Application of Gaussian Process and Maximum Entropy Sampling in Methane Plume Prediction

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Mathematics by

Raquel Anastasia Barata

May 2014
The thesis of Raquel Anastasia Barata is approved:

________________________________________  

Dr. Ramin Vakilian  Date

________________________________________  

Dr. David Thompson  Date

________________________________________  

Dr. Mark Schilling, Chair  Date

California State University, Northridge
Dedications

Dedicated to Todd, Tasha and Texas.
Acknowledgments

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and made possible by the JPL Visiting Student Researcher Program (JVSRP) and the National Aeronautics and Space Administration (NASA).

I would like to thank Dr. David Thompson and Dr. Mark Schilling for their continued patience and guidance throughout this process. Specifically, David Thompson for the opportunity to work on this project in coloration with the innovative scientists at JPL.

I would also like to thank the Louis Stokes Alliance for Minority Participation Bridge to the Doctorate program that made all this possible.
# Table of Contents

Signature Page ii
Dedications iii
Acknowledgements iv
Abstract vi

1 Introduction 1
   1.1 Statistical Review 2

2 The Gaussian Process 7
   2.1 The Mean and Covariance Function 7
   2.2 Bayesian Inference via Inclusion of Sampled Data 9
   2.3 Optimization of Hyperparameters 11

3 Refinements to the Gaussian Process 13
   3.1 Inclusion of Wind direction in the Covariance Function 13
   3.2 The log-Gaussian Process 13

4 Maximum Entropy Sampling 15

5 Simulated Results 19
   5.1 Maximum Entropy Sampling Simulations 23
   5.2 Field Data Results 25

6 Conclusion 27

Bibliography 28

Appendix A: Gaussian Properties 29
Appendix B: Numerical Analysis 33
Appendix C: Additional Proofs 36
Abstract

Application of Gaussian Process and Maximum Entropy Sampling in Methane Plume Prediction

by

Raquel Anastasia Barata
Master of Science in Mathematics

We implement a Gaussian Process model in the prediction and visualization of a methane field. Starting with an initial training data set, we fit the Gaussian Process to the data by adjusting the hyperparameters of the covariance function using maximum likelihood estimation. Further refinements include the consideration of wind direction in the covariance function and implementation of the log-Gaussian Process to more accurately model extreme fluctuations in methane measurements. We couple this modeling approach with a sampling technique known as Maximum-Entropy Sampling. This method selects measurements that provide the most information about the Gaussian Process. Ultimately, the research will be applied in autonomous exploration where the user cannot specify their measurement objectives but the exploratory robot can perform adaptive sampling that accounts for previous measurement distributions or prior expectations.
Chapter 1

Introduction

Methane and other gas leaks are a prevalent problem in the modern world. Current techniques for sampling and visualizing these leaks include everything from hand held manual methane meters to thermodynamic modeling. This thesis explores methods of sampling, predicting, and visualizing a methane field. This is accomplished by utilizing statistical, machine learning and information theory techniques.

Gaussian process models are well known in the geostatistical field as kriging. These models are used for spatial prediction of terrain in both two and three dimensional input spaces. They are also used in one dimensional regression applied to the analysis and design of computer experiments.

Specifically in the machine learning community, Gaussian Processes models are widely used for Bayesian regression. They are parameteric models which define a distribution over a function space. They provide a general framework that can be applied not only to regression but to interpolation and prediction where a deterministic model is not obtainable. Gaussian Processes have been used in everything from robotic adaptive path planning (Low, K. H., et al. 2009) [1] to modeling instrumental systematics in application to transmission spectroscopy (Gibson, N. P., et al. 2012) [2]. Although our Gaussian Process regression utilizes many of the same techniques as the above applications, this specific application to plume prediction has not been seen before.

We suggest the use of Gaussian Process models in methane plume prediction. Specification of the mean and covariance function with parameters allows us to tailor the process to our environmental assumptions and observations. The “learning” of the Gaussian Process is included in the posterior after initial observations are taken and the hyperparameters are optimized using maximum likelihood estimation. Refinements specific to methane plume prediction such as more accurately modeling high concentrations using a log-Gaussian Process and accounting for wind direction using Automatic Relevance Detection (ARD) are easily included in our model.

Finally, we couple this prediction technique with an information theoretic sampling method known as Maximum Entropy Sampling (MES). The information gained in an experiment is optimized by sampling in spatial locations specifically chosen by this MES method. Coupled with the ordinary Gaussian Process, the entropy of each interpolated spatial location ends up being the predicted variance. However, coupled with the log-Gaussian Process, a scalar of the predicted mean is added to the entropy resulting in higher sampling in areas of higher predicted concentrations. These additional measurements feed back into our prediction model, further refining our process.
Preliminary testing was performed on a simulated methane plume data depicted as grayscale images. We simulated many combinations of MES sampling techniques paired with different Gaussian Processes and with various refinements. Real data collection was also done in the field using an orthocopter. The field experiment was done in Wyoming in late June of 2013. The orthocopter was able to sample orthogonally to a methane plume at three different altitudes. We assume the data to be from a Gaussian Prior process where the inputs are the three dimensional spatial coordinates and the outputs are the observed methane values. We were able to interpolate and visualize a three dimensional methane plume for concentrations greater than or equal to 2ppm. This paper concludes with the results of our sampling and prediction models on both simulations and real field data.

We finish this chapter with a quick statistical review. Then in chapter 2 we explain the statistical details of the Gaussian Process including the choice of mean and covariance functions, the inclusion of the training data through Bayesian inference, and the adjustment of covariance hyperparameters using maximum likelihood estimation. In chapter 3, consideration of wind direction in the covariance function and implementation of the log-Gaussian Process are detailed as further refinements to the Gaussian Process. In chapter 4, we introduce a sampling method known as Maximum Entropy Sampling, which selects measurements that provide the most amount of data. Chapter 5 compares sampling and inference techniques on simulated methane plumes, where Chapter 6 shows plume inference from field data collected by orthocopter. Lastly, Chapter 7 concludes with a discussion of results and future work.

1.1 Statistical Review

Definition 1 A probability space is a measure space usually denoted \((\Omega, \mathcal{F}, \mathbb{P})\) where:

- \(\Omega\) is a set known as the sample space, elements of which are denoted \(\omega\).
- \(\mathcal{F}\) is a \(\sigma\)-algebra with elements \(E_1, E_2, E_3, \ldots\) called events.
- \(\mathbb{P}\) is a function such that \(\mathbb{P} : \mathcal{F} \rightarrow [0, 1]\) with \(\mathbb{P}(\Omega) = 1\) and

\[
\mathbb{P} \left( \bigcup_{j=1}^{\infty} E_j \right) = \sum_{j=1}^{\infty} \mathbb{P}(E_j)
\]

where \(E_1, E_2, E_3, \ldots\) are disjoint sets of \(\mathcal{F}\). [3]

Definition 2 A random variable \(X\) is a measurable function from a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) to \(\mathbb{R}\)

\[
X : \Omega \rightarrow (-\infty, \infty)
\]
such that for every Borel set $B$, if $\{X \in B\}$ then there exists $\{\omega \in \Omega : X(\omega) \in B\}$. [3]

Intuitively, a random variable can be thought of as a variable that takes on its values by chance, typically denoted by a capital letter such as $X, Y, Z$. [3] For this project, we restrict our class of random variables to those with finite expectation and defined density.

We adopt the notation $\{Y \leq y\}$ to express the event when a random variable $Y$ takes a value less than or equal to a real number $y$. Similarly, we denote the probability of this event’s occurrence $P\{Y \leq y\}$.

**Definition 3** The cumulative distribution function of the random variable $Y$, denoted $F(\cdot)$ is defined as the function with domain $\mathbb{R}$ where $F(y) = P\{Y \leq y\}$ for $\forall y \in \mathbb{R}$. [3]

**Definition 4** A random variable is said to be a Continuous Random Variable if there exists a function $f(\cdot)$ such that $F(y) = \int_{-\infty}^{y} f(u) du$ for $\forall y \in \mathbb{R}$. [3]

**Definition 5** If $Y$ is a continuous random variable, the function $f(\cdot)$ in $F(y) = \int_{-\infty}^{y} f(u) du$ is called the probability density function. [3] Furthermore, $f(\cdot)$ is a probability density function if and only if:

- $f(y) \geq 0$ for all $y$.
- $\int_{-\infty}^{\infty} f(y) dy$

**Definition 6** The expected value or mean of a continuous random variable $Y$ is denoted $\mu = E[Y]$ is $E[Y] = \int_{-\infty}^{+\infty} y f(y) dy$ where $f(\cdot)$ is the probability density function of the random variable. [3]

**Definition 7** The variance of a continuous random variable $Y$ denoted $\sigma^2 = \text{var}[Y]$ is defined to be $E[(Y - \mu)^2] = \int_{-\infty}^{+\infty} (y - \mu)^2 f(\cdot) dy$. [3]

**Definition 8** The $n$th moment of a random variable $Y$ is defined as $E[Y^n]$. [4]

For this project, we assume all random variables have finite first and second moments.

**Definition 9** A parameter is a value (usually unknown) used to represent a certain numerical characteristic of a probabilistic model. For example, the mean $\mu$ and variance $\sigma^2$ defined above are parameters of the random variable $Y$ and its associated probability distribution.

**Definition 10** (The Univariate Gaussian Distribution) The Gaussian distribution with parameters $\mu$ and $\sigma^2 > 0$ is given by the probability density function

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

for $-\infty < y < \infty$. [3]
Definition 11 Let \( Y_1, \ldots, Y_n \) be a set of continuous random variables. The joint cumulative distribution function of \( Y_1, \ldots, Y_n \) is denoted \( F(\cdot, \ldots, \cdot) \) and defined as \( P[Y_1 \leq y_1, \ldots, Y_n \leq y_n] \) for all \( \{y_1, \ldots, y_n\} \in \mathbb{R} \). [4]

Definition 12 Let \([Y_1, \ldots, Y_n]'\) be a vector of random variables (a random vector). \([Y_1, \ldots, Y_n]'\) is a continuous random vector if and only if there exists a \( f(\cdot, \ldots, \cdot) \geq 0 \) such that \( F(\cdot, \ldots, \cdot) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f(u_1, \ldots, u_n) du_1 \cdots du_n \) for all \( \{y_1, \ldots, y_n\} \in \mathbb{R} \). \( f(\cdot, \ldots, \cdot) \) is defined to be the joint probability density function. [4]

Definition 13 Let \([Y_1, \ldots, Y_n]'\) be a random vector with joint probability density function \( f(\cdot, \ldots, \cdot) \). [4] The marginal probability density function of \( Y_i \) is

\[
f_{Y_i}(y_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n) du_1 \cdots du_{i-1} du_{i+1} \cdots du_n
\]

Definition 14 Let \([Y_1, \ldots, Y_n]'\) be a random vector with joint probability density function \( f(\cdot, \ldots, \cdot) \). [4] The conditional probability density function of \( Y_i \) given observations \( y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n \) is defined as

\[
f(y_i | y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) = \frac{f(y_1, \ldots, y_n)}{f_{Y_i}(y_i)}
\]

Definition 15 A random vector \( Y = [Y_1, \ldots, Y_n]' \) is said to have a multivariate Gaussian distribution or joint Gaussian distribution if every linear combination \( \alpha_1 Y_1 + \cdots + \alpha_n Y_n \) where \( \alpha_i \in \mathbb{R} \) has a univariate Gaussian distribution. [5]

A multivariate Gaussian distribution is specified by two parameter structures, the mean vector \( \mu = E[Y] \) and the covariance matrix \( \Sigma = E[(Y - \mu)(Y - \mu)^T] \) where each \( i, j \)th component is \( \text{cov}(Y_i, Y_j) = E[(Y_i - \mu_i)(Y_j - \mu_j)] \).

Definition 16 (The Multivariate Gaussian Distribution) The \( n \)-dimensional Gaussian distribution for a random vector \( Y = [Y_1, \ldots, Y_n]' \) with mean vector \( \mu \) and covariance matrix \( \Sigma \) has probability density function

\[
f(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2}
\]

for \( -\infty \leq y_i \leq \infty \) for \( i = 1, 2, \ldots, n \). [5]

Definition 17 A Stochastic Process is a collection of random variables \( \{f(t) : t \in T\} \) where \( T \) is the index set of the process. [3]

For this project, our index set is a spatial coordinate in \( \mathbb{R}^n \). Formally, we will have a collection of random variables \( \{f(x) : x \in \mathbb{R}^n\} \). For each \( x \in \mathbb{R}^n \), \( f(x) \) is a random variable.

Definition 18 The Mean Function, \( m(x) \), of a stochastic process, \( f(x) \), is defined as \( m(x) = E[f(x)] \). [3]
Definition 19 The Covariance Function, \( k(x, x') \), for a stochastic process, \( f(x) \), is defined as 
\[ k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]. \] [3]

Example 1 (Gaussian Process) A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. Formally, let \( \{f(x) : x \in \mathbb{R}^n\} \) be a set of random variables. We call \( \{f(x) : x \in \mathbb{R}^n\} \) a Gaussian Process if for all \( n \in \mathbb{N} \) and every finite subset \( \{x_1, \ldots, x_n\} \) of \( \mathbb{R}^n \), the random vector \( (f(x_1), \ldots, f(x_n)) \) has a multivariate Gaussian distribution. The covariance and mean functions parameterize our Gaussian Process denoted \( f(x) \sim GP(m(x), k(x, x')) \). [3]

Definition 20 Let \( Y_1, \ldots, Y_n \) be a set of samples from density \( f(\cdot | \theta) \), where \( \theta \) is the value of random variable \( \Theta \) with density function \( g(\cdot) \). The density \( g(\cdot) \) represents our assumptions about the random variable \( \Theta \) before observed data is taken into consideration and is called the prior distribution. [4]

Definition 21 Building off the previous definition, the conditional density of \( \Theta \) given observations \( y_1, \ldots, y_n \) denoted \( f(\theta | y_1, \ldots, y_n) \) is called the posterior distribution. This function takes into account previous assumptions and evidence from observed data. [4]

Definition 22 Bayesian Inference derives a posterior probability distribution utilizing Bayes’ Rule, a prior probability distribution and a likelihood function derived from a probability model for the observed data. [4]

Theorem 1 Bayes’ Rule Let \( \{Y_1, \ldots, Y_n\} \) be a set of random variables with joint probability density \( f(\cdot, \ldots, \cdot) \). Let \( A \) and \( B \) both be subsets of \( \{Y_1, \ldots, Y_n\} \)
\[ f(A|B) = \frac{f(A)f(B|A)}{f(B)} \]

Definition 23 A hyperparameter is a parameter of the prior, not to be confused with the parameters of the model.

Definition 24 If \( X_1, \ldots, X_n \) is a set of \( n \) random samples from the joint density function \( f(\cdot, \ldots, \cdot | \theta) \) with parameters \( \theta \), the Likelihood Function is defined to be
\[ L(\theta | x_1, \ldots, x_n) = P(x_1, \ldots, x_n | \theta) = f(x_1, \ldots, x_n | \theta) = f(x_1 | \theta) \times \cdots \times f(x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta) \]
This function gives the likelihood that the random variables will be the particular observed values \( x_1, \ldots, x_n \). [4]
Example 2 (Multivariate Gaussian Likelihood Function) Let $Y = (Y_1, \ldots, Y_n)$ denote a random vector from a Gaussian Process with multivariate Gaussian distribution. Let $\Sigma$ be the covariance matrix and $\mu$ be the expected value of $Y$, as defined above. Then the likelihood function is

$$L(\mu, \Sigma|Y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{n/2}} e^{-(Y-\mu)'\Sigma^{-1}(Y-\mu)/2}$$

Definition 25 Marginal Likelihood If $X_1, \ldots, X_n$ is a set of $n$ random samples from the joint density function $f(\cdot, \ldots, \cdot|\theta)$ with parameters $\theta$ from density function $g(\cdot)$, the Marginal Likelihood is defined to be

$$P(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} f(x_1, \ldots, x_n|\theta) g(\theta) d\theta$$
Chapter 2

The Gaussian Process

Recall that a Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. Utilizing the information gained from a training dataset \{(x_i, y_i) | i = 1, \ldots, n\} where \(x_i\) is the \(i\)th row of \(X\) and \(y_i\) is the \(i\)th value of \(y\), we wish to make output predictions for all new inputs \(x_*\). In our case the random variables stand for the values of the function \(f(x)\) we wish to predict at input locations \(x\). More specifically, \(f(x_i)\) is the random variable corresponding to \(y_i\) in \((x_i, y_i)\). We will do this by constructing a continuous function or real process, known as a Gaussian Process. To do this, we first make some assumptions about the underlying function.

2.1 The Mean and Covariance Function

A particular Gaussian Process is specified solely by its mean and covariance function. It is common to take the mean function to be the zero function (for notational simplicity).

\[
m(x) = 0 \tag{2.1}
\]

We adopt this practice with further justification that our observed and interpolated methane values will tend towards zero; assuming that there will be negligible methane levels outside of a plume.

A common covariance function is the squared exponential covariance function (see equation 2.2).

\[
cov(f(x_i), f(x_j)) = k(x_i, x_j) = \exp\left(-\frac{1}{2}(x_i - x_j)^2\right) \tag{2.2}
\]

Let \(k(x_i, x_j)\) define the \(i, j\)th element of the covariance matrix \(K = K(X, X)\). Notice that the covariance between outputs \(f(x_i)\) and \(f(x_j)\) is only a function of the inputs \(x_i\) and \(x_j\). Furthermore, covariance defined in equation 2.2 is a decreasing function of distance resulting in a high correlation between inputs that are close and a negligible correlation between inputs that are far apart.

Modifying this covariance function slightly with the addition of the hyperparameter \(l\) (a positive constant, as in equation 2.3) allows for more control of the correlation between inputs (see equation 2.3). This hyperparameter \(l\) is called the length-scale. The length-scale dictates the characteristic of the covariance which Rasmussen and
Williams [6] informally describe as the distance that needs to be traveled in the input space before the interpolated function value can significantly change.

\[ k(f(x_i), f(x_j)) = \exp\left(-\frac{1}{2} \frac{(x_i - x_j)^2}{\ell^2}\right) \]  

(2.3)

Figure 2.1 shows a two-dimensional analysis of the influence of this hyperparameter on the interpolated output values. Note that the smaller the hyperparameter is, the further the distance needed for the function value to dramatically change. Conversely, the larger the hyperparameter, the shorter the distance needed. The addition of this hyperparameter allows for more accurate representation of the fluctuation of methane values within our model.

Figure 2.1: Figure taken from Rasmussen and Williams [6]. The blue crosses represent sample points, the gray shaded areas represent two standard deviations and the blue line represents the interpolated mean. Figures (a), (b) and (c) are the realizations of a Gaussian Process with hyperparameters 1, 0.3, and 3 respectively. Notice the smaller the hyperparameter is, the further distance needed for the function value to dramatically change. Conversely, the larger the hyperparameter, the shorter the distance needed.

Note since the covariance function specifies elements for the covariance matrix, for
any subset $X$ of the input space $\mathcal{X}$ will also have a joint Gaussian distribution $f(X) \sim N(0, K(X, X))$, thus $f \sim GP(0, K(\mathcal{X}, \mathcal{X}))$.

The selection of the zero function and the squared exponential covariance function with the addition of the characteristic length-scale for our mean and covariance functions fully define our Gaussian process. Both are selected to represent environmental factors prevalent when sampling methane plumes. We have defined a Gaussian process which embodies our assumptions as a joint prior distribution over our observations.

### 2.2 Bayesian Inference via Inclusion of Sampled Data

First for simplicity, consider a one dimensional input $x$ and output $f(x)$ with assumed mean 0 and variance $\sigma^2$. Now, consider a two element observed dataset $\{(x_1, y_1), (x_2, y_2)\}$. Bayesian inference is a well known statistical method for including observed information into prior assumptions about the data. This is done utilizing Bayes’ rule

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$

(2.4)

The marginal likelihood acts as the normalization constant making the result a probability distribution i.e. the posterior distribution. The posterior’s uncertainty decreases closer to the observations. A pictorial representation of this, taken from Rasmussen and Williams [6], can be seen in figure 2.2.

![Figure 2.2: Figure taken from Rasmussen and Williams [6].](image)
Within our model, we assume our observations to be noisy. More specifically, 
\( f(x_i) \) is the random variable corresponding to \((x_i, y_i)\) such that 
\( y_i = f(x_i) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2_n) \). This will alter our covariance function and covariance matrix as follows:

\[
\text{cov}(y_i, y_j) = k(x_i, x_j) + \sigma^2_n \delta_{ij} \tag{2.5}
\]

Or equivalently,

\[
\text{cov}(y) = K(X, X) + \sigma^2_n I \tag{2.6}
\]

Here, \( \delta_{ij}, I, \sigma^2_n \) are the Kronecker delta, identity matrix and noise variance respectively. We consider the joint distribution of the observed data \( y \) and function values \( f^* = f(x^*) \) at unobserved locations \( x^* \).

\[
\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma^2_n I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \tag{2.7}
\]

We use Bayesian inference to condition the joint Gaussian prior distribution on our observations. This is done by using Bayes’ rule,

\[
p(f^*|X, y, X^*) = \frac{L(y|X, f^*, X^*) \times p(f^*)}{p(y|X)} \tag{2.8}
\]

where \( L(y|X, f^*, X^*) \) is the Likelihood Function of the observations \( y \) given locations \( X \) and test locations \( X^* \). That is

\[
L(y|X, f^*, X^*) = p(y|X, f^*, X^*) \\
= p(y_1|x_1, f^*, X^*) \times \cdots \times p(y_n|x_n, f^*, X^*) \\
= \prod_{i=1}^{n} p(y_i|x_i, f^*, X^*) \tag{2.9}
\]

Additionally the denominator of equation 2.8, \( p(y|X) \), is the Marginal Likelihood.

\[
p(y|X) = \int_{-\infty}^{\infty} p(y|f, X)p(f|X) \, df \tag{2.10}
\]

These equations result in the predictive function (see Appendix A for further de-
tails):

\[
f_*(X, y, X_*) \sim N(K(X_*, X)[K(X, X) + \sigma^2_n I]^{-1} y, K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma^2_n I]^{-1} K(X, X_*))
\]

(2.11)

Note that if we have \( n \) observed data points, these calculations will require the inversion of an \( n \) by \( n \) matrix. There are several numerical analysis techniques used to minimize computational time for this that are discussed in Appendix B.

### 2.3 Optimization of Hyperparameters

Many characteristics of our predictive function are determined solely by the chosen covariance function. For this reason, it is important to pick a covariance function that can be tailored to the observations. Rasmussen and Williams refer to finding suitable properties for our covariance function as learning the Gaussian Process. Many covariance functions have adjustable parameters such as length scale \( l \), noise variance \( \sigma^2_n \) (both of which we briefly talked about above), and signal variance \( \sigma^2_f \) which we introduce in this section. The parameters can be inferred from the observed data utilizing marginal likelihood. We refer to this Bayesian method as the optimization of the hyperparameters.

Consider again the covariance function and covariance matrix that include the hyperparameters listed above.

\[
cov(y_i, y_j) = \sigma^2_f \exp\left(-\frac{1}{2} \frac{(x_i - x_j)^2}{l^2}\right) + \sigma^2_n \delta_{ij}
\]

(2.12)

\[
cov(y) = \sigma^2_f K(X, X) + \sigma^2_n I = \sigma^2_f K + \sigma^2_n I
\]

(2.13)

Maximizing the marginal likelihood \( p(y|X) \) in this case is equivalent to maximizing the equation below.

\[
\log p(y|X) = -\frac{1}{2} y^T (\sigma^2_f K + \sigma^2_n I)^{-1} y - \frac{1}{2} \log |\sigma^2_f K + \sigma^2_n I| - \frac{n}{2} \log 2\pi
\]

(2.14)

Maximization is done by taking partial derivatives of the above equation with respect to the three different parameters and solving the system.

\[
\frac{\partial}{\partial \sigma^2_f} \log p(y|X) = -\frac{1}{2} y^T \frac{\partial}{\partial \sigma^2_f} (\sigma^2_f K + \sigma^2_n I)^{-1} y - \frac{1}{2} \log |\sigma^2_f K + \sigma^2_n I| = 0
\]

(2.15)
\[
\frac{\partial}{\partial \sigma_n^2} \log p(y|X) = -\frac{1}{2} y^T \frac{\partial}{\partial \sigma_n^2} (\sigma_f^2 K + \sigma_n^2 I)^{-1} y - \frac{1}{2} \log \frac{\partial}{\partial \sigma_n^2} |\sigma_f^2 K + \sigma_n^2 I| = 0 \quad (2.16)
\]

\[
\frac{\partial}{\partial l} \log p(y|X) = -\frac{1}{2} y^T \frac{\partial}{\partial l} (\sigma_f^2 K + \sigma_n^2 I)^{-1} y - \frac{1}{2} \log \frac{\partial}{\partial l} |\sigma_f^2 K + \sigma_n^2 I| = 0 \quad (2.17)
\]

By maximizing the marginal likelihood with respect to these hyperparameters, we maximize the amount of evidence found for \( y \) given \( X \) marginalized over the function values \( f \). This fits our Gaussian Process and predictive functions to our observed data.
Chapter 3

Refinements to the Gaussian Process

3.1 Inclusion of Wind direction in the Covariance Function

For almost any natural gas plume, methane values will have higher correlation in certain directions depending on environmental conditions. For example, consider the case when the plume is elongated in a specific direction by the wind. If samples are taken perpendicular to the direction the methane plume is headed then there will be a more rapid fluctuation in methane values as the sampling position changes. If samples are taken parallel to the plume direction then there will be less fluctuation. As discussed in the previous chapter, the hyperparameter known as the length scale dictates how far you need to move in the input space for the function values to become essentially uncorrelated. In the standard squared exponential covariance function shown in equation 2.2, there is one length scale for all directions. We introduce in equation 3.1 the squared exponential covariance function with a separate length scale $l_i$ for each $i$th dimension of the input space,

$$
cov(y_i, y_j) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_i - x_j)^TM(x_i - x_j)\right) + \sigma_n^2 \delta_{i,j} \tag{3.1}
$$

where $M = diag(l)$ and $l = (l_1, l_2, \ldots, l_D)$ and $D$ is the dimension of the input space.

Inclusion of this covariance function in our model allows for flexibility in the correlation between Gaussian Process values along the different directional axes. By rotating our initial training data so that the wind direction is aligned with an axis, the minimization of the hyperparameters (specifically the length scales) will automatically incorporate the correlation structure. This is known as Automatic Relevance Detection (ARD). At any set distance between spatial points, a lower correlation will be seen perpendicular to the wind and a higher correlation parallel to the wind. This innovative idea allows for more accurate representation of environmental conditions, thus making our prediction model that much more flexible.

3.2 The log-Gaussian Process

When sampling methane, intuitively we know a majority of our measurements will have very small values with a limited number of extreme valued measurements closer to the plume source. This will make for a positive skew in the distribution of our sample values. (See histograms in Figure 3.2 of a gridded sample from the plume simulation in Figure 5) Analyzing the log of the data is a common statistical method for dealing with positively skewed data [5], thus our motivation for introducing the log-Gaussian Process.
Figure 3.1: Left: Histogram of sample values from Figure 5. Right: Histogram of the log of sample values from Figure 5.

**Definition 26** Let \( Y(x_i) \) be a random variable at location \( x_i \). For all possible locations, the set \( Y = \{Y(x_i) : x_i \in X\} \) denotes a log-Gaussian Process if \( \{Z\} = \log\{Y\} \) is a Gaussian Process. Thus \( Y(x_i) = \exp Z(x_i) \) is the original random measurement at location \( x_i \).

Taking the log of the data amplifies the differences between smaller measurements and dampens the differences between extreme measurements, therefore removing the skew. The log-Gaussian Process deals with the extreme measurement values better than the ordinary Gaussian Process. It creates lower spatial variability in a majority of the spatial field while allowing higher spatial variability in “hotspots”, such as near the plume emission source. [1]
Information theory is based off defining a measure on the amount of information in a system. Further, entropy is a common information theoretic measure of the variability of a random variable. Entropy is defined to be the negative information in a system (see definitions 27 and 28 below). By minimizing the amount of entropy in a system, we are maximizing the amount of information known. Maximum entropy sampling (MES) is formulated around sampling the areas with maximum entropy, eliminating the variability in that area and thus maximizing the amount of information known. [7] We wish to sample where there is a maximum amount of variability so that the unsampled areas have minimal variability, making MES the ideal method.

**Definition 27** Consider the continuous random variable \( X \) with density function \( p(x) \). The differential information of \( X \) is defined as

\[
I(x) = \int_{-\infty}^{\infty} p(x) \log p(x) \, dx \quad (4.1)
\]

[9]

**Definition 28** Consider the continuous random variable \( X \) with density function \( p(x) \). The differential entropy of \( X \) is defined as

\[
H(x) = -I(x) \quad (4.2)
\]

[9]

Say we have a sampling area \( S = [1, \ldots, n] \) containing \( n \) possible sampling locations with values represented by random variables \( Y_1, \ldots, Y_n \). Notice that \( H(Y_S) \) will be a fixed value. Let \( s \) be a initially sampled area of \( S \) and \( \bar{s} \) be the unsampled area such that \( s \cup \bar{s} = S \). From common information theory results,

\[
H(Y_S) = H(Y_s) + E_{Y_S}\{H(Y_{\bar{s}}|Y_s)\}
\]

As mentioned previously we aim to minimize the unknown information about the area. This is the same as minimizing the entropy of the unsampled random variables given the sampled random variables, or minimizing \( E_{Y_S}\{H(Y_{\bar{s}}|Y_s)\} \). Since \( H(Y_S) \) is fixed, minimizing \( E_{Y_S}\{H(Y_{\bar{s}}|Y_s)\} \) is equivalent to maximizing \( H(Y_s) \), hence the term Maximum Entropy Sampling. (For more theoretical support for the idea of MES, see Shewry and Wynn. [7])

Considering each unsampled spatial point individually, we only need to derive the entropy of a single random variable with a univariate Gaussian distribution.
Lemmas 1, 2 and 3 are needed to derive the entropy for random variables with both Gaussian and log-Gaussian univariate distributions. Proofs are included in Appendix C.

Lemma 1
\[ \int_{-\infty}^{\infty} t e^{-t^2} = 0 \] (4.3)

Lemma 2
\[ \int_{-\infty}^{\infty} e^{-t^2} = \sqrt{\pi} \] (4.4)

Lemma 3
\[ \int_{-\infty}^{\infty} t^2 e^{-t^2} = \frac{1}{2} \sqrt{\pi} \] (4.5)

Theorem 2 The entropy of a random variable \( X \) with a Gaussian Distribution with mean \( \mu \) and variance \( \sigma^2 \) is \( \frac{1}{2} \log(2\pi\sigma^2e) \). The logarithm can be in any base (i.e. 10, \( e \), etc.)

Proof.

\[
H(x) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx \\
= - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \log\left(\frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}\right) dx \\
= - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \left[- \log(\sqrt{2\pi\sigma}) - \frac{(x-\mu)^2}{2\sigma^2} \log(e)\right] dx \\
= \int_{-\infty}^{\infty} \frac{\log(\sqrt{2\pi\sigma})}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} dx + \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} \log(e) e^{-(x-\mu)^2/2\sigma^2} dx \\
= \int_{-\infty}^{\infty} \frac{\log(\sqrt{2\pi\sigma})}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} dx + \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} \log(e) e^{-(x-\mu)^2/2\sigma^2} dx
\]

Let:
\[
t = \frac{x - \mu}{\sqrt{2\sigma}} \\
x = \sqrt{2\sigma} t + \mu \\
dx = \sqrt{2\sigma} dt
\[
\begin{align*}
&\int_{-\infty}^{\infty} \log(\sqrt{2\pi\sigma}) e^{-t^2/2\sigma^2} dt + \int_{-\infty}^{\infty} t^2 \log(e) e^{-t^2/2\sigma^2} dt \\
&= \int_{-\infty}^{\infty} \log(\sqrt{2\pi\sigma}) e^{-t^2/2\sigma^2} dt + \int_{-\infty}^{\infty} \log(e) t^2 e^{-t^2/2\sigma^2} dt \\
&= \log(\sqrt{2\pi\sigma}) \int_{-\infty}^{\infty} e^{-t^2/2\sigma^2} dt + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2/2\sigma^2} dt \\
&= \log(\sqrt{2\pi\sigma}) + \frac{1}{2} \log(e) \\
&= \frac{1}{2} \log(\sqrt{2\pi\sigma})^2 + \frac{1}{2} \log(e) \\
&= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \log(e) \\
&= \frac{1}{2} \log(2\pi\sigma^2 e)
\end{align*}
\]

The result is a strictly increasing function of variance, thus it will suffice to choose the spatial location with maximal calculated variance as the next sample destination. Since variance will be the highest in areas that are sparsely sampled, our exploration will be directed towards those areas.

**Theorem 3** The entropy of a random variable with a log-Gaussian Distribution with mean \( \mu \) and variance \( \sigma^2 \) is \( \frac{1}{2} \log(2\pi\sigma^2 e) + \mu \). The logarithm can be in any base (i.e. 10, \( e \), etc.)

**Proof.**

\[
H(x) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx \\
= -\int_{-\infty}^{\infty} \frac{1}{x\sqrt{2\pi\sigma}} e^{-(\ln(x) - \mu)^2/2\sigma^2} \log\left[ \frac{1}{x\sqrt{2\pi\sigma}} e^{-(\ln(x) - \mu)^2/2\sigma^2} \right] dx \\
= -\int_{-\infty}^{\infty} \frac{1}{x\sqrt{2\pi\sigma}} e^{-(\ln(x) - \mu)^2/2\sigma^2} \left[ - \log(x\sqrt{2\pi\sigma}) - \frac{(\ln(x) - \mu)^2}{2\sigma^2} \log(e) \right] dx \\
= \int_{-\infty}^{\infty} \frac{\log(x\sqrt{2\pi\sigma})}{x\sqrt{2\pi\sigma}} e^{-(\ln(x) - \mu)^2/2\sigma^2} dx + \int_{-\infty}^{\infty} \frac{(\ln(x) - \mu)^2}{2\sigma^2} \log(e) e^{-(\ln(x) - \mu)^2/2\sigma^2} dx \\
= \int_{-\infty}^{\infty} \frac{\log(x\sqrt{2\pi\sigma})}{x\sqrt{2\pi\sigma}} e^{-(\ln(x) - \mu)^2/2\sigma^2} dx + \int_{-\infty}^{\infty} \frac{(\ln(x) - \mu)^2}{2\sigma^2} \log(e) e^{-(\ln(x) - \mu)^2/2\sigma^2} dx
\]
Let:
\[ t = \ln(x) - \mu \]
\[ x = e^{\sqrt{2} \sigma t + \mu} \]
\[ dx = \sqrt{2} \sigma x dt \]

\[ = \int_{-\infty}^{\infty} \log(e^{\sqrt{2} \sigma t + \mu}) \frac{e^{-t^2}}{2\sigma \sqrt{\pi}} dt + \int_{-\infty}^{\infty} \frac{t^2 \log(e)}{2\sigma \sqrt{\pi}} e^{-t^2} \sqrt{2} \sigma x dt \]
\[ = \int_{-\infty}^{\infty} \frac{\log(e^{\sqrt{2} \sigma t + \mu})}{\sqrt{\pi}} \frac{e^{-t^2}}{2\sigma \sqrt{\pi}} dt + \int_{-\infty}^{\infty} \frac{\log(e)}{\sqrt{\pi}} t^2 e^{-t^2} dt \]
\[ = \int_{-\infty}^{\infty} \left\{ \sqrt{2} \sigma t + \mu \right\} e^{-t^2} dt + \frac{\log(\sqrt{2} \pi \sigma)}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{\log(e)}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2} dt \]
\[ = \frac{\sqrt{2} \sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} t e^{-t^2} dt + \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{\log(\sqrt{2} \pi \sigma)}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{\log(e)}{\sqrt{\pi}} \frac{1}{2} \sqrt{\pi} \]
\[ = 0 + \frac{\mu}{\sqrt{\pi}} + \log(\sqrt{2} \pi \sigma) + \frac{1}{2} \log(e) \]
\[ = \mu + \frac{1}{2} \log(\sqrt{2} \pi \sigma)^2 + \frac{1}{2} \log(e) \]
\[ = \mu + \frac{1}{2} \log(2 \pi \sigma^2) + \frac{1}{2} \log(e) \]
\[ = \mu + \frac{1}{2} \log(2 \pi \sigma^2 e) \]

The entropy of a random variable with a univariate log-Gaussian distribution is similar to that of the Gaussian distribution with the addition of the mean. We refer to “hotspots” as the locations with the highest mean values, thus this mean term directs us to use what is commonly called “hotspot sampling”. Low et al. defines a hotspot to be an area exhibiting extreme measurements and much higher spatial variability than the rest of the spatial field [1]. Physically, we know the spatial locations with the greatest fluctuations in methane values will be near the highest concentrations, which motivates greater sampling in those areas. Selection of the location with the maximal value of this new entropy will allow for not only exploration in sparsely sampled regions, but will also emphasize sampling in those “hotspot” areas.
Chapter 5

Simulated Results

Figure 5.1: Horizontal slice from a three dimensional simulated methane plume.

We began testing our algorithm on a three dimensional simulated methane plume. By taking horizontal slices at several different constant elevations, we were able to test our interpolation method on multiple two dimensional simulations of a methane plume. We discuss the results specific to the two dimensional slice taken at a constant elevation seen in Figure 5.1.

Our initial simulations sampled both a finely and coarsely spaced grids. Letting the sample locations be \( X \) and the pixel values (representing methane values) be \( y \), we assume \((X,y)\) are generated by some Gaussian Prior process with specified mean and covariance function. The hyperparameters were minimized using the gridded samples taken from the images as initial training data. The algorithm then interpolated the mean and covariance values for the pixel locations between those sampled. We reconstructed the mean values into the images seen in Figures 5.2 and 5.3. Figure 5.2 shows the interpolated means of every 2nd pixel in the image for the finely spaced sample grid (every 50th pixel value). Figure 5.3 shows the interpolated means of every 2nd pixel in the image for the more coarsely spaced sample grid (every 75th pixel value).
For the images seen in the top left corners of Figure 5.2 and 5.3, we assume \((X, y)\) are generated by a Gaussian Prior process with mean function 0 and standard covariance function given by equation 2.3.

Maintaining the method but altering our assumptions such that \((X, y)\) are assumed to be samples from a Gaussian prior Process with mean 0 and covariance function equation 3.1 with wind aligned axis, we obtained the images seen in the top right of Figure 5.2 and 5.3.

Further altering our assumption such that \((X, y)\) are assumed to be samples from a log-Gaussian prior Process with mean function 0 and covariance function equation 2.2, we reconstructed the images in the bottom left corners of Figure 5.2 and 5.3.

Lastly, when \((X, y)\) are assumed to be samples from a log-Gaussian prior Process with mean function 0 and covariance function equation (3.1) with wind aligned axis,
we obtained the images in the bottom right corners of Figure 5.2 and 5.3.

For both the fine and coarse sampled grids, the greatest correlation of values are seen in the simulations where covariance function is assumed to be equation (3.1) with wind-aligned axis. More accurate representation of the extreme fluctuations in methane values are also seen in the simulations which assume $(X, y)$ to be from a log-Gaussian prior distribution. Note when the coarse sample is assumed to be from a Gaussian prior distribution with covariance function equation (2.2), the simulation was not able to reconstruct the plume. However the plume was able to be reconstructed when the sample was assumed to be from a log-Gaussian prior or the covariance was assumed to be equation (3.1) with wind alignment.

For a more concrete evaluation of the simulations we include a residual analysis in Table 5.1. The values shown are the sums of the errors of the interpolated pixel values compared to the actual pixel values.
<table>
<thead>
<tr>
<th>Grid Spacing</th>
<th>Regular</th>
<th>Wind</th>
<th>log</th>
<th>Wind and log</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine</td>
<td>17.10</td>
<td>16.74</td>
<td>18.47</td>
<td>16.04</td>
</tr>
</tbody>
</table>

Table 5.1: Residuals of Simulations

In addition we show plots of the residuals for the finely spaced grid in Figure 5.5 to illustrate the areas of greatest accuracy for each simulation method. In each plot, the difference between the interpolated images (figure 5.2) and the original image (figure 5.1) is indicated by color at each point (red being higher, blue being lower).
5.1 Maximum Entropy Sampling Simulations

Beginning with one initial sample from the image in Figure 5.6, we calculate the entropy for each spatial point. This has almost no additional computational cost since entropy is a function of each point’s variance and mean (which are already interpolated). Our algorithm is constructed to select the next closest location with the maximum entropy for sampling, indicated in Figures 5.7 and 5.8 by green dots. We constructed our algorithm to sample methane values along the path between successive points of maximal entropy as well, shown in Figures 5.7 and 5.8 by yellow lines. Each time a point of maximal entropy is reached, the Gaussian Process (or log-Gaussian Process) is relearned (the hyperparameters are re-minimized) from the new comprehensive set of samples. The next closest location with the maximum entropy is calculated and the algorithm continues until 500 samples are taken.

In Figure 5.7, the image on the left shows part of the sampling path based on MES of a Gaussian Process. We arbitrarily selected the initial sampling location to be in the bottom right corner of the image. Notice that the sampling takes on an even pattern spreading out over the entire spatial field regardless of methane values. This is a result of entropy being a function of variance (see Theorem 2) where variance was assumed to be a decreasing function of distance. Thus the farther away from sampled locations, the higher the entropy. Although this method does sample the entire spatial field relatively evenly, it does not sample the plume sufficiently for the Gaussian Process algorithm to interpolate it well, seen in the right image of Figure

Figure 5.5: Plume used to simulate MES.
5.7.

Figure 5.6: MES applied when sample assumed to be Gaussian Process.

Alternatively in Figure 5.8, the image on the left shows the sampling path based on MES of a log-Gaussian Process. We arbitrarily selected the initial sampling location to be in the middle left of the image. Notice that the sampling does explore some of the spatial field but focuses on the areas with higher sampled methane values. This is a result of entropy being a function of variance and mean (see Theorem 3). As discussed previously, the addition of the mean term will make the entropy greater in places where the mean is interpolated to be more extreme. Although the log-Gaussian Process approach does sample one methane plume sufficiently to interpolate it with accuracy, the other methane plume is not found at all, as seen in the right image of Figure 5.8.

Figure 5.7: MES applied when sample assumed to be log-Gaussian Process.
5.2 Field Data Results

A field experiment was done in Wyoming in late June of 2013. The orthocopter was able to sample a methane plume at 3 different altitudes. Figure 5.10 shows the data as dot colors to illustrate the methane values, green being low to red being high.

Excluding the first 100 measurements or so (which are assumed to be unstable due to take off conditions) we assume the data \((X, y)\) to be from a Gaussian prior Process where \(X\) is the three dimensional spatial coordinate and \(y\) are the observed methane values. Figure 5.11 shows the interpolated three dimensional methane plume for values greater than or equal to 2ppm. The red dots indicate the sampling locations while the plume is depicted by a solid blue surface in the left image and a scatter plot of methane values (red being high, blue being low) in the right image.
Figure 5.9: Collected Data

Figure 5.10: Interpolated methane plume of values greater than or equal to 2ppm
Chapter 6
Conclusion

This thesis describes approaches to sampling, predicting and visualizing a methane field utilizing statistical, machine learning and information theory techniques. Statistical background was provided before introducing the Gaussian Process as a prior distribution over interpolated functions. Gaussian Processes are completely specified by the choice of the mean and covariance functions. We introduced several covariance functions which represent physical attributes of methane plumes. Using Bayesian methods, observed data is included in the model through a posterior distribution and the optimization of hyperparameters is achieved utilizing the marginal likelihood function.

Innovations to the model include the consideration of wind direction in the covariance functions and hyperparameters. In addition, the log-Gaussian process is utilized to deal with the positive skew in our data and more accurately model hotspots within the methane field. We couple this statistical interpolation method with an information theoretic approach to determining where to take samples known as maximum entropy sampling. When paired with a Gaussian Process, entropy is a strictly increasing function of variance thus resulting in relatively even sampling throughout the field. However when paired with a log-Gaussian process, entropy is a function of both the variance and the mean resulting in more sampling in hotspot areas where a higher mean is interpolated.

We discuss and compare these techniques applied to sampling and reconstruct simulated methane plumes. The results suggest greater accuracy in sampling when utilizing the wind aligned axis and log-Gaussian Process. Applying the prediction scheme to observed data, we are able to successfully reconstruct and visualize a methane plume.

Future work for the project includes modification to our suggested maximum entropy sampling technique to consider the entropy of a collection of points in the spatial field simultaneously as oppose to analyzing each one individually. Field experiments coupling our interpolation method with the maximum entropy sampling technique are also yet to be done. These techniques in combination with future modifications will provide a robust and innovative approach to sampling, predicting and visualizing methane gas within a spatial field.
Bibliography


Appendix A: Gaussian Properties

The goal of Appendix A is to establish the regression function. The following theorems are necessary for the derivation of the predictive function previously presented in equation 2.11.

Theorem 4

\[(A + CBD)^{-1} = A^{-1} - A^{-1}C(B^{-1} + DA^{-1}C)^{-1}DA^{-1}\]

**Proof.**

\[
(A + CBD)(A + CBD)^{-1} = (A + CBD)[A^{-1} - A^{-1}C(B^{-1} + DA^{-1}C)^{-1}DA^{-1}]

= (A + CBD)A^{-1} - (A + CBD)A^{-1}C(B^{-1} + DA^{-1}C)^{-1}DA^{-1}

= I + CBDA^{-1} - (C + CBDA^{-1}C)(B^{-1} + DA^{-1}C)^{-1}DA^{-1}

= I + CBDA^{-1} - CB(B^{-1} + DA^{-1}C)(B^{-1} + DA^{-1}C)^{-1}DA^{-1}

= I + CBDA^{-1} - CBDA^{-1}

= I
\]

Theorem 5

Let a matrix and its inverse be defined as follows:

\[
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}
\]

\[
A^{-1} = B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix}
\]

Then

\[
B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{12}^T)^{-1}
\]

\[
B_{22} = (A_{22} - A_{12}^T A_{11}^{-1} A_{12})^{-1}
\]

\[
B_{12}^T = -A_{22}^{-1} A_{12}^T (A_{11} - A_{12}A_{22}^{-1} A_{12}^T)^{-1}
\]

\[
B_{22} = -A_{11}^{-1} A_{12} (A_{22} - A_{12}^T A_{11}^{-1} A_{12})^{-1}
\]

**Proof.**

\[
AA^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix}
\]

\[
= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{12}^T & A_{11}B_{12} + A_{12}B_{22} \\ A_{12}^T B_{11} + A_{22}B_{12}^T & A_{12}^T B_{12} + A_{22}B_{22} \end{pmatrix}
\]

\[
= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}
\]
This gives the equations:

\[ A_{11}B_{11} + A_{12}B_{12}^T = I \]
\[ A_{11}B_{12} + A_{12}B_{22} = 0 \]
\[ A_{12}^T B_{11} + A_{22}B_{12}^T = 0 \]
\[ A_{12}^T B_{12} + A_{22}B_{22} = I \]

These can be manipulated to be the following:

\[ B_{11} = A_{11}^{-1} - A_{11}^{-1}A_{12}B_{12}^T \]
\[ B_{12} = -A_{11}^{-1}A_{12}B_{22} \]
\[ B_{12}^T = -A_{22}^{-1}A_{12}^TB_{11} \]
\[ B_{22} = A_{22}^{-1} + A_{22}^{-1}A_{12}^TB_{12} \]

Plugging \( B_{12}^T \) into \( B_{11} \):

\[ B_{11} = A_{11}^{-1} - A_{11}^{-1}A_{12}(-A_{22}^{-1}A_{12}^TB_{11}) \]
\[ B_{11} - A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^TB_{11} = A_{11}^{-1} \]
\[ (I - A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^T)B_{11} = A_{11}^{-1} \]
\[ A_{11}(I - A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^T)B_{11} = A_{11}A_{11}^{-1} \]
\[ (A_{11} - A_{12}A_{22}^{-1}A_{12}^T)B_{11} = I \]
\[ B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{12}^T)^{-1} \]

A similar process can be done to find the following:

\[ B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{12}^T)^{-1} \]
\[ B_{22} = (A_{22} - A_{12}^TA_{11}^{-1}A_{12})^{-1} \]
\[ B_{12}^T = -A_{22}^{-1}A_{12}^T(A_{11} - A_{12}A_{22}^{-1}A_{12}^T)^{-1} \]
\[ B_{22} = -A_{11}^{-1}A_{12}(A_{22} - A_{12}^TA_{11}^{-1}A_{12})^{-1} \]

Theorem 6 \( |A| = |A_{11}||A_{22} - A_{12}^TA_{11}^{-1}A_{12}| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{12}^T| \)

Proof.

\[ |A| = \begin{vmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{vmatrix} \]
\[ = \begin{vmatrix} A_{11} & 0 \\ A_{12}^T & I \end{vmatrix} \begin{vmatrix} I & A_{12}^{-1}A_{12} \\ 0 & A_{22} - A_{12}^TA_{11}^{-1}A_{12} \end{vmatrix} \]
\[ = \begin{vmatrix} A_{11} & 0 \\ A_{12}^T & I \end{vmatrix} | \begin{vmatrix} I & A_{12}^{-1}A_{12} \\ 0 & A_{22} - A_{12}^TA_{11}^{-1}A_{12} \end{vmatrix} | \]
\[ = |A_{11}||A_{22} - A_{12}^TA_{11}^{-1}A_{12}| \]
Furthermore:

\[ |A| = \begin{vmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{vmatrix} \]

\[ = \begin{vmatrix} I & A_{12} \\ 0 & A_{22} \end{vmatrix} \begin{vmatrix} A_{11} - A_{12}A_{22}^{-1}A_{12}^T & 0 \\ A_{22}^{-1}A_{21} & I \end{vmatrix} \]

\[ = \begin{vmatrix} I & A_{12} \\ 0 & A_{22} \end{vmatrix} \begin{vmatrix} A_{11} - A_{12}A_{22}^{-1}A_{12}^T & 0 \\ A_{22}^{-1}A_{21} & I \end{vmatrix} \]

\[ = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{12}^T| \]

**Theorem 7** Let \( x = [x_1 \, x_2] \sim N (\mu_1, \Sigma ; \Sigma_{21} \, \Sigma_{22}) \)

**Proof.** The joint density of \( x \) is:

\[ f(x) = f(x_1, x_2) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]

We denote the inverse as the following: \([\Sigma_{11} \, \Sigma_{12} \, \Sigma_{21} \, \Sigma_{22}]^{-1} = [\Sigma_{11}^{11} \, \Sigma_{12}^{12}] \]

Consider the exponent,

\[ (x - \mu)^T \Sigma^{-1} (x - \mu) = [(x_1 - \mu_1)^T, (x_2 - \mu_2)^T] [\Sigma_{11}, \Sigma_{21}, \Sigma_{21}^T, \Sigma_{22}] [(x_1 - \mu_1), (x_2 - \mu_2)] \]

\[ = (x_1 - \mu_1)^T \Sigma_{11} (x_1 - \mu_1) + 2(x_1 - \mu_1)^T \Sigma_{12} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma_{22} (x_2 - \mu_2) \]

Applying Theorems 4 and 5,

\[ \Sigma_{11} = (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^T)^{-1} = \Sigma_{11}^{-1} - \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12} \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12} \Sigma_{11}^{-1} \Sigma_{12}^T \Sigma_{11}^{-1} \]

\[ \Sigma_{22} = (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} = \Sigma_{22}^{-1} - \Sigma_{22}^{-1} \Sigma_{12}^T (\Sigma_{11} + \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12})^{-1} \Sigma_{12} \Sigma_{22}^{-1} \]

\[ \Sigma_{21} = -\Sigma_{22}^{-1} \Sigma_{12}^T (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12})^{-1} \]

\[ \Sigma_{12} = -\Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12})^{-1} = (\Sigma_{21})^T \]

Replacing these in the above exponent,
\[(x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + 2(x_1 - \mu_1)^T \Sigma_{12}^{-1} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) = (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)
+ 2(x_1 - \mu_1)^T \Sigma_{12}^{-1} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)
\]

\[= (x_1 - \mu_1)^T \Sigma_{11}^{-1} + \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12} \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12} \Sigma_{11}^{-1}(x_1 - \mu_1)
+ 2(x_1 - \mu_1)^T \Sigma_{12}^{-1} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)
\]

\[= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)
+ 2(x_1 - \mu_1)^T \Sigma_{12}^{-1} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)
\]

\[= (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)
+ 2(x_1 - \mu_1)^T \Sigma_{12}^{-1} (x_2 - \mu_2) + (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2)
\]

Plugging this in for the exponent, we get the joint density to be:

\[f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + \Sigma_{12}^{-1} (x_1 - \mu_1 - (x_2 - \mu_2))^T (\Sigma_{22} - \Sigma_{12} \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\Sigma_{12} \Sigma_{11}^{-1} (x_1 - \mu_1) - (x_2 - \mu_2))\right)
\]

\[= \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1) + \Sigma_{12}^{-1} (x_1 - \mu_1 - (x_2 - \mu_2))^T (\Sigma_{22} - \Sigma_{12} \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\Sigma_{12} \Sigma_{11}^{-1} (x_1 - \mu_1) - (x_2 - \mu_2))\right)
\]
According to Theorem 6, we know $|\Sigma| = |\Sigma_{11}| |\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}|$. This leads to the following:

$$f(x) = f(x_1)f(x_2|x_1) = \frac{1}{(2\pi)^n/2 |\Sigma_{11}|^{1/2}} e^{-\frac{1}{2} (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)} \times$$

$$\frac{1}{(2\pi)^n/2 |\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}|^{1/2}} e^{-\frac{1}{2} (\Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1) - (x_2 - \mu_2))^T (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1) - (x_2 - \mu_2))}$$

$$= \frac{1}{(2\pi)^n/2 |\Sigma_{11}|^{1/2}} e^{-\frac{1}{2} (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)} \times$$

$$\frac{1}{(2\pi)^n/2 |\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}|^{1/2}} e^{-\frac{1}{2} ((x_2 - \mu_2) - \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1))^T (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} ((x_2 - \mu_2) - \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1))}$$

$$= \frac{1}{(2\pi)^n/2 |\Sigma_{11}|^{1/2}} e^{-\frac{1}{2} (x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)} \times$$

$$\frac{1}{(2\pi)^n/2 |\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}|^{1/2}} e^{-\frac{1}{2} (x_2 - \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1))^T (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})^{-1} (x_2 - \mu_2 - \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1))}$$

Therefore we have:

$$x_2|x_1 \sim N(\mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (x_1 - \mu_1), \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})$$
Appendix B: Numerical Analysis

As mentioned above, computing the marginal likelihood requires the inversion of a \( n \) by \( n \) covariance matrix which becomes more and more computationally heavy as \( n \) increases. Fortunately, there are several methods for matrix inversion of symmetric positive definite matrices, one being Cholesky Decomposition which we will discuss here. First, we provide background and prove that a multivariate Gaussian covariance matrix \( \Sigma \) is symmetric positive definite.

**Definition 29** A matrix \( A \in \mathbb{R}^{n \times n} \) is defined to be **Symmetric** if \( A = A^T \). More specifically if \( a_{i,j} \) denotes the \( i,j \)th element of \( A \), then \( a_{i,j} = a_{j,i} \) for all integers \( i \) and \( j \) such that \( 1 \leq i, j \leq n \).

**Definition 30** A symmetric matrix \( A \in \mathbb{R}^{n \times n} \) is said to be **Positive Semi-Definite** if the following is true:

- \( x^T Ax \geq 0 \) for all \( x \in \mathbb{R}^n \)

Similarly a symmetric matrix \( A \in \mathbb{R}^{n \times n} \) is said to be **Positive Definite** if:

- \( x^T Ax > 0 \) for all \( x \neq 0 \)

**Theorem 8** All covariance matrices are symmetric, positive semi-definite.

**Proof.** Let \( A \in \mathbb{R}^{n \times n} \) be a covariance matrix for random vector \( Y = Y_1, \ldots, Y_n \). By definition of a covariance matrix, the \( i,j \)th component of \( A \) denoted \( a_{i,j} \) is defined to be

\[
\text{cov}(Y_i, Y_j) = E[(Y_i - \mu_i)(Y_j - \mu_j)]
= E[(Y_j - \mu_j)(Y_i - \mu_i)]
= \text{cov}(Y_j, Y_i)
\]

For all integers \( i \) and \( j \) such that \( 1 \leq i, j \leq n, a_{i,j} = a_{j,i} \). This implies \( A = A^T \).

Therefore, covariance matrix \( A \) is symmetric.

Let \( x \in \mathbb{R}^n \) and \( U \in \mathbb{R} \) such that \( U = (Y - E[Y])x \)

\[
x^T Ax = x^T E[(Y - E[Y])(Y - E[Y])]x 
= E[x^T (Y - E[Y])^T (Y - E[Y])x]
= E[U^T U]
= E[U^2]
\geq 0
\]

Therefore the covariance matrix \( A \) is positive semi-definite.
Theorem 9 If \( A \in \mathbb{R}^{n \times n} \) be a symmetric matrix with strictly positive eigenvalues, then \( A \) is positive definite.

**Proof.** Assume all eigenvalues of \( A \) are strictly positive \( \lambda_1, \ldots, \lambda_n > 0 \). Since \( A \) is real and symmetric then \( \lambda_1, \ldots, \lambda_n \in \mathbb{R} \) and we can choose orthogonal eigenvalues \( o_i \) for each \( \lambda \) (if an eigenvalue has multiplicity \( m \), then \( m \) orthogonal eigenvectors can be chosen for that eigenvalue). Finally, we can normalize each eigenvector such that \( o_i^T o_j = 1 \) for \( i = j \) and \( o_i^T o_j = 0 \) for \( i \neq j \). By definition of eigenvectors and eigenvalues \( Ao_i = \lambda o_i \) for integers \( i \in [1, n] \). Therefore we can define a diagonal matrix \( D = I[\lambda_1, \ldots, \lambda_n]^T \) and orthogonal matrix \( O \) with \( i \)th column equal to \( o_i \) such that \( D = O^T A O \) and \( A = ODO^T \). Let \( x \in \mathbb{R}^n \) not equal to the zero vector and consider

\[
x^T A x = x^T O D O^T x = y^T D y \quad \text{where} \quad y = O^T x \neq 0 = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 > 0
\]

Therefore, \( A \) is positive definite.

Theorem 10 A multivariate Gaussian covariance matrix \( \Sigma \) is symmetric, positive definite.

**Proof.** By definition of a multivariate Gaussian distribution, the covariance matrix \( \Sigma \) is non-singular. \( \Sigma \) being non-singular implies that its eigenvalues are distinct and non-zero. Furthermore for each eigenvalue \( \lambda_i \) there is a normalized eigenvector \( x_i \) such that

\[
x_i^T \Sigma x_i = \lambda_i x_i^T x_i
\]

\( \Sigma \) is semi-positive definite thus this implies all eigenvalues of \( \Sigma \) must greater than or equal to zero, making them strictly positive. Therefore, by the above theorem \( \Sigma \) is positive definite.

Theorem 11 LU-Decomposition A non-singular, positive definite matrix \( A \) can be decomposed into a product of a lower triangular matrix \( L \) and an upper matrix \( U \).

\[
A = LU
\]

Note we are only considering the case when \( A \) is non-singular (since we know our Gaussian covariance matrix to be non-singular).

**Proof.** Since \( A \) is invertible, we can perform the matrix form of Gaussian elimination by multiplying \( A \) by lower triangular elementary matrices \( L_1, \ldots, L_n \) until
$L_n \cdots L_1 A = U$ where $U$ is an upper triangular matrix. Thus $A = L_1^{-1} \cdots L_n^{-1} U = LU$ where $L = L_1^{-1} \cdots L_n^{-1}$ is a lower triangular matrix.

Theorem 12 Cholesky Decomposition A non-singular, symmetric, positive definite matrix $A$ can be decomposed into a product of a lower triangular matrix $B$ and its transpose $B^T$.

$$A = BB^T$$

Proof. Let $A = LU$ be the LU-decomposition of the matrix $A$ where $L$ is a lower and $U$ is an upper triangular matrix. Since $A = A^T$ by definition of symmetric

$$LU = U^T L^T$$

$$U(L^T)^{-1} = (L^{-1})U^T$$

Note the left side of the equality is an upper triangular matrix and the right side is a lower triangular matrix. Let $D$ be a diagonal matrix.

$$D = U(L^T)^{-1} = (L^{-1})U^T$$

$$D = U(L^T)^{-1}$$

$$DL^T = U$$

$$A = LU = LDL^T$$

Let $D = \sqrt{D}$ and $B = L\sqrt{D}$ a lower triangular matrix.

$$A = L\sqrt{D} L^T$$

$$A = L\sqrt{D} L^T$$

$$A = BB^T$$
Appendix C: Additional Proofs

The following Lemmas are needed to derive the entropy for random variables with both Gaussian and log-Gaussian univariate distributions.

Proof. Lemma 1
Let:
\[ u = -x^2 \]
\[ du = -2x \, dx \]
\[ dx = -\frac{1}{2x} \, du \]
\[
\int_{-\infty}^{\infty} x e^{-x^2} \, dx = -\frac{1}{2} \int_{-\infty}^{\infty} e^u \, du \\
= -\frac{1}{2} e^u \bigg|_{x=-\infty}^{x=\infty} \\
= -\frac{1}{2} e^{-x^2} \bigg|_{x=-\infty}^{x=\infty} \\
= -\frac{1}{2} (0 - 0) = 0
\]

Proof. Lemma 2 The Gaussian Integral
Let
\[
I = \int_{-\infty}^{\infty} e^{-x^2} \, dx \\
I^2 = \int_{-\infty}^{\infty} e^{-x^2} \, dx \int_{-\infty}^{\infty} e^{-y^2} \, dy \\
I^2 = \int_{-\infty}^{\infty} e^{-x^2} \, dx \int_{-\infty}^{\infty} e^{-y^2} \, dy \\
I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} \, dx \, dy
\]
Let:
\[ r^2 = x^2 + y^2 \]
Jacobian
\[ = r \, dr \, d\theta \]
\[
I^2 = \int_{0}^{2\pi} \int_{0}^{\infty} r e^{-r^2} \, dr \, d\theta
\]
Let:
\[ u = -r^2 \]
\[ du = -2rdr \]
\[ dr = -\frac{1}{2r} du \]

\[ I^2 = \int_0^{2\pi} \left[ -\frac{1}{2} \int_0^\infty e^u du \right] d\theta \]
\[ I^2 = \int_0^{2\pi} -\frac{1}{2} [e^u \big|_0^\infty] d\theta \]
\[ I^2 = -\frac{1}{2} \int_0^{2\pi} [0 - 1] d\theta \]
\[ I^2 = \frac{1}{2} \int_0^{2\pi} 1 d\theta \]
\[ I^2 = \frac{1}{2} (\theta |_0^{2\pi}) \]
\[ I^2 = \frac{1}{2} (2\pi) \]
\[ I^2 = \pi \]
\[ I = \sqrt{\pi} \]

**Proof.** Lemma 3

\[ \int_{-\infty}^{\infty} e^{-x^2} dx = -\frac{1}{2} \int_{-\infty}^{\infty} x(-2xe^{-x^2}) dx \]

Let:

\( u = x \)
\( v = e^{-x^2} \)
\( \frac{du}{dx} = dx \)
\( \frac{dv}{dx} = -2xe^{-x^2} \)

\[ -\frac{1}{2} \int_{-\infty}^{\infty} u \frac{dv}{dx} dx = -\frac{1}{2} [uv|_{-\infty}^{\infty}] + \frac{1}{2} \int_{-\infty}^{\infty} v \frac{du}{dx} dx \]
\[ = -\frac{1}{2} [uv|_{-\infty}^{\infty}] + \frac{1}{2} \int_{-\infty}^{\infty} v \frac{du}{dx} dx \]
\[ = -\frac{1}{2} xe^{-x^2}|_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} e^{-x^2} dx \]
\[ = 0 + \frac{1}{2} \sqrt{\pi} \]
\[ = \frac{1}{2} \sqrt{\pi} \]