussian process as a prior on $f(t)$ while $\epsilon(t)$ is left to have a white noise. By a skew Gaussian process they mean a family of random variables where for which every subset of size n, the joint pdf of these random variables follows equation (1.1). In the second case, they assumed a Gaussian process as a prior on $f(t)$ while $\epsilon(t)$ is assumed to have a skew Gaussian process. In both cases they derived the predictive density of a new observation. In their analysis of GPR, they extended the distribution of Arnold and Beaver (2002) and then they showed that it is closed under convolution with normal distribution. This closure property has convenienced the mathematical derivation and produced a tractable predictive distribution.

Since the Gaussian family is a sub-family of the skew Gaussian family, then the skew Gaussian process, as a prior on $f(t)$, allows us to define a distribution over a more rich family of functions than the Gaussian one. Also, it allows us to extend the error term in the above regression model to have a skew distribution which closer to real world than its Gaussian counterpart.

It appears from extensive literature on Gaussian process for regression that the GP has a significant applications in various fields of science. It has been applied to model noisy data and to classify problems arise in Machine learning for learning the inverse dynamics of a robot arm (Rasmussen and Williams, 2006).

Brahim-Belhouari and Bermak (2004) applied the Gaussian Process Regression model to predict the future value of a non-stationary time series.

Schmidt et al. (2008) studied the sensitivity of Gaussian process to the choice of correlation function. Based on a numerical study, they concluded that the predictions did not differ much amongst the different correlation functions.

Vanhatalo et al. (2009) proposed a GPR with student-t likelihood by approximating the joint distribution of process values by a student distribution.

The idea beyond this approximation is to make the GPR model robust against outliers. The model they proposed is analytically intractable.

Kuss (2006) proposed other robust models as alternatives for GPR.

Macke et al. (2010) applied the Gaussian process for regression to estimate the cortical map of the human brain. They modeled an image arises in their experiment by a Gaussian process where the activity at each voxel is measured.

Fyfe et al. (2008) have applied the GPR to Canonical correlation analysis with application to neuron data.

The problem of treating the prediction problem of the non linear regression $Y(\boldsymbol{t}) = f(\boldsymbol{t}) + \epsilon(\boldsymbol{t})$ from a Bayesian view point when both $f(\boldsymbol{t})$ and $\epsilon(\boldsymbol{t})$ follow skew Gaussian processes has not yet been a dressed in the literature. In this thesis, we will extend their work by assuming two skew Gaussian normal processes on both $f(\boldsymbol{t})$ and $\epsilon(\boldsymbol{t})$.

In this thesis, we consider the non-linear regression model $Y_i = f(\boldsymbol{t}_i) + \epsilon(\boldsymbol{t}_i)$, $i = 1, 2, \dots, n$, where $f(\boldsymbol{t}_i)$'s are the output values of $f(\boldsymbol{t})$ and $\epsilon(\boldsymbol{t}_i)$'s are iid $N(0, \tau^2)$. We put a skew-Gaussian process prior on the function $f(\boldsymbol{t})$. Moreover, we consider the following two prediction problems

(i) Prediction of $f(\boldsymbol{t})$ at a fixed input $\boldsymbol{t}$.

(ii) Prediction of $f(\boldsymbol{t})$ at a random input $\boldsymbol{t}^*$.

In both cases, it is expected that the predictive densities have no closed forms. Hence, numerical and analytical approximations are needed. For this we will use the approximations in section (3.6) to approximate the predictive densities. Also we will compare the accuracy of these approximations.

The rest of this thesis is organized as follows. In chapter two we give an introduction to Gaussian processes for regression. Also we give some definitions on skew normal distribution and skew-Gaussian processes. In chapter three, we generalize the skew Gaussian process for regression by assuming a skew Gaussian process on $f(t)$ and another skew Gaussian process on $\epsilon(t)$. Then we derive the predictive density of the output function at new input. Also, we derive the mean and the variance of the predictive distribution. Finally, we study the bias mean square error of the prediction for two special cases under assumption violation. In chapter four, we conduct a simulation study to compare the new model to the Gaussian one. Also, we apply our finding to real data. In chapter five, we report our conclusion and we propose future work.

CHAPTER TWO

GAUSSIAN AND SKEW-GAUSSIAN PROCESS FOR REGRESSION

In this chapter, we introduce the reader to both Gaussian processes for regression and multivariate skew normal theory needed in the next chapters for generalizing the Gaussian processes for regression.

2.1. Gaussian Process for Regression

A family $\{X(\boldsymbol{t}), \boldsymbol{t} \in \mathcal{C}\}, \mathcal{C} \subseteq \mathbb{R}^n$ of random variables is said to constitute a Gaussian process if for n and $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n \in \mathcal{C}$, the random variables $X_1(\boldsymbol{t}), \dots, X_n(\boldsymbol{t})$ have n-dimentional multivariate normal distribution.

A Gaussian process allows us to do a non-parametric treatment of a non-linear regression from a Bayesian point of view. O'Hagan (1978), was the first to employ the Gaussian process in a non-parametric frame work, while an application of O'Hagan's work to Bayesian learning in networks has been appeared in Neal (1995).

The Gaussian process for regression, as proposed by Neil (1995), can be illustrated as follows: O'Hagan (1978); Consider a set of training data $\mathcal{D} = \{(\boldsymbol{t}_1, Y_1), \dots, (\boldsymbol{t}_n, Y_n)\}$, where the input vectors $\boldsymbol{t}_1$, $\boldsymbol{t}_2$, $\boldsymbol{t}_3, \dots, \boldsymbol{t}_n \in \mathcal{C} \subseteq \mathbb{R}^n$ and their output values $Y_1, Y_2, \dots, Y_n$ are governed by the non-linear regression model $Y_i = f(\boldsymbol{t}_i) + \epsilon(\boldsymbol{t}_i)$, where $\epsilon(\boldsymbol{t}_1), \epsilon(\boldsymbol{t}_2), \dots, \epsilon(\boldsymbol{t}_n)$ are iid Gaussian noises on $\mathcal{C}$ of mean 0 and variance τ^2, and $f(.)$ is an unknown function. The main question is "what is the predicted value of $f^* = f(\boldsymbol{t}^*)$, the value of $f(\boldsymbol{t})$ at a new input $\boldsymbol{t}^*$, say?". To answer this question, a prior distribution is needed on $f(\mathbf{t})$ i.e., A distribution over a set of functions is needed. This prior distribution should be defined on the class of all functions defined on the space of $\boldsymbol{t}$. The set of all sample paths of a Gaussian process on $\mathcal{C}$ provides us with a rich class of such functions.

Assume that $f(\boldsymbol{t}), \boldsymbol{t} \in \mathcal{C}$ is a Gaussian process with covariance function $k(.,.)$; i.e., for every n and $\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \dots, \boldsymbol{t}_n \in \mathcal{C}$, we have $\boldsymbol{f} = \left(f(\boldsymbol{t}_1), \dots, f(\boldsymbol{t}_n)\right)^T \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} k(\boldsymbol{t}_1, \boldsymbol{t}_1) & \cdots & k(\boldsymbol{t}_1, \boldsymbol{t}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{t}_n, \boldsymbol{t}_1) & \cdots & k(\boldsymbol{t}_n, \boldsymbol{t}_n) \end{pmatrix}.$$

A suitable choice for $k(.,.)$ is the following covariance function

$$k(\boldsymbol{t}_i, \boldsymbol{t}_j) = exp\left(-\tfrac{1}{2}(\boldsymbol{t}_i - \boldsymbol{t}_j)^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} (\boldsymbol{t}_i - \boldsymbol{t}_j)\right) \tag{2.1}$$

For simplicity we may consider $\Lambda = diag(\lambda_1^2, \dots, \lambda_n^2)$, where λ_i's are the skewness parameters. A covariance function $k(.,.)$ is said to be isotropic if $k(\boldsymbol{t}_i, \boldsymbol{t}_j)$ depends only on the distance $\|\boldsymbol{t}_i - \boldsymbol{t}_j\|$. For more information about other types of covariance functions see Girard et al. (2004).

Since $f(\boldsymbol{t})$ is a Gaussian process, then the joint pdf of $f(\boldsymbol{t})$ and $f(\boldsymbol{t}^*)$ (Quinonero-Candela et al., 2003) is

$$\begin{pmatrix} f(\boldsymbol{t}_1) \\ \vdots \\ f(\boldsymbol{t}_n) \\ f(\boldsymbol{t}^*) \end{pmatrix} \sim \mathcal{N}_{n+1}(\boldsymbol{0}, \boldsymbol{\Psi}),$$

with

$$\boldsymbol{\Psi} = \begin{pmatrix} k(\boldsymbol{t}_1, \boldsymbol{t}_1) & \cdots & k(\boldsymbol{t}_1, \boldsymbol{t}_n) & k(\boldsymbol{t}_1, \boldsymbol{t}^*) \\ \vdots & \ddots & \vdots & \vdots \\ k(\boldsymbol{t}_n, \boldsymbol{t}_1) & \cdots & k(\boldsymbol{t}_n, \boldsymbol{t}_n) & k(\boldsymbol{t}_n, \boldsymbol{t}^*) \\ k(\boldsymbol{t}^*, \boldsymbol{t}_1) & \cdots & k(\boldsymbol{t}^*, \boldsymbol{t}_n) & k(\boldsymbol{t}^*, \boldsymbol{t}^*) \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{k}(\boldsymbol{t}^*) \\ \boldsymbol{k}^T(\boldsymbol{t}^*) & \mathscr{k} \end{pmatrix},$$

where

$$\boldsymbol{\Sigma} = \{k(\boldsymbol{t}_i, \boldsymbol{t}_j)\}_{i,j=1}^{n},$$

$$\boldsymbol{k}(\boldsymbol{t}^*) = \left(k(\boldsymbol{t}_1, \boldsymbol{t}^*), \dots, k(\boldsymbol{t}_n, \boldsymbol{t}^*)\right)^T$$

and

$$\mathscr{k} = k(\boldsymbol{t}^*, \boldsymbol{t}^*) .$$

Rasmussen (1996) shows that the prediction distribution of f^* given $\boldsymbol{t}^*$ and $\mathcal{D}$ is Gaussian and is given by:

$$p(f^*|\boldsymbol{t}^*,\mathcal{D}) \sim \mathcal{N}\big(\mu(\boldsymbol{t}^*),\sigma^2(\boldsymbol{t}^*)\big), \qquad (2.2)$$

where $\mu(\boldsymbol{t}^*)$ and $\sigma^2(\boldsymbol{t}^*)$ are the mean and the variance of the Gaussian predictive distribution are given by:

$$\mu(\boldsymbol{t}^*) = \boldsymbol{k}^T(\boldsymbol{t}^*)(\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n)^{-1}\boldsymbol{Y}, \text{ where}$$

$$\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^T,$$

and

$$\sigma^2(\boldsymbol{t}^*) = k(\boldsymbol{t}^*,\boldsymbol{t}^*) - \boldsymbol{k}^T(\boldsymbol{t}^*)(\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n)^{-1}\boldsymbol{k}(\boldsymbol{t}^*).$$

The distribution (2.2) can be used to draw several inferential statements about $f(\boldsymbol{t}^*)$. For example, when $p = 1$, a $100(1-\alpha)\%$ prediction interval for $f(\boldsymbol{t}^*)$ is given by $[L, U]$, where L and U are the solution of

$$\int_0^L p(f^*|\boldsymbol{t}^*,\boldsymbol{\mathcal{D}})\, df^* = \frac{\alpha}{2}$$

and

$$\int_U^\infty p(f^*|\boldsymbol{t}^*,\boldsymbol{\mathcal{D}})\, df^* = \frac{\alpha}{2}.$$

For GPR, a $100(1-\alpha)\%$ prediction interval for f^* is

$$\mu(\boldsymbol{t}^*) \pm Z_{1-\frac{\alpha}{2}}\, \sigma(\boldsymbol{t}^*),$$

where $Z_{1-\frac{\alpha}{2}}$ is the $100(1-\alpha)$ quantile of $N(0,1)$. Moreover, the mean $\mu(\boldsymbol{t}^*)$ serves as a predictor for $f(\boldsymbol{t}^*)$ given the data $\boldsymbol{\mathcal{D}}$, while the variance $\sigma^2(\boldsymbol{t}^*)$ is the measure of uncertainty in it.

In the next section, we present the multivariate skew-normal distribution which will be used to define a skew-Gaussian process.

2.2 Multivariate Skew-Normal Distribution

Following Genton et al. (2004), a random vector $\boldsymbol{X} = \left(X_1, X_2, \dots, X_p\right)^T$ is said to have a p-dimensional skew normal distribution, denoted by $\boldsymbol{X} \sim SN_p(\Omega, \boldsymbol{\alpha})$, if it is absolutely continuous and has the pdf

$$f(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x};\Omega)\Phi(\boldsymbol{\alpha}^T\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^p, \qquad (2.3)$$

where $\phi_p(\boldsymbol{x};\Omega)$ denotes the pdf of p-dimensional multivariate normal distribution with standardized marginals, correlation matrix Ω and $\Phi(.)$ is the cumulative distribution function of standard normal variate.

To construct a $SN_p(\Omega, \boldsymbol{\alpha})$, Genton (2004) proposed the following procedure. Let $\boldsymbol{Y} = \left(Y_1, Y_2, \dots, Y_p\right)^T$ have a multivariate normal distributions with standardized marginals, zero mean and correlation matrix $\boldsymbol{\Psi}$. If $Y_0 \sim \mathcal{N}(0,1)$ is independent of $\boldsymbol{Y}$, such that $\begin{bmatrix} Y_0 \\ \boldsymbol{Y} \end{bmatrix} \sim N_{p+1}\left(\boldsymbol{0}, \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi} \end{pmatrix}\right)$, and $\delta_1, \delta_2, \dots, \delta_p \in$

$(-1,1)$, then the random variables $Z_j = \delta_j |Y_0| + \sqrt{1-\delta_j^2}\, Y_j$, have the joint pdf (2.3), with

$$\boldsymbol{\alpha}^T = \frac{\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\Delta}^{-1}}{(1+\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda})^{\frac{1}{2}}},$$

$$\boldsymbol{\Delta} = \operatorname{diag}\left(\sqrt{1-\delta_1^2}, \dots, \sqrt{1-\delta_p^2}\right),$$

$$\Omega = \boldsymbol{\Delta}(\boldsymbol{\Psi} + \boldsymbol{\lambda}\boldsymbol{\lambda}^T)\boldsymbol{\Delta},$$

where $\boldsymbol{\Psi}$ is a $p \times p$ correlation matrix and $\boldsymbol{\lambda} = \left(\alpha_1, \dots, \alpha_p\right)^T$, $\alpha_j = \frac{\delta_j}{\sqrt{1-\delta_j^2}}$, for $j = 1, 2, \dots, p$.

Now, we go to present a generalization to the Gaussian process. This generalization is called the skew-Gaussian process.

2.3 Basic Results on the Multivariate CSN Distribution

Gonza'lez-farias et al. (2003) defines the closed skew-normal distribution as follows:

Definition 2.1: consider $p \geq 1, q \geq 1$, $\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{D}$ an arbitrary $q \times p$ matrix, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta}$ positive definite matrices of dimensions $p \times p$ and $q \times q$, respectively. Then the probability density function (pdf) of the CSN distribution is given by:

$$g_{p,q}(\boldsymbol{y}) = C\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi_q(\boldsymbol{D}(\boldsymbol{y}-\boldsymbol{\mu});\boldsymbol{v},\boldsymbol{\Delta}), \quad \boldsymbol{y}\in\mathbb{R}^p,$$

where C is defined via

$$C^{-1} = \Phi_q(\boldsymbol{0};\boldsymbol{v},\boldsymbol{\Delta}+\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^{\mathrm{T}}),$$

where $\phi_p(.;\boldsymbol{\eta},\boldsymbol{\psi})$, $\Phi_p(.;\boldsymbol{\eta},\boldsymbol{\psi})$ are the pdf and cumulative distribution function (cdf) of a p −dimentional normal distribution. Here $\boldsymbol{\eta}\in\mathbb{R}^p$ denotes the mean and $\boldsymbol{\psi}$ is a $p\times p$ covariance matrix. We denote a p −dimentional random vector distributed according to a CSN distribution with parameters $q,\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{D},\boldsymbol{v},\boldsymbol{\Delta}$ by $\boldsymbol{Y}\sim CSN_{p,q}(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{D},\boldsymbol{v},\boldsymbol{\Delta})$.

The next four lemmas; concerning the multivariate CSN distribution, will be used extensively in the sequel. For proofs, see Genton (2004).

Proposition 2.1 If $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ are independent random vectors with $\boldsymbol{Y}_i\sim CSN_{p_i,q_i}(\boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i,\ \boldsymbol{D}_i,\ \boldsymbol{v}_i,\Delta_i)$, Then the joint distribution of $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ is

$$\boldsymbol{Y} = ({\boldsymbol{Y}_1}^T,\dots,{\boldsymbol{Y}_n}^T)^T\sim CSN_{p^+,q^+}(\boldsymbol{\mu}^+,\boldsymbol{\Sigma}^+,\boldsymbol{D}^+,\boldsymbol{v}^+,\Delta^+),$$

where

$$p^+ = \sum_{i=1}^{n} p_i \ , q^+ = \sum_{i=1}^{n} q_i\ ,\ \boldsymbol{\mu}^+ = (\boldsymbol{\mu}_1^T,\dots,\boldsymbol{\mu}_n^T)^T, \qquad \boldsymbol{\Sigma}^+ = \oplus_{i=1}^{n}\ \boldsymbol{\Sigma}_i$$

and

$$D^{+} = \oplus_{i=1}^{n} D_i,\ v^{+} = (v_1^T, \dots, v_n^T)^T,\ \Delta^{+} = \oplus_{i=1}^{n} \Delta_i\,.$$

also

$$A \oplus B = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

Proposition 2.2 Let $Y \sim CSN_{p,q}(\mu, \Sigma, D, v, \Delta)$ and A be an $n \times p (n \le p)$ matrix of rank n. Then $Ay \sim CSN_{p,q}(\mu_A, \Sigma_A, D_A, v, \Delta_A)$, where

$$\mu_A = A\mu, \quad \Sigma_A = A\Sigma A^T, \quad D_A = D\Sigma A^T \Sigma_A^{-1},$$

and

$$\Delta_A = \Delta + D\Sigma D^T - D\Sigma A^T \Sigma_A^{-1} A\Sigma D^T$$

Proposition 2.3 If $Y \sim CSN_{p,q}(\mu, \Sigma, D, v, \Delta)$, then for two sub vectors Y_1 and Y_2 where $Y^T = (Y_1^T, Y_2^T)$, Y_1 is k –dimensional, $1 \le k \le p$, and μ, Σ, D are partitioned as follows:

$$\mu = \begin{array}{cc} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & \begin{array}{c} k \\ p-k \end{array} \end{array} \quad , \Sigma = \begin{array}{cc} \begin{array}{cc} k & p-k \end{array} & \\ \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} & \begin{array}{c} k \\ p-k \end{array} \end{array}$$

and D_1 , D_2 come from

$$D = \begin{array}{cc} \begin{array}{cc} k & p-k \end{array} & \\ \begin{pmatrix} D_1 & D_2 \end{pmatrix} & q. \end{array}$$

Then the conditional distribution of Y_2 given Y_1 is

$$CSN_{p-k,q}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(y_{10} - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22.1},\ \boldsymbol{D}_2, \boldsymbol{v} - \boldsymbol{D}^*(y_{10} - \boldsymbol{\mu}_1), \Delta),$$

where

$$\boldsymbol{D}^* = \boldsymbol{D}_1 + \boldsymbol{D}_2\, \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1},$$

and

$$\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\, \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}.$$

Proposition 2.4 If $\boldsymbol{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{v}, \Delta)$, then the moment generating function of $\boldsymbol{Y}$ is:

$$M_Y(\boldsymbol{s}) = \frac{\Phi_q(\boldsymbol{D\Sigma s}; \boldsymbol{v}, \Delta + \boldsymbol{D\Sigma D}^{\mathrm{T}})}{\Phi_q(0; \boldsymbol{v}, \Delta + \boldsymbol{D\Sigma D}^{\mathrm{T}})} e^{s^T\mu + \frac{1}{2}s^T\Sigma s}, \quad \boldsymbol{s} \in \mathbb{R}^p.$$

Proposition 2.5 **If** $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are independent random vectors with $\boldsymbol{y}_i \sim CSN_{p,q_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{D}_i, \boldsymbol{v}_i, \boldsymbol{\Delta}_i)$, $i = 1, \dots, n$, then

$$\sum_{i=1}^{n} \boldsymbol{y}_i \sim CSN_{p,q^\circ}(\boldsymbol{\mu}^\circ, \boldsymbol{\Sigma}^\circ, \boldsymbol{D}^\circ, \boldsymbol{v}^\circ, \Delta^\circ),$$

where

$$q^\circ = \textstyle\sum_{i=1}^n q_i, \quad \boldsymbol{\mu}^\circ = \sum_{i=1}^n \boldsymbol{\mu}_i, \quad \boldsymbol{\Sigma}^\circ = \sum_{i=1}^n \boldsymbol{\Sigma}_i,$$

$$\boldsymbol{D}^\circ = \left(\boldsymbol{\Sigma}_i \boldsymbol{D}_i{}^T, \dots, \boldsymbol{\Sigma}_n \boldsymbol{D}_n{}^T\right)^T (\textstyle\sum_{i=1}^n \boldsymbol{\Sigma}_i)^{-1}, \quad \boldsymbol{v}^\circ = (\boldsymbol{v}_i{}^T, \dots, \boldsymbol{v}_n{}^T)^T,$$

and:

$$\Delta^{\circ} = \Delta^{+} + \boldsymbol{D}^{+}\boldsymbol{\Sigma}^{+}\boldsymbol{D}^{+^{T}} - \left(\oplus_{i=1}^{n}(\mathbf{D}_i\boldsymbol{\Sigma}_i)\right)\left(\sum_{i=1}^{n}\boldsymbol{\Sigma}_i\right)^{-1}\left(\oplus_{i=1}^{n}(\boldsymbol{\Sigma}_i\mathbf{D}_i{}^{T})\right),$$

where $\Delta^{+} = \oplus_{i=1}^{n}\Delta_i$, $\boldsymbol{D}^{+} = \oplus_{i=1}^{n}\boldsymbol{D}_i$ and $\boldsymbol{\Sigma}^{+} = \oplus_{i=1}^{n}\boldsymbol{\Sigma}_{\mathrm{i}}$.

In spatial case if $n = 2$ and $\boldsymbol{y}_i \sim CSN_{p,q_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{D}_i, \boldsymbol{v}_i, \boldsymbol{\Delta}_i)$, $i = 1, 2$, we get

$$\boldsymbol{y}_1 + \boldsymbol{y}_2 \sim CSN_{p,q_1+q_2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2, \boldsymbol{D}^{\circ}, \boldsymbol{v}^{\circ}, \Delta^{\circ}),$$

where:

$$\boldsymbol{D}^{\circ} = \begin{pmatrix} \boldsymbol{D}_1\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} \\ \boldsymbol{D}_2\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} \end{pmatrix}, \ \Delta^{\circ} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

and

$$A_{11} = \boldsymbol{\Delta}_1 + \boldsymbol{D}_1\boldsymbol{\Sigma}_1\boldsymbol{D}_1{}^{T} - \boldsymbol{D}_1\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_1\boldsymbol{D}_1{}^{T},$$

$$A_{22} = \boldsymbol{\Delta}_2 + \boldsymbol{D}_2\boldsymbol{\Sigma}_2\boldsymbol{D}_2{}^{T} - \boldsymbol{D}_2\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2\boldsymbol{D}_2{}^{T},$$

$$A_{12} = -\boldsymbol{D}_1\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2\boldsymbol{D}_2{}^{T}$$

$$\boldsymbol{v}^{\circ} = (\boldsymbol{v}_1{}^{T}, \boldsymbol{v}_2{}^{T})^{T}$$

Lemma 2.1: If $\boldsymbol{X} \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, then

$$(i)\ EX = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\delta}, \quad \boldsymbol{\delta} = \frac{\boldsymbol{\Sigma\alpha}}{\sqrt{1+\boldsymbol{\alpha}^{T}\boldsymbol{\Sigma\alpha}}}.$$

$(ii)\ Cov(X) = \Sigma - \frac{2}{\pi}\delta\delta^T.$

2.4 Skew Gaussian Process for Regression and Monte Carlo Approximation

Let X_0 be a standard normal random variable, which is independent of $X(\boldsymbol{t})$, a Gaussian process with mean 0 and variance 1, Alodat and AL-Rawwash (2009) define the skew Gaussian process $Y(\boldsymbol{t})$ as follows

$$Y(\boldsymbol{t}) = \delta(\boldsymbol{t})|X_0| + \sqrt{1-\delta^2(\boldsymbol{t})}X(\boldsymbol{t}), \boldsymbol{t} \in \mathcal{C} \subseteq \mathbb{R}^p,$$

where $\delta: \mathcal{C} \to (-1,1)$ is a function which controls the skewness of the processes $Y(\boldsymbol{t})$, i.e., the skewness of the finite dimensional distribution of $Y(\boldsymbol{t})$.

It can be shown that for every $\boldsymbol{t}_1,\ \boldsymbol{t}_2,\ \boldsymbol{t}_3, \dots, \boldsymbol{t}_p \in \mathcal{C}$, the random vector $\boldsymbol{Y} = (Y(\boldsymbol{t}_1), Y(\boldsymbol{t}_2), \dots, Y(\boldsymbol{t}_p))^T$ has a multivariate skew Gaussian distribution. Hence, we may extend the Gaussian process for regression by assuming a skew normal processes on the function $f(\boldsymbol{t})$ and on $\epsilon(\boldsymbol{t})$. We will refer to the new model as skew Gaussian process for regression.

Assume that we are interested in predicting $f(\boldsymbol{t})$ at $\boldsymbol{t}^*$, where $\boldsymbol{t}^*$ is a random variable such that $\boldsymbol{t}^* \sim \mathcal{N}_p(\mu_*, \Sigma_*)$, i.e., we are interested in prediction at a random input, in this case (Girard et al. 2004), the predictive pdf for f^* given that μ_*, Σ_* is

$$p(f^*|\mu_*, \Sigma_*) = \int p(f^*|\boldsymbol{t}^*, \mathcal{D})p(\boldsymbol{t}^*)d\boldsymbol{t}^* \quad (2.3)$$

The integral in equation (2.3) does not have a closed form. Hence, a closed form approximation to this integral is needed in order to find inferential statements about f^*.

Moreover the main computational problem in Gaussian process for regression is the inversion of the matrix $\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n$ and in obtaining the mean and variance of the predictive distribution of f^* at a random input $\boldsymbol{t}^*$. Under Gaussian process for regression; we do have a closed formula for the predictive density (2.3). For this reason, the next section presents several approximation methods to approximate the predictive density (2.3) at a random input $\boldsymbol{t}^*$.

Next, we propose a Monte carlo approximation to (2.3) as follows:

If we assume that the input variable $\boldsymbol{t}^*$ has a Gaussian distribution, i.e., $\boldsymbol{t}^* \sim \mathcal{N}_p(\mu_*, \Sigma_*)$, then the predictive distribution

$$p(f^*|\mu_*, \Sigma_*, \boldsymbol{\mathcal{D}}) = \int p(f^*|\boldsymbol{t}^*, \boldsymbol{\mathcal{D}})p(\boldsymbol{t}^*)d\boldsymbol{t}^*$$

can be obtained by performing a numerical approximation of the integral, using the simple Monte-Carlo approach:

$$p(f^*|\mu_*,\Sigma_*,\mathcal{D}) = \int p(f^*|\boldsymbol{t}^*,\mathcal{D})p(\boldsymbol{t}^*)d\boldsymbol{t}^* \simeq \frac{1}{N}\sum_{r=1}^{N} p\left(f^*,\mathcal{D}|\boldsymbol{t}^{*(r)}\right),$$

where $\boldsymbol{t}^{*(1)},\ldots,\boldsymbol{t}^{*(N)}$ are independent samples from $p(\boldsymbol{t}^*)$.

Before closing this section, we refer to Girard et al. (2002) and Williams and Rassmussen and (2006) when the reader can find several analytical approximation techniques to approximate the predictive density of GPR.

CHPTER THREE

GENERLIZATION OF SKEW-GAUSSIAN PROCESS FOR NON LINEAR REGRESION

In this chapter, generalization to the skew Gaussian process and skew Gaussian process for regression are given. Also, the predictive densities at new inputs under the new processes are derived. Finally, the Gaussian process predictor is studied under the assumption that the error term $\epsilon(t)$ violates the Gaussianity.

3.1 Skew-Gaussian and Skew-White Noise Processes

In this section, we follow an alternative approach to that of section (2.4) to define as a skew Gaussian process. We rely on the set of finite dimensional distribution of a process to define it. The definition of skew Gaussian process via its finite dimensional will ease the mathematical calculations.

Definition 3.1: A random process $Y(\boldsymbol{t}), \boldsymbol{t} \in C \subseteq \mathbb{R}^p$ is said to be a skew-Gaussian process if for every $n \in \{1, 2, 3, \dots\}$ and every $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n \in C$, the vector $(Y(\boldsymbol{t}_1), \dots, Y(\boldsymbol{t}_n))^T$ follows the density (2.3), i.e., $Y(\boldsymbol{t})$ is skew-Gaussian process if its set of finite dimensional distributions is a subfamily of the family of distributions defined by (2.3).

Definition 3.2: A skew-Gaussian process $Y(\boldsymbol{t})$ possesses fixed skewness in all directions if for every n and $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n$, the parameter $\boldsymbol{\alpha}$ in (2.3) takes the form $\boldsymbol{\alpha} = \alpha \boldsymbol{1}_n,\ \alpha \in \mathbb{R}$.

We assume that for each n and $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n \in C$, the parameter $\Omega = \left(k(\boldsymbol{t}_i, \boldsymbol{t}_j)\right)_{i,j=1}^{n}$, where $k(.,.)$ is a given covariance function.

Definition 3.3: A skew-Gaussian process is called skew-white noise if for every n and $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n \in C \subseteq \mathbb{R}^n$, $\epsilon = (\epsilon(\boldsymbol{t}_1), \dots, \epsilon(\boldsymbol{t}_n))^T$ has $SN_n(\boldsymbol{0}, \tau^2 \boldsymbol{I}_n, \beta \boldsymbol{1}_n^T)$.

3.2 Joint Density of Data and Output

The aim of this section is to derive the joint density of $f^* = f(\boldsymbol{t}^*)$ and the data. For simplicity, we assume that the skew Gaussian processes used here possess fixed skew Gaussian in all directions. Since $f(\boldsymbol{t})$ is assumed to have a skew Gaussian process prior, then

$$\begin{pmatrix} \boldsymbol{f} \\ f^* \end{pmatrix} \sim CSN_{n+1,1}(\boldsymbol{0}, \boldsymbol{\Psi}, \alpha \boldsymbol{1}_{n+1}^T, 0, 1),\ \ \boldsymbol{\epsilon} \sim CSN_{n,1}(\boldsymbol{0}, \tau^2 \boldsymbol{I}_n, \beta \boldsymbol{1}_n^T, 0, 1),$$

where $\boldsymbol{1}_{n+1}$ denotes the column of one's of size $(n+1)$, and $\mathbf{I}_\mathrm{n}$ is the identity matrix of size $n \times n$. Since $\begin{pmatrix} \boldsymbol{f} \\ f^* \end{pmatrix}$ is independent of $\boldsymbol{\epsilon}(\boldsymbol{t})$ then by proposition 2.1 we have that

$$\begin{pmatrix} \boldsymbol{f} \\ f^* \\ \boldsymbol{\epsilon} \end{pmatrix} \sim CSN_{2n+1,2}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+, \boldsymbol{D}^+, \boldsymbol{v}^+, \Delta^+),$$

where

$$\boldsymbol{\mu}^+ = (\mathbf{0}_{1\times n}^T,\ 0,\ \mathbf{0}_{1\times n}^T)^T \ ,\ \boldsymbol{v}^+ = (0,0)^T \ ,\ \ \Delta^+ = \boldsymbol{I}_2\ ,$$

where $\mathbf{0}_{n\times 1}$ is the zero vector of size $n \times 1$, and

$$\boldsymbol{D}^+ = \begin{pmatrix} \alpha\mathbf{1}_{n+1}^t & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{(n+1)\times 1}^T & \beta\mathbf{1}_n^t \end{pmatrix}_{2\times(2n+1)} \ ,\ \boldsymbol{\Sigma}^+ = \begin{pmatrix} \boldsymbol{\Psi}_{(n+1)\times(n+1)} & \mathbf{0}_{(n+1)\times n} \\ \mathbf{0}_{(n+1)\times n}^T & \tau^2\boldsymbol{I}_n \end{pmatrix}.$$

The first step is to find the conditional distribution of f^* and $\boldsymbol{Y}$ is to find the joint pdf of f^* and $\boldsymbol{Y}$. To proceed, we write $(\boldsymbol{Y}^T, f^*)^T$ as a linear combination of $(\boldsymbol{f}^T\ \ f^*\ \ \boldsymbol{\epsilon}^T)^T$, i.e.,

$$\begin{pmatrix} \boldsymbol{Y} \\ f^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} + \boldsymbol{\epsilon} \\ f^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_n & \mathbf{0}_{n\times 1} & \boldsymbol{I}_n \\ \mathbf{0}_{n\times 1}^T & 1 & \mathbf{0}_{n\times 1}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{f} \\ f^* \\ \boldsymbol{\epsilon} \end{pmatrix}.$$

To simplify the notation, let $\boldsymbol{A}_{(n+1)\times(2n+1)} = \begin{pmatrix} \boldsymbol{I}_n & \mathbf{0}_{n\times 1} & \boldsymbol{I}_n \\ \mathbf{0}_{n\times 1}^T & 1 & \mathbf{0}_{n\times 1}^T \end{pmatrix}$. It is straight forward to check that the matrix $\boldsymbol{A}$ is of rank$(n+1)$. Now, we are ready to apply proposition 2.2. Hence,

$$\begin{pmatrix} \boldsymbol{Y} \\ f^* \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} \boldsymbol{f} \\ f^* \\ \boldsymbol{\epsilon} \end{pmatrix} \sim CSN_{n+1,2}(\boldsymbol{\mu}_{\boldsymbol{A}},\ \boldsymbol{\Sigma}_{\boldsymbol{A}},\ \boldsymbol{D}_{\boldsymbol{A}}, \boldsymbol{v}^+, \Delta_{\boldsymbol{A}}),$$

where

$$\boldsymbol{\mu}_A = \boldsymbol{A}\boldsymbol{\mu}^+ = \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0}_{n\times1} & \boldsymbol{I}_n \\ \boldsymbol{0}^T_{n\times1} & 1 & \boldsymbol{0}^T_{n\times1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T_{1\times n} \\ 0 \\ \boldsymbol{0}^T_{1\times n} \end{pmatrix} = \boldsymbol{0}_{(n+1)\times1}$$

$$\boldsymbol{\Sigma}_A = \boldsymbol{A}\boldsymbol{\Sigma}^+\boldsymbol{A}^T$$

$$= \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0}_{n\times1} & \boldsymbol{I}_n \\ \boldsymbol{0}^T_{n\times1} & 1 & \boldsymbol{0}^T_{n\times1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}_{(n+1)\times(n+1)} & \boldsymbol{0}_{(n+1)\times n} \\ \boldsymbol{0}^T_{(n+1)\times n} & \tau^2\boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0}_{n\times1} \\ \boldsymbol{0}^T_{n\times1} & 1 \\ \boldsymbol{I}_n & \boldsymbol{0}_{n\times1} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n & \boldsymbol{k} \\ \boldsymbol{k}^T & k^* \end{pmatrix}_{(n+1)\times(n+1)}.$$

To proceed, we need to apply the following matrix identity. Let $\boldsymbol{A}$ be a matrix which is partitioned as follows:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix},$$

where $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are invertible square matrices. Then

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \left(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right)^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\left(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\right)^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\left(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right)^{-1} & \left(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\right)^{-1} \end{pmatrix}.$$

For proof, we refer to Schott (1997). Hence, we find $\boldsymbol{\Sigma}_A^{-1}$ as follow:

$$\boldsymbol{\Sigma}_A^{-1}$$

$$= \begin{pmatrix} \left(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n - \boldsymbol{k}k^{*-1}\boldsymbol{k}^T\right)^{-1} & -(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k}(k^* - \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k})^{-1} \\ -k^{*-1}\boldsymbol{k}^T\left(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n - \boldsymbol{k}k^{*-1}\boldsymbol{k}^T\right)^{-1} & (k^* - \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k})^{-1} \end{pmatrix}_{(n+1)\times(n+1)}.$$

The parameter $\boldsymbol{D}_A$ is given by

$$\boldsymbol{D_A = D^{+}\Sigma^{+}A^{T}\Sigma_A^{-1}},$$

$$= \begin{pmatrix} \alpha\mathbf{1}_{n+1}^{t} & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n}^{T} & \beta\mathbf{1}_{n}^{t} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}_{(n+1)\times(n+1)} & \mathbf{0}_{(n+1)\times n} \\ \mathbf{0}_{(n+1)\times n}^{T} & \tau^2 \boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{n\times 1}^{T} & 1 \\ \boldsymbol{I}_n & \mathbf{0}_{n\times 1} \end{pmatrix} \boldsymbol{\Sigma}_A^{-1}$$

$$= \begin{pmatrix} \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma},\boldsymbol{k})^{T} & \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T},k^{*})^{T} \\ \beta\tau^2\mathbf{1}_n^{T} & 0 \end{pmatrix} \boldsymbol{\Sigma}_A^{-1},$$

$$= \begin{pmatrix} \boldsymbol{D}_{11(1\times n)} & D_{12} \\ \boldsymbol{D}_{21(1\times n)} & D_{22} \end{pmatrix}_{2\times(n+1)},$$

where

$$\boldsymbol{D}_{11} = \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma},\boldsymbol{k})^{T}\left(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n - \boldsymbol{k}k^{*^{-1}}\boldsymbol{k}^{T}\right)^{-1} -$$

$$\alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T},k^{*})^{\mathrm{T}}k^{*^{-1}}\boldsymbol{k}^{T}\left(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n - \boldsymbol{k}k^{*^{-1}}\boldsymbol{k}^{T}\right)^{-1},$$

$$D_{12} = -\,\alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma},\boldsymbol{k})^{T}(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k})^{-1} +$$

$$\alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T},k^{*})^{T}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k})^{-1},$$

$$\boldsymbol{D}_{21} = \beta\tau^2\mathbf{1}_n^{T}\left(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n - \boldsymbol{k}k^{*^{-1}}\boldsymbol{k}^{T}\right)^{-1}$$

and

$$D_{22} = -\,\beta\tau^2\mathbf{1}_n^{T}(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k})^{-1}.$$

Also the parameter $\Delta_{\boldsymbol{A}} = \boldsymbol{\Delta}^{+} + \mathbf{D}^{+}\boldsymbol{\Sigma}^{+}\mathbf{D}^{+^{T}} - \mathbf{D}^{+}\boldsymbol{\Sigma}^{+}\boldsymbol{A}^{T}\boldsymbol{\Sigma}_{\boldsymbol{A}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{+}\mathbf{D}^{+^{T}}$,

where $\boldsymbol{\Delta}^{+} = \boldsymbol{I}_2$, can be simplified as follows:

$$\mathbf{D}^{+}\boldsymbol{\Sigma}^{+}\mathbf{D}^{+^{T}} = \begin{pmatrix} \alpha\mathbf{1}_{n+1}^{T}\boldsymbol{\Psi}\mathbf{1}_{\mathrm{n+1}} & 0 \\ 0 & n\beta^2\tau^2 \end{pmatrix}_{2\times 2},$$

$$\mathbf{D}^{+}\boldsymbol{\Sigma}^{+}\boldsymbol{A}^{T} = \begin{pmatrix} \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma},\boldsymbol{k})^{T} & \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T},k^{*})^{T} \\ \beta\tau^2\mathbf{1}_{n}^{T} & 0 \end{pmatrix}_{2\times(n+1)},$$

and

$$\boldsymbol{A}\boldsymbol{\Sigma}^{+}\mathbf{D}^{+^{T}} = \begin{pmatrix} \alpha(\boldsymbol{\Sigma},\boldsymbol{k})\mathbf{1}_{n+1} & \beta\tau^2\mathbf{1}_{n} \\ \alpha(\boldsymbol{k}^{T},k^{*})\mathbf{1}_{n+1} & 0 \end{pmatrix}_{(n+1)\times 2}.$$

Finally $\Delta_{\boldsymbol{A}}$ takes the following form

$$\Delta_{\boldsymbol{A}} = \boldsymbol{I}_2 + \begin{pmatrix} \alpha\mathbf{1}_{n+1}^{T}\boldsymbol{\Psi}\mathbf{1}_{n+1} & 0 \\ 0 & n\beta^2\tau^2 \end{pmatrix}$$

$$- \begin{pmatrix} \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma},\boldsymbol{k})^{T} & \alpha\mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T},k^{*})^{T} \\ \beta\tau^2\mathbf{1}_{n}^{T} & 0 \end{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{A}}^{-1} \begin{pmatrix} \alpha(\boldsymbol{\Sigma},\boldsymbol{k})\mathbf{1}_{n+1} & \beta\tau^2\mathbf{1}_{n} \\ \alpha(\boldsymbol{k}^{T},k^{*})\mathbf{1}_{n+1} & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 + \alpha\mathbf{1}_{n+1}^{T}\boldsymbol{\Psi}\mathbf{1}_{n+1} - W_{11} & -W_{12} \\ -W_{21} & 1 + n\beta^2\tau^2 - W_{22} \end{pmatrix},$$

where

$$W_{11} = \Big(\alpha \mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma}, \boldsymbol{k})^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}$$

$$- \alpha \mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T}, k^{*})^{T}k^{*-1}\boldsymbol{k}^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n}$$

$$- \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}\Big)(\alpha(\boldsymbol{\Sigma}, \boldsymbol{k})\mathbf{1}_{n+1})$$

$$+ (-\alpha \mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma}, \boldsymbol{k})^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k})^{-1}$$

$$+ \alpha \mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T}, k^{*})^{T}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k})^{-1})(\alpha(\boldsymbol{k}^{T}, k^{*})\mathbf{1}_{n+1})$$

$$= \alpha^{2}\Big(\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma}, \boldsymbol{k})^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}$$

$$- \mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T}, k^{*})^{T}k^{*-1}\boldsymbol{k}^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}\Big)(\boldsymbol{\Sigma}, \boldsymbol{k})\mathbf{1}_{n+1}$$

$$+ \alpha^{2}(-\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma}, \boldsymbol{k})^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k})^{-1}$$

$$+ \mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T}, k^{*})^{T}(k^{*} - \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k})^{-1})(\boldsymbol{k}^{T}, k^{*})\mathbf{1}_{n+1}$$

$$W_{12} = \Big(\alpha \mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma}, \boldsymbol{k}^{T})^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}$$

$$- \alpha \mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T}, k^{*})^{T}k^{*-1}\boldsymbol{k}^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}\Big)\beta\tau^{2}\mathbf{1}_{n}$$

$$= \alpha\beta\tau^{2}\Big(\mathbf{1}_{n+1}^{T}(\boldsymbol{\Sigma}, \boldsymbol{k}^{T})^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}$$

$$- \mathbf{1}_{n+1}^{T}(\boldsymbol{k}^{T}, k^{*})^{T}k^{*-1}\boldsymbol{k}^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}\Big)\mathbf{1}_{n}$$

$$W_{21} = \beta\tau^{2}\mathbf{1}_{n}^{T}\big(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n} - \boldsymbol{k}k^{*-1}\boldsymbol{k}^{T}\big)^{-1}\alpha(\boldsymbol{\Sigma}, \boldsymbol{k})\mathbf{1}_{n+1}$$

$$- \beta\tau^{2}\mathbf{1}_{n}^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k}(k^{*}$$

$$- \boldsymbol{k}^{T}(\boldsymbol{\Sigma} + \tau^{2}\boldsymbol{I}_{n})^{-1}\boldsymbol{k})^{-1}\alpha(\boldsymbol{k}^{T}, k^{*})\mathbf{1}_{n+1}$$

$$= \alpha\beta\tau^2 \left(\mathbf{1}_n^T \left(\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n - \boldsymbol{k}k^{*-1}\boldsymbol{k}^T \right)^{-1} (\boldsymbol{\Sigma}, \boldsymbol{k}) \mathbf{1}_{n+1} \right.$$

$$\left. - \tau^2 \mathbf{1}_n^T (\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n)^{-1} \boldsymbol{k} (k^* - \boldsymbol{k}^T (\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n)^{-1} \boldsymbol{k})^{-1} (\boldsymbol{k}^T, k^*) \mathbf{1}_{n+1} \right)$$

$$W_{22} = n\beta^2\tau^4 \mathbf{1}_n^T \left(\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n - \boldsymbol{k}k^{*-1}\boldsymbol{k}^T \right)^{-1} \mathbf{1}_n \ .$$

3.3 The predictive density at fixed input

The predictive density of f^* given $\boldsymbol{Y}$ is obtained by direct application of proposition 2.3 with $p = n + 1$ and $q = 2$.

To proceed, consider the following partitions for $\boldsymbol{\mu}_A$, $\boldsymbol{\Sigma}_A$, $\boldsymbol{D}_A$, $\boldsymbol{v}^+$, Δ_A :

$$\boldsymbol{\Sigma}_A = \begin{pmatrix} \boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n & \boldsymbol{k} \\ \boldsymbol{k}' & k^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \Sigma_{22} \end{pmatrix},$$

$$\Delta_A = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$$

and

$$\boldsymbol{D}_A = \begin{pmatrix} \boldsymbol{D}_{11} & D_{12} \\ \boldsymbol{D}_{21} & D_{22} \end{pmatrix} = (\boldsymbol{D}_1 \quad \boldsymbol{D}_2),$$

where

$\boldsymbol{D}_1 = \begin{pmatrix} \boldsymbol{D}_{11} \\ \boldsymbol{D}_{21} \end{pmatrix}_{2\times n}$ and $\boldsymbol{D}_2 = \begin{pmatrix} D_{12} \\ D_{22} \end{pmatrix}_{2\times 1}$. So the conditional distribution of f^* given $\boldsymbol{Y}$ is

$$f^*|Y{\sim}CSN_{1,2}(\boldsymbol{k}^T(\boldsymbol{\Sigma}+\tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{Y},\ k^*-\boldsymbol{k}^T(\boldsymbol{\Sigma}+\tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k},\ \boldsymbol{D}_2, -\boldsymbol{D}^*\boldsymbol{Y}, \Delta_A), \tag{3.1}$$

where

$$\boldsymbol{D}^* = \boldsymbol{D}_1 + \boldsymbol{D}_2\boldsymbol{k}^T(\boldsymbol{\Sigma}+\tau^2\boldsymbol{I}_n)^{-1}.$$

The above analysis shows that the predictive distribution of a new output follows a closed skew Gaussian distribution. As a special case, this predictive distribution reduces to (2.2) if the skewness is absent, i.e., if $\alpha = \beta = 0$.

The mean of the predictive distribution can serve as a predictor for $f(t^*)$ while its variance $var(f^*|Y)$ can be used as a measure of uncertainty of this predictor.

Another predictor of $f(\boldsymbol{t}^*)$ is the median of the conditional distribution of f^* given $\boldsymbol{Y}$.

Neither mean nor the median of the conditional distribution in our case have simple closed form. In the next section, we derive formulas for the predictive mean and variance based on the moment generating function.

3.4 Predictive Mean and Variance

Here we have to find the mean and the variance of $f^*|\boldsymbol{Y}$ by applying Proposition 2.4; to complete this mission, we find the moment generating function of $f^*|\boldsymbol{Y}$, hence the moment generating function of $f^*|\boldsymbol{Y}$ is equal to

$$M_{f^*|Y}(s) = \frac{\Phi_2(\boldsymbol{D}_2\sigma^{*2}s; -\boldsymbol{D}^*\boldsymbol{Y}, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T)}{\Phi_2(\boldsymbol{D}_2\sigma^{*2}s; -\boldsymbol{D}^*\boldsymbol{Y}, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T)} e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}, \quad s \in \mathbb{R}$$

where $\sigma^{*2} = k^* - \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{k},$ and $\mu^* = \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{Y}.$ Let $\Phi_2^{(j)}(.,.)$ denote the first prtial derivative of $\Phi_2(.,.)$ with respect to the j^{th} component for $j = 1, 2$. Also, let $\Phi_2^{(ij)}(.,.)$ denote the mixed second partial derivative of Φ_2.

Now we have to find the mean and the variance of $f^*|\boldsymbol{Y}$ as

$$\mathbb{E}(f^*|Y) = \frac{\partial}{\partial s} M_{f^*|Y}(s)\;|_{s=0}$$

$$= \frac{\partial}{\partial s}\left(\frac{\Phi_2(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T)}{\Phi_2(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T)} e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}\right)|_{s=0}$$

$$= \frac{\partial}{\partial s}\left(\frac{\Phi_2\left(\begin{pmatrix} D_{12}\sigma^{*2}s \\ D_{22}\sigma^{*2}s \end{pmatrix};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\right)}{\Phi_2(\boldsymbol{0}_{2\times 1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T)} e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}\right)|_{s=0}$$

$$= \left(\frac{e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}}{\Phi_2\left(\mathbf{0}_{2\times1};\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right)} \left(\Phi_2^{(1)}\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A \right.\right.\right.$$

$$\left. + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right) D_{12}\sigma^{*2}$$

$$\left.\left. + \Phi_2^{(2)}\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right) D_{22}\sigma^{*2}\right)\right)|_{s=0}$$

$$+ \left(\frac{\Phi_2\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right)}{\Phi_2\left(\mathbf{0}_{2\times1};\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right)} (\mu^* \right.$$

$$\left. + \ \sigma^{*2}s)e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}\right)|_{s=0}$$

$$= \left(\ D_{12}\sigma^{*2}\Phi_2^{(1)}\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right)\right.$$

$$\left. + \ D_{22}\sigma^{*2}\Phi_2^{(2)}\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right)\right) + \mu^*$$

$$=$$

$$\mu^* + \sigma^{*2}\left(\ D_{12}\Phi_2^{(1)}\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \ \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T\right) + \ D_{22}\Phi_2^{(2)}\left(\ \boldsymbol{D}_2\sigma^{*2}s;\ -\boldsymbol{D}^*\boldsymbol{Y},\ \Delta_A + \right.\right.$$

$$\left.\left. \sigma^{*2}\boldsymbol{D}_2\ \boldsymbol{D}_2{}^T)\right)\right),$$

where $\Phi_2^{(1)}$ is the first derivative of Φ_2 with respect to the first component, and $\Phi_2^{(2)}$ is the first derivative of Φ_2 with respect to the second component.

Also we need to find $\mathbb{E}\left(f^{*2}|\boldsymbol{Y}\right)$ to calculate the variance of $f^*|\boldsymbol{Y}$.

So

$$\mathbb{E}\left(f^{*2}|\boldsymbol{Y}\right) = \frac{\partial^2}{\partial s^2}\left(\frac{\Phi_2\left(\begin{pmatrix} D_{12}\sigma^{*2}s \\ D_{22}\sigma^{*2}s \end{pmatrix}; -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)}{\Phi_2\left(\boldsymbol{0}_{2\times1}; -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)} e^{s\mu^{*}+\frac{1}{2}\sigma^{*2}s^2}\right)|_{s=0}$$

$$= \frac{\partial}{\partial s}\Bigg(\Bigg(\frac{e^{s\mu^{*}+\frac{1}{2}\sigma^{*2}s^2}}{\Phi_2\left(\boldsymbol{0}_{2\times1}; -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)}\Big(\Phi_2^{(1)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A$$

$$+\ \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\big)\, D_{12}\sigma^{*2}$$

$$+\, \Phi_2^{(2)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\big)\, D_{22}\sigma^{*2}\Big)\Bigg)$$

$$+\left(\frac{\Phi_2\left(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)}{\Phi_2\left(\boldsymbol{0}_{2\times1}; -\boldsymbol{D}^{*}\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)}\right.(\mu^{*}$$

$$+\, \sigma^{*2}s)e^{s\mu^{*}+\frac{1}{2}\sigma^{*2}s^2}\Big)\Bigg)|_{s=0}$$

$$= \Bigg(\frac{\partial}{\partial s} \Bigg(\frac{1}{\Phi_2\left(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\right)} \Big(\Phi_2^{(1)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A$$

$$+\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}$$

$$+\, \Phi_2^{(2)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\big)\, D_{22}\sigma^{*2}\big) \Bigg) e^{s\mu^*+\frac{1}{2}\sigma^{*2}s^2}$$

$$+ \Bigg(\frac{1}{\Phi_2\left(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\right)} \Big(\Phi_2^{(1)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A$$

$$+\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}$$

$$+\, \Phi_2^{(2)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\big)\, D_{22}\sigma^{*2}\big) \Bigg) (\mu^*$$

$$+\, \sigma^{*2}s) e^{s\mu^*+\frac{1}{2}\sigma^{*2}s^2}$$

$$+ \frac{\partial}{\partial s} \left(\frac{\Phi_2\left(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\right)}{\Phi_2\left(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\right)} \right) \left((\mu^* \,+ \sigma^{*2}s) e^{s\mu^*+\frac{1}{2}\sigma^{*2}s^2} \right)$$

$$+ \left(\frac{\Phi_2\left(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\right)}{\Phi_2\left(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +\, \sigma^{*2}\boldsymbol{D}_2\, \boldsymbol{D}_2{}^T\right)} \right) \Big((\mu^* \,+ \sigma^{*2}s)^2 e^{s\mu^*+\frac{1}{2}\sigma^{*2}s^2}$$

$$+\, \sigma^{*2} e^{s\mu^*+\frac{1}{2}\sigma^{*2}s^2} \Big) \Bigg) |_{s=0}$$

$$= \Bigg(\Bigg(\frac{e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}}{\Phi_2\big(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)}\Bigg)\Big(\Phi_2^{(11)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A$$

$$+\ \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\big(\,D_{12}\sigma^{*2}\big)^2$$

$$+\,\Phi_2^{(12)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}\ D_{22}\sigma^{*2}$$

$$+\,\Phi_2^{(21)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}\ D_{22}\sigma^{*2}$$

$$+\,\Phi_2^{(22)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\big(\,D_{22}\sigma^{*2}\big)^2\Big)$$

$$+\Bigg(\frac{1}{\Phi_2\big(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)}\Big(\Phi_2^{(1)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A$$

$$+\ \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}$$

$$+\,\Phi_2^{(2)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{22}\sigma^{*2}\Big)\Bigg)(\mu^*$$

$$+\,\sigma^{*2}s)e^{s\mu^* + \frac{1}{2}\sigma^{*2}s^2}$$

$$+\Bigg(\frac{1}{\Phi_2\big(\mathbf{0}_{2\times1};\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)}\Big(\Phi_2^{(1)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A$$

$$+\ \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}$$

$$+\,\Phi_2^{(2)}\big(\,\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{22}\sigma^{*2}\Big)\Bigg)\big((\mu^*$$

$$+\sigma^{*2}s)e^{s\mu^{*}+\frac{1}{2}\sigma^{*2}s^{2}}\Big)$$

$$+\left(\frac{\Phi_2\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)}{\Phi_2\left(\boldsymbol{0}_{2\times 1};\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)}\right)\Big(\big(\mu^{*}$$

$$+\sigma^{*2}s)^{2}e^{s\mu^{*}+\frac{1}{2}\sigma^{*2}s^{2}}+\sigma^{*2}e^{s\mu^{*}+\frac{1}{2}\sigma^{*2}s^{2}}\Big)\Bigg)\Big|_{s=0}$$

$$=$$

$$2\left(\Phi_2^{(11)}\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)\left(D_{12}\sigma^{*2}\right)^{2}+\right.$$

$$\Phi_2^{(12)}\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)D_{12}\sigma^{*2}\,D_{22}\sigma^{*2}+$$

$$\Phi_2^{(21)}\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)D_{12}\sigma^{*2}\,D_{22}\sigma^{*2}+$$

$$\left.\Phi_2^{(22)}\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)\left(D_{22}\sigma^{*2}\right)^{2}\right)+$$

$$4\left(\Phi_2^{(1)}\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)D_{12}\sigma^{*2}\quad+\right.$$

$$\left.\Phi_2^{(2)}\left(\boldsymbol{D}_2\sigma^{*2}s;\,-\boldsymbol{D}^{*}\boldsymbol{Y},\,\Delta_A+\,\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^{T}\right)D_{22}\sigma^{*2}\right)\mu^{*}+\mu^{*\,2}+\sigma^{*2},$$

hence

$$var(f^{*}|\boldsymbol{Y})=\mathbb{E}\left(f^{*2}|\boldsymbol{Y}\right)-\left(\mathbb{E}(f^{*}|\boldsymbol{Y})\right)^{2}$$

$$= 2\Big(\Phi_2^{(11)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\big(D_{12}\sigma^{*2}\big)^2$$

$$+ \Phi_2^{(12)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}\, D_{22}\sigma^{*2}$$

$$+ \Phi_2^{(21)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}\, D_{22}\sigma^{*2}$$

$$+ \Phi_2^{(22)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\big(D_{22}\sigma^{*2}\big)^2\Big)$$

$$+ 4\Big(\Phi_2^{(1)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\sigma^{*2}$$

$$+ \Phi_2^{(2)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{22}\sigma^{*2}\Big)\mu^* + \mu^{*\,2} + \sigma^{*2}$$

$$-\Big(\mu^*$$

$$+ \sigma^{*2}\Big(D_{12}\Phi_2^{(1)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)$$

$$+ D_{22}\Phi_2^{(2)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\Big)\Big)^2$$

$$= 2\sigma^{*4}\Big(\Phi_2^{(11)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}{}^2 + \Phi_2^{(12)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A +$$

$$\sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\, D_{22} + \Phi_2^{(21)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{12}\, D_{22} +$$

$$\Phi_2^{(22)}\big(\boldsymbol{D}_2\sigma^{*2}s;\, -\boldsymbol{D}^*\boldsymbol{Y},\, \Delta_A + \sigma^{*2}\boldsymbol{D}_2\,\boldsymbol{D}_2{}^T\big)\, D_{22}{}^2\Big),$$

where $\Phi_2^{(11)}$ is the derivative of $\Phi_2^{(1)}$ with respect to the first component, and $\Phi_2^{(12)}$ is the derivative of $\Phi_2^{(1)}$ with respect to the second component, and $\Phi_2^{(21)}$ is the derivative of $\Phi_2^{(2)}$ with respect to the first component, and $\Phi_2^{(22)}$ is the derivative of $\Phi_2^{(2)}$ with respect to the second component.

3.5 Gaussian Process for Regression under Skew-Normal Errors

In this section, it is assumed that the error term $\boldsymbol{\epsilon}(\mathbf{t})$ follows a skew white noise. Under this assumption, we study the effect of this assumption on the mean and mean squared error of the GPR predictor. To proceed, let $\hat{f} = \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{Y}$. In the sequel, the mean, the variance and the bias of $\hat{f}$ where $\boldsymbol{Y}$ is replaced by $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \tau^2\boldsymbol{I}_n)$ are denoted by $E^G(\hat{f})$, $var^G(\hat{f})$ and $bias^G(\hat{f})$, respectively. Also if Y is replaced by $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim SN_n(\mathbf{0}, \tau^2\boldsymbol{I}_n,\ \beta\mathbf{1}_n)$, then the mean, the variance and the bias are replaced by $E^{SG}(\hat{f})$, $var^{SG}(\hat{f})$ and $bias^{SG}(\hat{f})$, respectively. Under white noise, i.e., $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \tau^2\boldsymbol{I}_n)$ we have

$$E^G(\hat{f}) = \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}E(\boldsymbol{f} + \boldsymbol{\epsilon}),$$

$$= \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{f}.$$

While under the assumption $\boldsymbol{\epsilon} \sim SN_n(\mathbf{0}, \tau^2\boldsymbol{I}_n,\ \beta\mathbf{1}_n)$,

$$E^{SG}(\hat{f}) = \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}E(\boldsymbol{f} + \boldsymbol{\epsilon}),$$

$$= \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{f} + \boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}E\boldsymbol{\epsilon}.$$

Since $\boldsymbol{\epsilon} \sim SN_n(\mathbf{0}, \tau^2\boldsymbol{I}_n,\ \beta\mathbf{1}_n)$, then

$$E(\epsilon) = \sqrt{\frac{2}{\pi}}\frac{\tau^2\,\beta \mathbf{1}_n}{\sqrt{1+\beta^2\tau^2 n}}.$$

Hence

$$E^{SG}(\hat{f}) = E^{G}(\hat{f}) + \boldsymbol{k}^T(\boldsymbol{\Sigma}+\tau^2\boldsymbol{I}_n)^{-1}\sqrt{\frac{2}{\pi}}\frac{\tau^2\,\beta \mathbf{1}_n}{\sqrt{1+\beta^2\tau^2 n}}\ ,$$

$$= E^{G}(\hat{f}) + b(\tau^2,\beta^2,n)\ \text{, say.} \qquad (2.4)$$

From the last equation, we conclude that the *GPR* predictor is increased or decreased by an amount of $|b(\tau^2,\beta^2,n)|$. We set the following theorem now more about the properties of the term (τ^2,β^2,n) :

Theorem 3.1 :

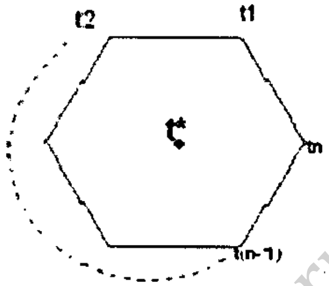
The term $b(\tau^2,\beta^2,n) = \boldsymbol{k}^T(\boldsymbol{\Sigma}+\tau^2\boldsymbol{I}_n)^{-1}\sqrt{\frac{2}{\pi}}\frac{\tau^2\,\beta \mathbf{1}_n}{\sqrt{1+\beta^2\tau^2 n}}$, has the following properties:

(i) $\lim_{\beta\to 0} b(\tau^2,\beta^2,n) = 0$.

(ii) $\lim_{\beta\to\pm\infty}|b(\tau^2,\beta^2,n)| = \sqrt{\frac{2}{\pi}}\,\frac{\tau}{\sqrt{n}}\,\boldsymbol{k}^T(\boldsymbol{\Sigma}+\tau^2\boldsymbol{I}_n)^{-1}\mathbf{1}_n$.

(iii) $\lim_{\tau\to 0} b(\tau^2,\beta^2,n) = 0$.

(iv) Assume that $\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_n$ are chosen so that they are the vertices of a regular polygon and $\boldsymbol{t}^*$ is located at its center. If $k(.,.)$ is an isotropic covariance function, then $|b(\tau^2, \beta^2, n)| > 0$ for all τ, n and $\beta \neq 0$, and $\lim_{\tau\to\infty} b(\tau^2, \beta^2, n) = 0$.



Moreover, if $\sum_{i=1}^{n} k(t_1, t_i) = n^{0.5}O(n)$, with $O(n) \to c \neq 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} b(\tau^2, \beta^2, n) = \sqrt{\frac{2}{\pi}}\frac{\tau\,\beta k_0}{|\beta|c}.$$

Proof : The proof of (i), (ii) and (iii) is easy so we leave it to the reader. To prove (iv), let $k_0 = k(t_i, t^*)$, since $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$ are vertices of regular polygon and $k(.,.)$ is isotropic, then the matrix $\Sigma = \left(k(\boldsymbol{t}_i, \boldsymbol{t}_j)\right)_{i,j=1}^{n}$ is circulant. Also $\boldsymbol{k} = \boldsymbol{k}(\boldsymbol{t}^*) = k_0\boldsymbol{1}_n$. moreover, the matrix $(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}$ is also circulant.

Therefore

$$b(\tau^2, \beta^2, n) = \sqrt{\frac{2}{\pi}}\frac{\tau^2\,\beta}{\sqrt{1+\beta^2\tau^2 n}}\boldsymbol{k}^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{1}_n,$$

$$= \sqrt{\frac{2}{\pi}}\frac{\tau^2\,\beta k_0}{\sqrt{1+\beta^2\tau^2 n}}\boldsymbol{1}_n^T(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{1}_n.$$

Since $(\mathbf{\Sigma} + \tau^2 \boldsymbol{I}_n)^{-1}$ is also circulant and $\mathbf{1}_n$ is an eigen vector of any circulant matrix, then

$$(\mathbf{\Sigma} + \tau^2 \boldsymbol{I}_n)^{-1}\mathbf{1}_n = \frac{1}{L_n}\mathbf{1}_n,$$

where

$$L_n = \tau^2 + \sum_{i=1}^{n} k(\boldsymbol{t}_1, \boldsymbol{t}_i).$$

Hence

$$b(\tau^2, \beta^2, n) = \sqrt{\frac{2}{\pi}} \frac{\tau^2 \, \beta k_0}{\sqrt{1 + \beta^2\tau^2 n}} \frac{n}{L_n}.$$

It is easy to see that $|b(\tau^2, \beta^2, n)| > 0$ for all non zero values of τ, β and k_0.

If $\sum_{i=1}^{n} k(\boldsymbol{t}_1, \boldsymbol{t}_i) = n^{0.5} O(n)$, where $O(n) \to c \neq 0$, then

$$b(\tau^2, \beta^2, n) = \sqrt{\frac{2}{\pi}} \frac{\tau^2 \, \beta k_0}{\sqrt{1+\beta^2\tau^2 n}} \frac{n}{\tau^2 + n^{0.5} O(n)},$$

$$= \sqrt{\frac{2}{\pi}} \frac{\tau^2 \, \beta k_0}{\sqrt{1+\beta^2\tau^2 n}} \frac{\sqrt{n}}{\frac{\tau^2}{\sqrt{n}} + O(n)}.$$

Hence

$$\lim_{n\to\infty} b(\tau^2, \beta^2, n) = \sqrt{\frac{2}{\pi}} \frac{\tau^2 \, \beta k_0}{\sqrt{\beta^2\tau^2}} \frac{1}{c}$$

$$= \sqrt{\frac{2}{\pi}}\,\frac{\tau\beta k_0}{c|\beta|}.$$

To show that $\lim_{\tau\to\infty} b(\tau^2,\beta^2,n) = 0$, we first note that $(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\mathbf{1}_n = \frac{1}{L_n}\mathbf{1}_n$. Hence

$$b(\tau^2,\beta^2,n) = \sqrt{\frac{2}{\pi}}\,\frac{\tau^2\,\beta k_0}{\sqrt{1+\beta^2\tau^2 n}}\,\frac{n}{L_n}$$

$$= \sqrt{\frac{2}{\pi}}\,\beta k_0 n\,\frac{\tau^2}{\left(\tau^2+\sum_{i=1}^{n} k(t_1,t_i)\right)\sqrt{1+\beta^2\tau^2 n}}.$$

Hence $\lim_{\tau\to\infty} b(\tau^2,\beta^2,n) = 0$.

3.6 Prediction at Random Input

In this section, we assume that the input vector $\boldsymbol{t}^*$has a normal distribution and we wish to predict $f^* = f(\boldsymbol{t}^*)$. Since $f^*|\boldsymbol{Y},\boldsymbol{t}^* \sim CSN_{1,2}\left(\mu^*,\ \sigma^{*2},\ \boldsymbol{D}_2, -\boldsymbol{D}^*\boldsymbol{Y}, \Delta_A\right)$ and $\boldsymbol{t}^* \sim N_n(\boldsymbol{a},\ \boldsymbol{B})$, then using the total probability law, we write the predictive density of f^* given $\boldsymbol{Y}$ as follows:

$$p(f^*|\boldsymbol{Y}) = \int_{\mathbb{R}^n} p(f^*|\boldsymbol{Y},\boldsymbol{t}^*)p(\boldsymbol{t}^*)d\,\boldsymbol{t}^*.$$

It is difficult, even for GPR, to find a closed form for the integral in the last equation, so an approximation for $p(f^*|\boldsymbol{Y})$ is needed. Here, we propose the following Monte Carlo approximation for the predictive distribution at random input:

$$p(f^*|\boldsymbol{Y}) = \int_{\mathbb{R}^n} p(f^*|\boldsymbol{Y}, \boldsymbol{t}^*)p(\boldsymbol{t}^*)d\,\boldsymbol{t}^* \cong \frac{1}{N}\sum_r^N p\left(f^*|\boldsymbol{Y}, \boldsymbol{t}^{*(r)}\right),$$

where $\boldsymbol{t}^{*(1)}, \dots, \boldsymbol{t}^{*(N)}$ are independent samples from $p(\boldsymbol{t}^*)$.

CHAPTER FOUR

SIMUULATION STUDY

In this chapter, we present an algorithm to simulate a realization from a skew-Gaussian process, i.e. by simulation from its finite dimensional distributions. Then the algorithm is implemented in a Matlab code to simulate from a GPR and a SGPR predictors.

4.1 Simulation from $SN_n(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$.

Simulation of a sample path from a skew Gaussian process can be achieved by sampling from a multivariate skew normal distribution on a smooth grid. To simulate a random vector from the pdf

$$f(\boldsymbol{x}; \mathbf{0}, \boldsymbol{\Sigma}) = 2\phi_n(\boldsymbol{x}; \mathbf{0}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^T \boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

we may employ the accept-reject method. The accept-reject method as given in Chrestian and Casella (2004) assumes that the pdf $f(\boldsymbol{x})$ can be written as

$$f(\boldsymbol{x}) = cg(\boldsymbol{x})h(\boldsymbol{x}),$$

where $c \geq 1, 0 < g(\boldsymbol{x}) \leq 1, \forall \boldsymbol{x}$ and $h(\boldsymbol{x})$ is a pdf. If this is the case, then a random observation from $f(\boldsymbol{x})$ is generated as follows:

1. Generate U from $u(0,1)$.
2. Generate $\boldsymbol{Y}$ from $h(\boldsymbol{x})$.

3. If $U \leq g(\boldsymbol{Y})$, then deliver $\boldsymbol{Y}$ as a realization of $f(\boldsymbol{x})$.

4. Go to step 1.

For the $SN_n(\mathbf{0}; \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ distribution, we may use this algorithm with $c = 2$, $g(\boldsymbol{x}) = \Phi(\boldsymbol{\lambda}^T \boldsymbol{x})$ and $h(\boldsymbol{x}) = \phi_n(\boldsymbol{x}; \mathbf{0}, \boldsymbol{\Sigma})$.

4.2 Simulation from $CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{v}, \Delta)$

To simulate a random observation from the $CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{v}, \Delta)$, it is difficult to achieve this via the accept-reject method due to the complexity of calculating $g(\boldsymbol{x}) = \Phi_q(\boldsymbol{D}^T(\boldsymbol{Y} - \boldsymbol{\mu}); \boldsymbol{v}, \boldsymbol{\Delta})$. Instead, we employ the following algorithm which is derived from the definition of the CSN distribution (Genton, 2004; Allard and Naveau, 2007).

(i) Simulate an observation from

$$\boldsymbol{U} = N_q(\boldsymbol{v}, \boldsymbol{\Delta} + \boldsymbol{D}^T \boldsymbol{\Sigma} \boldsymbol{D}) |\ \boldsymbol{U} \leq \mathbf{0}.$$

(ii) Given $\boldsymbol{U}$, simulate $\boldsymbol{Z}$ from

$$N_p(-\boldsymbol{\Sigma} \boldsymbol{D}(\boldsymbol{\Delta} + \boldsymbol{D}^T \boldsymbol{\Sigma} \boldsymbol{D})^{-1}(\boldsymbol{U} - \boldsymbol{v}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma} \boldsymbol{D}(\boldsymbol{\Delta} + \boldsymbol{D}^T \boldsymbol{\Sigma} \boldsymbol{D})^{-1} \boldsymbol{D}^T \boldsymbol{\Sigma}).$$

(iii) Deliver $\boldsymbol{Z}$ from $CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}, \boldsymbol{v}, \Delta)$.

Also, simulation from $\boldsymbol{U} | \boldsymbol{U} \leq \mathbf{0}$ is not an easy task, so an accept-reject method will be implemented.

4.3 Simulation results

In this simulation work, a realization of the sample path of the skew Gaussian process is generated for input function $f(t) = \frac{\sin(t)}{t}, t \neq 0$. Then the simulated data are substituted in both Gaussian and skew Gaussian predictors. To see effect of the departure from Gaussianity on the Gaussian predictor, we plot the distribution function for the two predictors. Figures (4.1) – (4.8) shows these distribution functions for different values of α, β and τ.

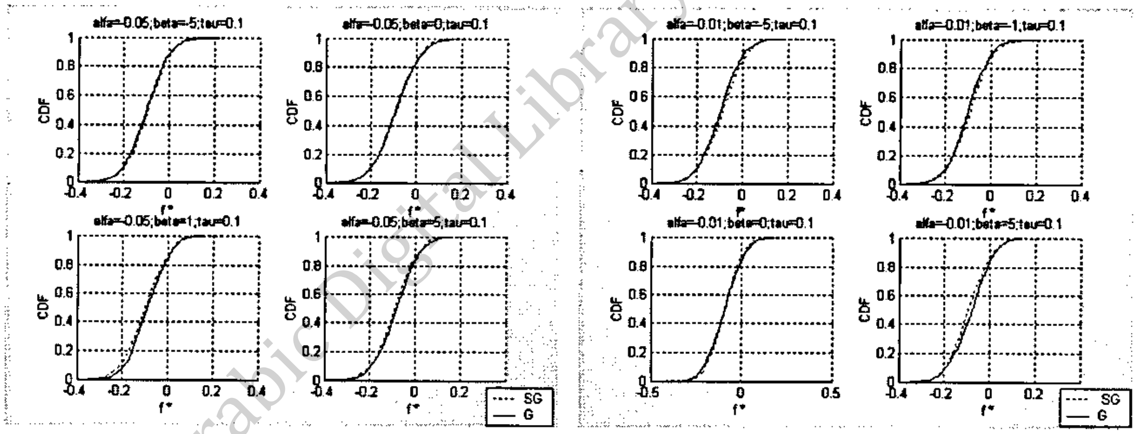


Figure (4.1):

(a) G and SG predictors with Parameters $\alpha = -0.05, \beta = -5, 0, 1, 5$ and $\tau = 0.1$.

(b) G and SG predictors with Parameters $\alpha = -0.01, \beta = -5, -1, 0, 5$ and $\tau = 0.1$.

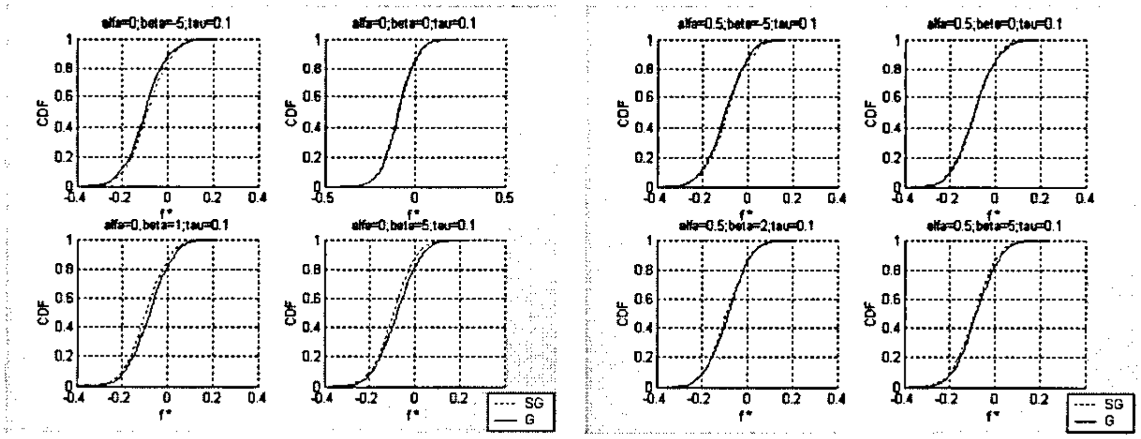


Figure (4.2):

(a) G and SG predictors with Parameters

$\alpha = 0,\ \beta = -5, 0, 1, 5$ and $\tau = 0.1.$

(b) G and SG predictors with Parameters

$\alpha = 0.05, \beta = -5, 0, 2, 5$ and $\tau = 0.1.$

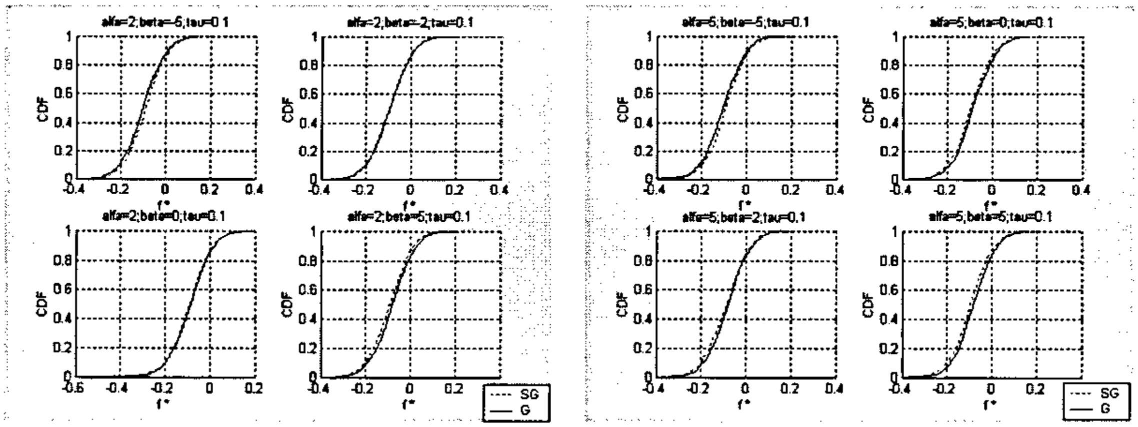


Figure (4.3):

(a) G and SG predictors with Parameters

$\alpha = 2, \beta = -5, -2, 0, 5$ and $\tau = 0.1.$

(b) G and SG predictors with Parameters

$\alpha = 5, \beta = -5, 0, 2, 5$ and $\tau = 0.1.$

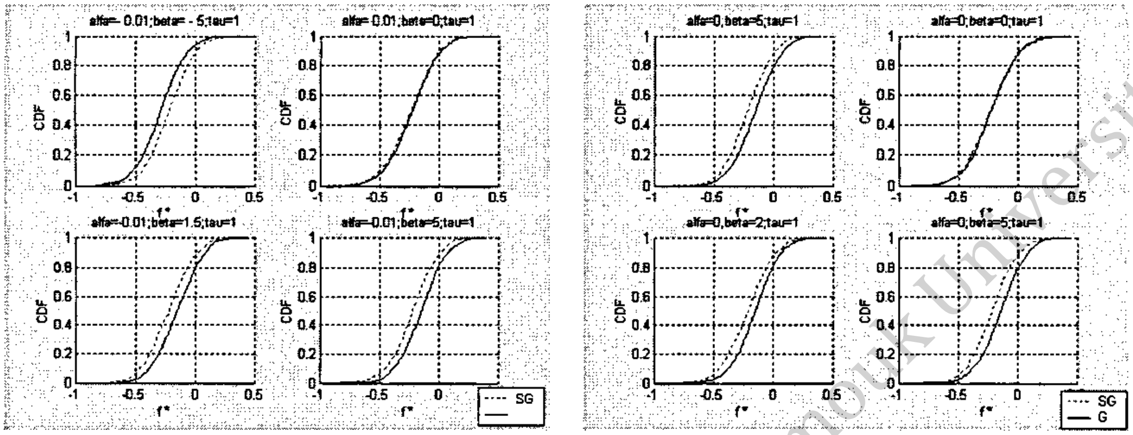


Figure (4.4):

(a) G and SG predictors with Parameters $\alpha = -0.01, \beta = -5, 0, 1.5, 5$ and $\tau = 1$.

(b) G and SG predictors with Parameters $\alpha = 0, \beta = 5, 0, 2, 5$ and $\tau = 1$.

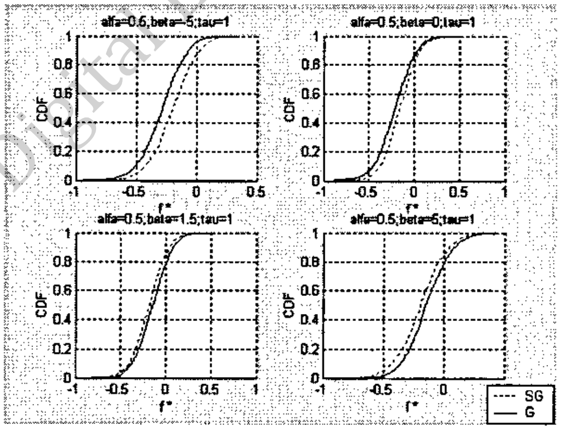


Figure (4.5) G and SG predictors with Parameters $\alpha = 0.5, \beta = -5, 0, 1.5, 5$ and $\tau = 1$.

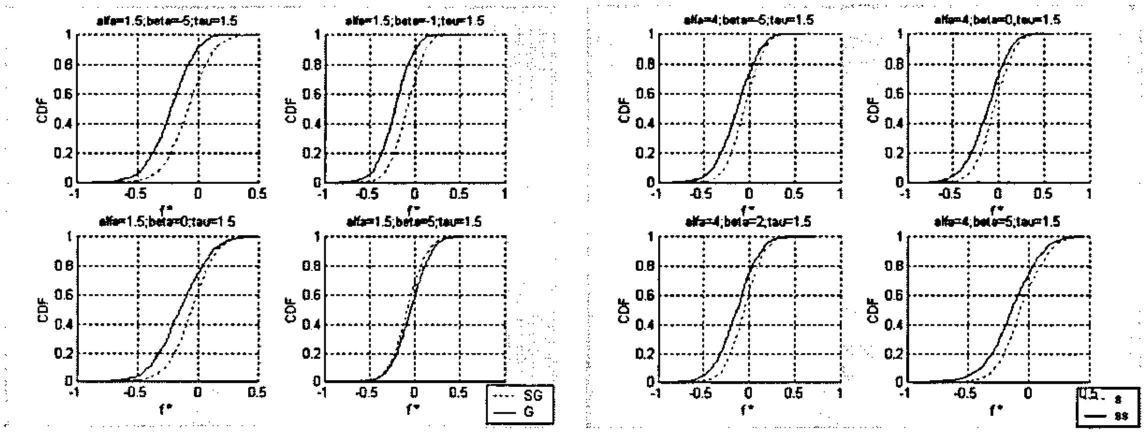


Figure (4.6):

(a) G and SG predictors with Parameters $\alpha = 1.5, \beta = -5, -1, 0, 5$ and $\tau = 1.5$.

(b) G and SG predictors with Parameters $\alpha = 4, \beta = -5, 0, 2, 5$ and $\tau = 1.5$.

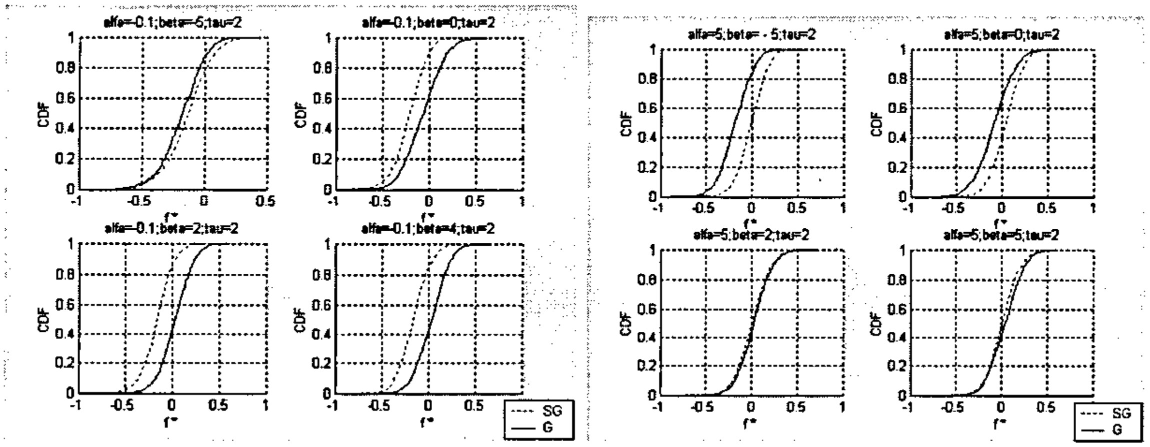


Figure (4.7):

(a) G and SG predictors with Parameters $\alpha = -0.1, \beta = -5, 0, 2, 4$ and $\tau = 2$.

(b) G and SG predictors with Parameters $\alpha = 5, \beta = -5, 0, 2, 5$ and $\tau = 2$.

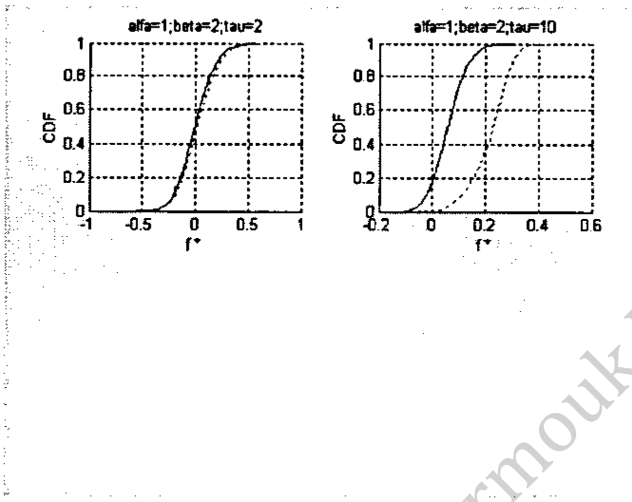


Figure (4.8) : G and SG predictors with Parameters $\alpha = 1, \beta = 2$ and $\tau = 2, 10$.

From figures (4.1) – (4.8), we report the following concluding remarks:

1. If a Gaussian process prior is used on the input function, i.e., $\alpha = 0$, then there is a small difference between the two distributions and this difference is increasing as a function of $|\beta|$. Moreover, the skew Gaussian predictor distribution is larger than the Gaussian predictor distribution if $\beta < 0$ and the converse is true if $\beta > 0$. (See Figure 4.2 (a)).

2. The two predictors have about the same distribution functions for small values of the skewness parameters τ, α and β . (See Figures 4.2 (a), (b)).

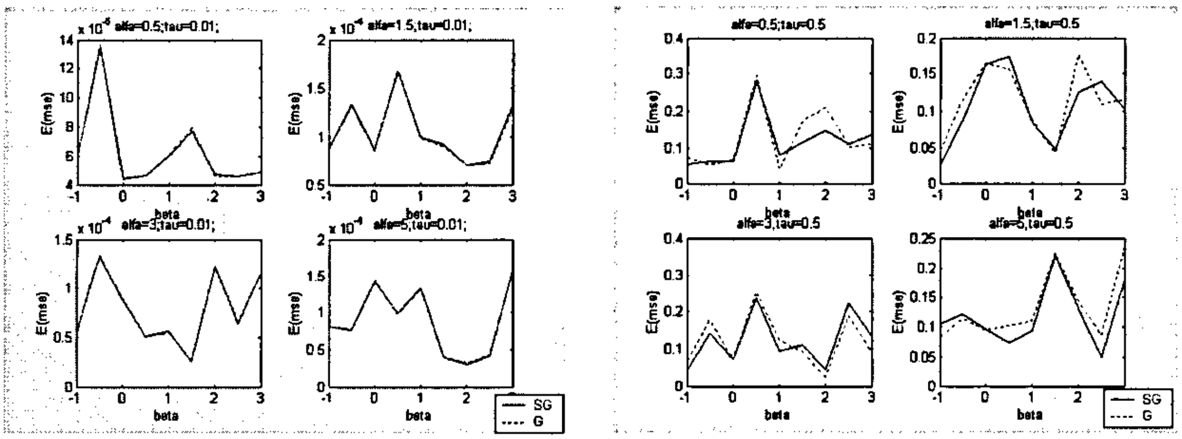
3. If a Gaussian process is used on the error, i.e., $\beta = 0$, then there is no difference between the two distributions when $\alpha \leq 0$, and τ is small. (See Figures 4.1, 4.2, 4.3 and 4.4).

4. For fixed values of α, and moderate values of τ, the difference between the two distribution is very clear and seems to be an increasing function in $|\beta|$.(See Figures 4.4, 4.5)

5. For fixed values of α, and large values of τ, there is a huge difference between the two distributions. (See Figure 4.8).

6. In general, the skew Gaussian predictor distribution is larger than the Gaussian predictor distribution. Also the Gaussian predictor is not robust against departure from Gaussianity.

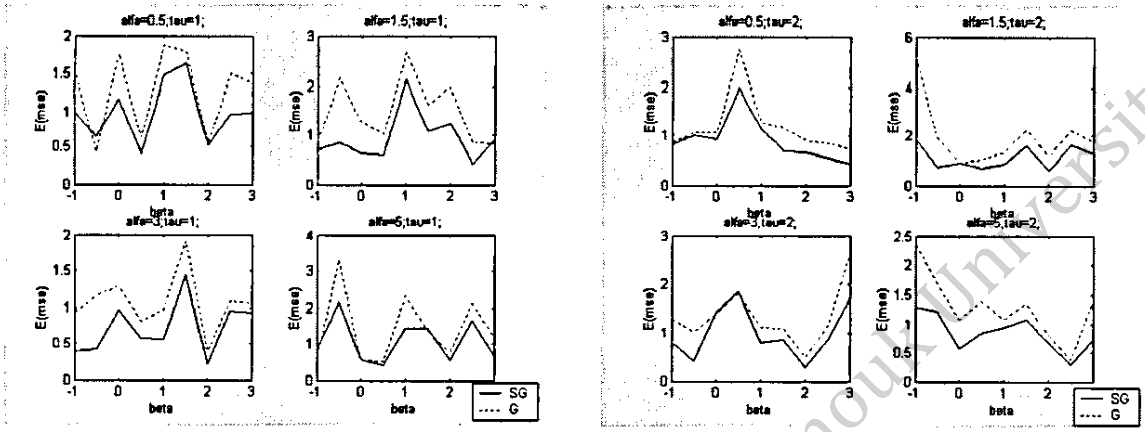
In figures (4.9) - (4.14) we plot the prediction errors of the Gaussian and the skew Gaussian predictors versus α, β and τ as follows:



(a) $\tau = 0.01$.versus $\beta = -1, \ldots, 3$.

(b) $\tau = 0.5$.versus $\beta = -1, \ldots, 3$.

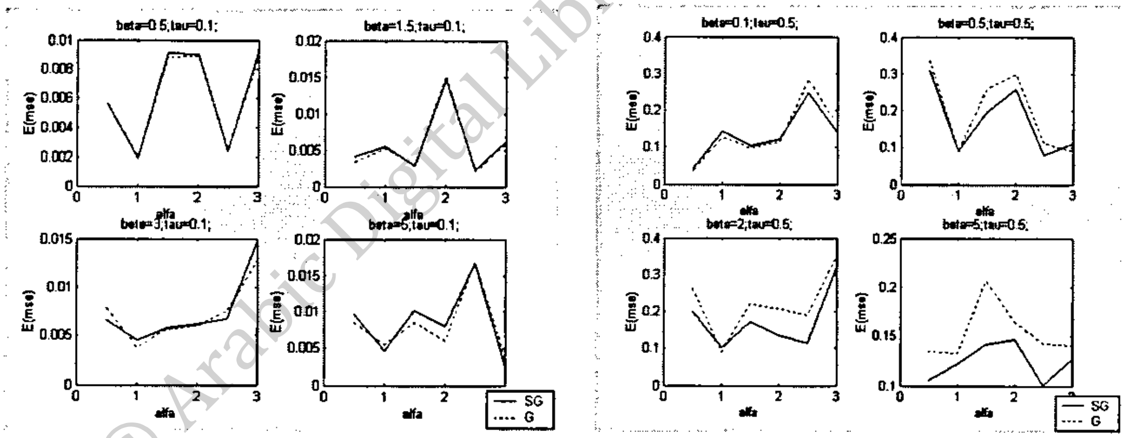
Figure (4.9): $E(mse)$ in G and SG predictors with $\alpha = 0.5, 1.5, 3, 5$.



(a) $\tau = 1$ and $\alpha = 0.5, 1.5, 3, 5.$

(b) $\tau = 2$ and $\alpha = 0.5, 1.5, 3, 5.$

Figure (4.10): $E(mse)$ in G and SG predictors versus $\beta = -1, \dots, 3.$



(a) $\tau = 0.1$ and $\beta = 0.5, 1.5, 3, 5.$

(b) $\tau = 0.5$ and $\beta = 0.1, 0.5, 2, 5.$

Figure (4.11): $E(mse)$ in G and SG predictors versus $\alpha = 0, \dots, 3.$

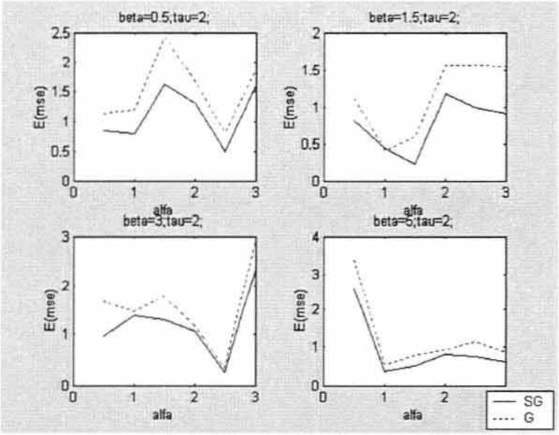
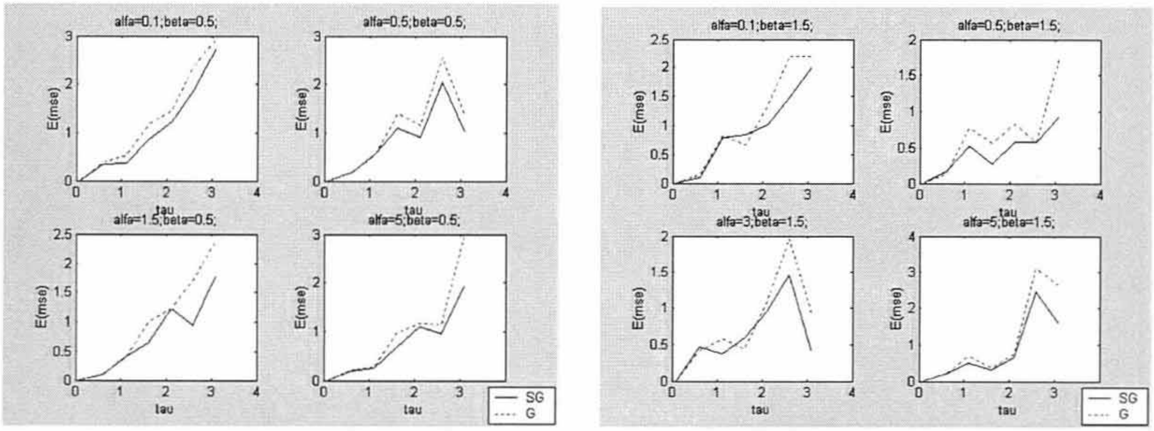


Figure (4.12) $E(mse)$ in G and SG predictors with $\tau = 2$ and $\beta = 0.5, 1.5, 3, 5$ versus $\alpha = 0, \ldots, 3$.



(a) $\alpha = 0.1, 0.5, 1.5, 5$ and $\beta = 0.5$.

(b) $\alpha = 0.1, 0.5, 3, 5$ and $\beta = 1.5$.

Figure (4.13): $E(mse)$ in G and SG predictors versus $\tau = 0, \ldots, 4$.

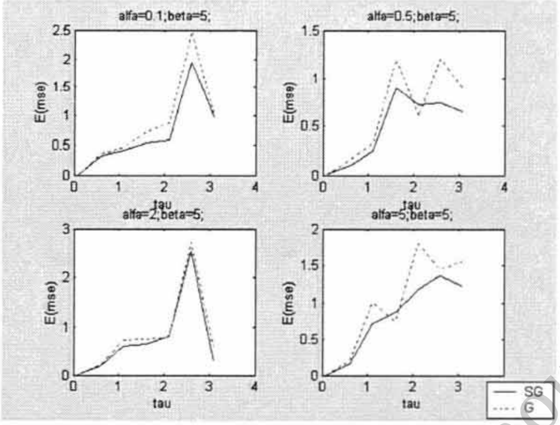


Figure (4.14) : $E(mse)$ in G and SG predictors with $\alpha = 0.1, 0.5, 2, 5$ and $\beta = 5$ versus $\tau = 0, \ldots, 4$.

From these figures, we write the following concluding remarks:

1. For small values of τ, the two predictions have about the same prediction errors. (See Figures 4.9, 4.11 (a)).

2. The prediction error of the skew Gaussian predictors is less than the prediction error of the Gaussian predictors when α is increasing and τ take a large value. (See Figure 4.10)

3. The difference between two predictors is increased when τ is increased. (See Figures 4.13, 4.14)

4. Note figure (6) it is clearly that the $E(mse)$ in skew Gaussian predictor is less than the $E(mse)$ in Gaussian predictor, and the difference between two values is increasing when τ increasing.

In general the value of $E(mse)$ in skew Gaussian predictor is approach from its value in Gaussian predictor when τ is small which appear clearly in figures (6), (7-a) and (8-a). Another fact that appears in all figures that the value of $E(mse)$ is small in skew Gaussian predictor and Gaussian predictor when the parameters are small, and its increasing when the parameters are increasing but its stile the value of the $E(mse)$ is smaller in skew Gaussian predictor than its value in Gaussian predictor.

4.4 Estimation of Hyper Parameters.

The Maximum Likelihood Estimation (MLE) is an estimation to estimate the parameters. Here we use the MLE to estimate the parameters $\tau, \sigma, \alpha, \beta, a$ and λ i.e. by maximizing the function $L(\tau, \sigma^2, \alpha, \beta)$ where L denote the pdf of .

Consider the model

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\epsilon}\,,$$

where $\boldsymbol{X} \sim SN_n(\mathbf{0}, \boldsymbol{\Sigma}, \alpha\mathbf{1}_n^{\mathrm{T}}, 0, 1)$, and $\boldsymbol{\epsilon} \sim SN_n(\mathbf{0}, \tau^2\boldsymbol{I}_n, \beta\mathbf{1}_n^{\mathrm{T}}, 0, 1)$, $\tau > 0$, and $\boldsymbol{X}, \boldsymbol{\epsilon}$ are independent random vectors then we can apply the special case of Proposition 2.5. When $q = 2$, we get

$$\boldsymbol{Y} \sim CSN_{n,2}(\mathbf{0}, \boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n, \boldsymbol{D}^{\circ}, \mathbf{0}, \Delta^{\circ}),$$

where

$$\boldsymbol{D}^{\circ} = \begin{pmatrix} \alpha\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}(\,\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1} \\ \beta\tau^2\mathbf{1}_n^{\mathrm{T}}(\,\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1} \end{pmatrix}, \quad \Delta^{\circ} = \begin{pmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{pmatrix},$$

and

$$A_{11} = 1 + \alpha^2\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{1}_n - \alpha^2\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}(\,\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\boldsymbol{\Sigma}\mathbf{1}_n,$$

$$A_{22} = 1 + n\beta^2\tau^2 - \beta^2\tau^4\mathbf{1}_n^{\mathrm{T}}(\,\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\mathbf{1}_n,$$

$$A_{12} = -\alpha\beta\tau^2\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n)^{-1}\mathbf{1}_n.$$

Now when we apply the definition 2.1 we get the likelihood function of the hyperparameters

$$g_{n,2}(\boldsymbol{Y}) = \frac{\Phi_2\left(\boldsymbol{D}^{\circ}\boldsymbol{Y}; 0, \Delta^{\circ}\right)}{\Phi_2\left(0; 0,\ \Delta^{\circ} + \boldsymbol{D}^{\circ}(\Sigma + \tau^2 \boldsymbol{I}_n)\boldsymbol{D}^{\circ T}\right)}\, \phi_n(\boldsymbol{Y}; \mathbf{0}, \boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n)\ .$$

CHAPTER FIVE

CONCLSIONS AND POSSIBILITYS FOR FUTURE STUDY

In this thesis, the non-linear regression model $\boldsymbol{Y(t) = f(t) + \epsilon(t)}$ has been tacked from a Bayesian viewpoint by assuming two skew Gaussian processes on $\boldsymbol{f(t)}$ and $\boldsymbol{\epsilon(t)}$. It is shown that, under this assumption, the predictive density at new input has a closed form. Also, we studied the GPR predictor under the assumption that the error violates the assumption of Gaussianity. If the error departs from Gaussianity to skew-Gaussianity, then the GPR predictor will be affected and may lead to unrealistic estimates. We know that skew Gaussian process for regression addressed in this thesis has several advantages over the GPR (see section 4.3). These advantages will attrac us to continue this work in future. We highlight some of that possible works:

1. Studying the effect of the choice of the covariance function on the skew Gaussian process predictor.
2. Developing methods for estimating the hyper-parameters of the model.
3. Prediction at several function inputs.

REFERENCES

[1] Allard, D. and Naveau, P. (2007). A New spatial skew-normal random field model. Comm. *In statistics-theory and methods*, vol(36), 1821-1834.

[2] Alodat, M. T. and AL-Rawwash, M. Y. (2009). Skew Gaussian random field. *Journal of Computational and applied mathematics,* **232, 2**, p 496-504.

[3] Alodat, M. T., Al-Rawwash, M. Y. and Al Jebrini, M. A. (2010). Duration distribution of the conjunction of two independent F process. *J. Appl. Probab., vol. (47), 179-190.*

[4] Alodat, M. T. and Aludaat, K. M. (2007). A skew Gaussian process. *Applied mathematical Science*, vol. (23(1)), 89-97.

[5] Anagreh,Y., Bataineh, A. and Al-Odat, M. (2010). Assessment of renewable energy potential, at Aqaba in Jordan. *Renewable and Sustainable Energy Reviews* 14(2010). 1347-1351.

[6] Arnold, B. C. and Beaver, R. J. (2002). Skew multivariate models related hidden truncation and/ or selective reporting t. test, 11(1), 7-54

[7] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171-178.

[8] Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* **46**, 199-208.

[9] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika 83*, 715-726.

[10]Azzlini, A. and Capitanio, A. (1999). Statistical application of the multivariate skew normal distributions. *J. R. Stat. Soc., ser. B* **61**, 579-602.

[11] Brahim-Belhouari, S. and Bermak, A. (2004). Gaussian process for non-stationary time series prediction. *Computational statistics and data analysis*, **47**, 705-712.

[12] Christian, P. R. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York.

[13]Buccianti, A. (2005), 'Meaning of the λ parameter of skew–normal and log–skew normal distributions in fluid geochemistry' a CODAWORK'05.

[14] Fyfe, C., Leen, G. and Lai, P.L. (2008).*Gaussian Processes for Canonical Correlation analysis*. Neuro computing, 71, 3077-3088.

[15] Genton, M. G., Wang, J., Boyer, J. and Marc, (2004). A Skew-Symmetric Representation of Multivariate Distributions. North Carolina State University.

[16] Girard, A. (2004). *Approximate methods for propagation of uncertainty with Gaussian process models*. Ph.D. thesis, department of computer science, University of Glasgow.

[17] Girard, A. Kocijan, J. Murray-Smith, R. and Rasmussen, C. E. (2004). Gaussian Process Model Based Predictive Control, proceeding of the American Control Conference Boston.

[18] Girard, A. Rasmussen, C. E. and Murray-Smith, R. M. (2002). *Gaussian Process priors with uncertain Inputs*: Multiple-Step-Ahead Prediction. Technical Report TR-2002-119, Department of computing Science, University of Glasgow.

[19] Gonzales-Farias, G., J. Domingusez-Molina and A.Gupta, 2004. Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference*, **126**, 521-534.

[20] Kuss, M. (2006). *Gaussian process models for Robust Regression, Classification, and Reinforcement learning*. Ph.D. thesis, Technische Universität Darmstadt.

[21] Macke, J. H., Gerwinn, S., White, L. E., Kaschube, M. and Bethge, M. (2010). Gaussian process methods for estimating Cortical maps.

[22] Neal, R. M. (1995). *Bayesian learning for Neural Networks.*, Ph.D., thesis, Dept. of Computer Science, University of Toronto.

[23] O'Hagan, A. (1978). On curve fitting and optimal design for prediction. *J. R.* soc., *B*40, 1- 42.

[24] Quinonero-Candela, J., Girard, A. and Rasmussen, C. E. (2003). Prediction at an Uncertain Input for Gaussian Processes and Relevance Vector Machines Application to Multiple-Step Ahead Time-Series Forecasting. Technical report, Technical University of Denmark.

[25] Rasmussen, C. E. and Williams, C. (2006). *Gaussian processes for machine learning*. Cambridge, Massachusetts, MIT press.

[26] Rasmussen, C. E. (1996). *Evaluation of Gaussian Processes and other methods for non-linear regression,* Ph.D. thesis, Dept. of Computer Science, University of Toronto.

[27] Schmidt, A. M., Concoicão, M. F. and Moreira, G. A. (2008). Investigating the sensitivity of Gaussian processes to the choice of their correlation functions and prior specifications. *Journal of statistical computation and simulation*, 78, 8, 681-699.

[28] Schott, J. R. (1997). *Matrix Analysis for statistics.* Wiley-Interscience.

[29] Vanhatalo, J., Jylanki, ·P. and Vehtari, A. (2009). Gaussian process regression with student. t likelihood. Advances in Neural Information processing systems **23**.

[30] Williams, C. K. I. and Rasmussen, C. E. (1996) Gaussian Processes for Regression. *Advances in Neural Information Processing Systems 8*, 514-520.

[31] Zhang, H., El-Shaarawi, A. (2009). On spatial skew Gaussian process applications. *Environmentics*, **10**, 982.

Appendix

A. MatLab Program for Simulation Study

This program is to finding the predictor values using the distribution in (3.1)

```
sigma=[];epsi=[];lam=7.6;alfa=1.5;beta=1.5;tau=1.5;a=.5;
miter=1000;
%-----Creating the covaraince matrices ---------
t=[-5:-1,1:.1:5];ts=6;
sigma=covm(t,lam,a);
tt=[t,ts];
epsi=covm(tt,lam,a);
ks=a^2;
 k=epsi(1:length(t),length(tt));
 sigmaA1=[sigma+tau^2*eye(length(t));k'];
 sigmaA2=[k;ks];
 sigmaA=[sigmaA1,sigmaA2];
 u11=alfa*ones(1,length(tt))*[sigma,k]';
 u12=alfa*ones(1,length(tt))*[k',ks]';
 u21=beta*tau^2*ones(1,length(t));
 u22=0;u1=[u11;u21];u2=[u12;u22];
 u=[u1,u2];
dA=u*inv(sigmaA);
delt=[1+alfa^2*ones(1,length(tt))*epsi*ones(1,length(tt))' 0;
     0 1+length(t)*beta^2*tau^2];
 deltaA=delt-u*inv(sigmaA)*u';
```

```
ft=sin(t)./t;
mu=zeros(1,length(t));
vbeta=beta*ones(1,length(t));

ss=[];s=[];smed=[];
for j=1:1000
 u=rand(1);
 z=tau*randn(1,length(t));
 while u>normcdf(beta*ones(1,length(t))*z',0,1);
  u=rand(1);
  z=tau*randn(1,length(t));
 end
  y=ft+z;
  mus=k'*inv(sigma+tau^2*eye(length(t)))*y';
  sigs=ks-k'*inv(sigma+tau^2*eye(length(t)))*k;
  D1=dA(:,1:length(t));D2=dA(:,length(tt));
  Ds=D1+D2*k'*inv(sigma+tau^2*eye(length(t)));
  DD=D2;
  vs=-Ds*y';
  fstar=[];
   for i=1:miter
     zz=mvnrnd(vs,deltaA,1);
     yy=mus+sqrt(sigs)*randn(1);
     w=zz'-DD*(yy-mus);
     while max(w)>0
       zz=mvnrnd(vs,deltaA,1);
       yy=mus+sqrt(sigs)*randn(1);
```

```
         w=zz'-DD*(yy-mus);
        end;
    fstar=[fstar;yy];
end;
s=[s;mean(fstar)];
smed=[smed;median(fstar)];

%%%%%%%Gussian Case
sigs=ks-k'*inv(sigma+tau^2*eye(length(t)))*k;
mus=k'*inv(sigma+tau^2*eye(length(t)))*y';
ss=[ss;mus];
end
[mean(s)  var(s)+(mean(s)-sin(ts)/ts)^2]
[mean(ss) var(ss)+(mean(ss)-sin(ts)/ts)^2]
[mean(smed) var(smed)+(mean(smed)-sin(ts)/ts)^2]
sin(ts)/ts

hold on
cdfplot(ss)
cdfplot(s)
cdfplot(smed)
hold off


 errorbar(Y, E)
 sqrt(2/pi)*tau^2*beta*sqrt(1+length(t)*beta^2*tau^2)^-
1*k'*inv(sigma+tau^2*eye(length(t)))*ones(1,length(t))'
```

```
ttt=[t,ts];
fff=[ft,sin(ts)./ts];
yyy=[y,mean(fstar)];
hold on
plot(ttt,fff)
plot(ttt,yyy,'+')
hold off
```

```
t=-5:.001:5;
y=sin(t);
l=.5*ones(1,length(t));
u=.5*ones(1,length(t));
hold on
errorbar(t,y-l,t,y+u,'+')
plot([t(1), t(100), t(200), t(300)],[y(1), y(100), y(200), y(300)],'+')
plot(t,y)
```

```
errorbar(t,y,ll,uu)
 for i=1:length(t)
  for j=1:length(t)
     sigma(i,j)=a^2*exp(-abs(t(i)-t(j))^2/2/lam);
     end;
  end;
 tt=[t,ts];
 for i=1:length(tt)
  for j=1:length(tt)
  epsi(i,j)=a^2*exp(-abs(tt(i)-tt(j))^2/2/lam);
  end
 end
```

B. MatLab Program for Computing the Bias

This program is to computing the bias in SG-case and G-case and compare between them.

```
sigma=[];epsi=[];lam=1.6;alfa=3.5;

beta=1.5;tau=1.;a=5.;

miter=1000;

lwr=-1;upr=3;inc=.5;

sk=[];g=[];

for beta=lwr:inc:upr

%-----Creating the covaraince matrices ---------

t=[-2:.5:-.05,.05:.5:2];

ys=[];yg=[];

for ii=1:length(t)

   ts=t(ii);

sigma=covm(t,lam,a);

tt=[t,ts];

epsi=covm(tt,lam,a);
```

```
ks=a^2;
 k=epsi(1:length(t),length(tt));
 sigmaA1=[sigma+tau^2*eye(length(t));k'];
 sigmaA2=[k;ks];
 sigmaA=[sigmaA1,sigmaA2];
 u11=alfa*ones(1,length(tt))*[sigma,k]';
 u12=alfa*ones(1,length(tt))*[k',ks]';
 u21=beta*tau^2*ones(1,length(t));
 u22=0;u1=[u11;u21];u2=[u12;u22];
 u=[u1,u2];
dA=u*inv(sigmaA);
delt=[1+alfa^2*ones(1,length(tt))*epsi*ones(1,length(tt))' 0;
     0 1+length(t)*beta^2*tau^2];
 deltaA=delt-u*inv(sigmaA)*u';
ft=sin(t)./t;
mu=zeros(1,length(t));
vbeta=beta*ones(1,length(t));
 u=rand(1);
 z=tau*randn(1,length(t));
```

```
while u>normcdf(beta*ones(1,length(t))*z',0,1);
  u=rand(1);
  z=tau*randn(1,length(t));
end
 y=ft+z;
 mus=k'*inv(sigma+tau^2*eye(length(t)))*y';
 sigs=ks-k'*inv(sigma+tau^2*eye(length(t)))*k;
 D1=dA(:,1:length(t));D2=dA(:,length(tt));
 Ds=D1+D2*k'*inv(sigma+tau^2*eye(length(t)));
 DD=D2;
 vs=-Ds*y';
 fstar=[];
  for i=1:miter
    zz=mvnrnd(vs,deltaA,1);
    yy=mus+sqrt(sigs)*randn(1);
    w=zz'-DD*(yy-mus);
    while max(w)>0
      zz=mvnrnd(vs,deltaA,1);
      yy=mus+sqrt(sigs)*randn(1);
```

```
         w=zz'-DD*(yy-mus);
       end;
    fstar=[fstar;yy];
   end;
 ys=[ys,mean(fstar)];
   sigs=ks-k'*inv(sigma+tau^2*eye(length(t)))*k;
   mus=k'*inv(sigma+tau^2*eye(length(t)))*y';
   yg=[yg;mus];
end;
sk=[sk;sum((ys-ft).^2)/(length(ys))];
g=[g;sum((yg'-ft).^2)/(length(yg))];
end
plot(lwr:inc:upr,sk,'-',lwr:inc:upr,g,'-')
```