Suitably impressive thesis title

Lennart Golks

Department of Physics
University of Otago

A thesis submitted for the degree of
Doctor of Philosophy

November 2025

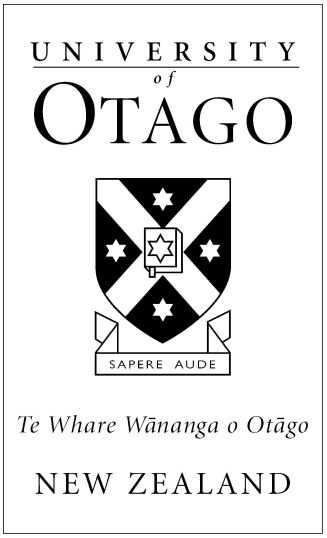
Abstract

Your abstract text goes here. Check your departmental regulations, but generally this should be less than 300 words. See the beginning of Chapter 1 for more.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet nibh volutpat, scelerisque nibh a, vehicula neque. Integer placerat nulla massa, et vestibulum velit dignissim id. Ut eget nisi elementum, consectetur nibh in, condimentum velit. Quisque sodales dui ut tempus mattis. Duis malesuada arcu at ligula egestas egestas. Phasellus interdum odio at sapien fringilla scelerisque. Mauris sagittis eleifend sapien, sit amet laoreet felis mollis quis. Pellentesque dui ante, finibus eget blandit sit amet, tincidunt eu neque. Vivamus rutrum dapibus ligula, ut imperdiet lectus tincidunt ac. Pellentesque ac lorem sed diam egestas lobortis.

Suspendisse leo purus, efficitur mattis urna a, maximus molestie nisl. Aenean porta semper tortor a vestibulum. Suspendisse viverra facilisis lorem, non pretium erat lacinia a. Vestibulum tempus, quam vitae placerat porta, magna risus euismod purus, in viverra lorem dui at metus. Sed ac sollicitudin nunc. In maximus ipsum nunc, placerat maximus tortor gravida varius. Suspendisse pretium, lorem at porttitor rhoncus, nulla urna condimentum tortor, sed suscipit nisi metus ac risus.

Suitably impressive thesis title



Lennart Golks
Department of Physics
University of Otago

A thesis submitted for the degree of
Doctor of Philosophy
November 2025

Acknowledgements

Personal

I would like to thank Alex Elliott for his wonderful help and support. None of this would be possible otherwise.

This is where you thank your advisor, colleagues, and family and friends.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum feugiat et est at accumsan. Praesent sed elit mattis, congue mi sed, porta ipsum. In non ullamcorper lacus. Quisque volutpat tempus ligula ac ultricies. Nam sed erat feugiat, elementum dolor sed, elementum neque. Aliquam eu iaculis est, a sollicitudin augue. Cras id lorem vel purus posuere tempor. Proin tincidunt, sapien non dictum aliquam, ex odio ornare mauris, ultrices viverra nisi magna in lacus. Fusce aliquet molestie massa, ut fringilla purus rutrum consectetur. Nam non nunc tincidunt, rutrum dui sit amet, ornare nunc. Donec cursus tortor vel odio molestie dignissim. Vivamus id mi erat. Duis porttitor diam tempor rutrum porttitor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed condimentum venenatis consectetur. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Aenean sit amet lectus nec tellus viverra ultrices vitae commodo nunc. Mauris at maximus arcu. Aliquam varius congue orci et ultrices. In non ipsum vel est scelerisque efficitur in at augue. Nullam rhoncus orci velit. Duis ultricies accumsan feugiat. Etiam consectetur ornare velit et eleifend.

Suspendisse sed enim lacinia, pharetra neque ac, ultricies urna. Phasellus sit amet cursus purus. Quisque non odio libero. Etiam iaculis odio a ex volutpat, eget pulvinar augue mollis. Mauris nibh lorem, mollis quis semper quis, consequat nec metus. Etiam dolor mi, cursus a ipsum aliquam, eleifend venenatis ipsum. Maecenas tempus, nibh eget scelerisque feugiat, leo nibh lobortis diam, id laoreet purus dolor eu mauris. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nulla eget tortor eu arcu sagittis euismod fermentum id neque. In sit amet justo ligula. Donec rutrum ex a aliquet egestas.

Institutional

If you want to separate out your thanks for funding and institutional support, I don't think there's any rule against it. Of course, you could also just remove the subsections and do one big traditional acknowledgement section.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut luctus tempor ex at pretium. Sed varius, mauris at dapibus lobortis, elit purus tempor neque, facilisis sollicitudin felis nunc a urna. Morbi mattis ante non augue blandit pulvinar. Quisque nec euismod mauris. Nulla et tellus eu nibh auctor malesuada quis imperdiet quam. Sed eget tincidunt velit. Cras molestie sem ipsum, at faucibus quam mattis vel. Quisque vel placerat orci, id tempor urna. Vivamus mollis, neque in aliquam consequat, dui sem volutpat lorem, sit amet tempor ipsum felis eget ante. Integer lacinia nulla vitae felis vulputate, at tincidunt ligula maximus. Aenean venenatis dolor ante, euismod ultrices nibh mollis ac. Ut malesuada aliquam urna, ac interdum magna malesuada posuere.

Abstract

Your abstract text goes here. Check your departmental regulations, but generally this should be less than 300 words. See the beginning of Chapter 1 for more.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet nibh volutpat, scelerisque nibh a, vehicula neque. Integer placerat nulla massa, et vestibulum velit dignissim id. Ut eget nisi elementum, consectetur nibh in, condimentum velit. Quisque sodales dui ut tempus mattis. Duis malesuada arcu at ligula egestas egestas. Phasellus interdum odio at sapien fringilla scelerisque. Mauris sagittis eleifend sapien, sit amet laoreet felis mollis quis. Pellentesque dui ante, finibus eget blandit sit amet, tincidunt eu neque. Vivamus rutrum dapibus ligula, ut imperdiet lectus tincidunt ac. Pellentesque ac lorem sed diam egestas lobortis.

Suspendisse leo purus, efficitur mattis urna a, maximus molestie nisl. Aenean porta semper tortor a vestibulum. Suspendisse viverra facilisis lorem, non pretium erat lacinia a. Vestibulum tempus, quam vitae placerat porta, magna risus euismod purus, in viverra lorem dui at metus. Sed ac sollicitudin nunc. In maximus ipsum nunc, placerat maximus tortor gravida varius. Suspendisse pretium, lorem at porttitor rhoncus, nulla urna condimentum tortor, sed suscipit nisi metus ac risus.

Contents

List of Figures

List of Abbreviations

i.i.d. . . . . . independent and identically distributed

MRF . . . . . Markov Random Field

GMRF . . . . Gaussian Markov Random Field

MTC . . . . . Marginal Then Conditional sampler

GOMOS . . . Global Ozone Monitoring by Occultation of Stars

MCMC . . . . Markov Chain Monte-Carlo

MH . . . . . Metropolis-Hastings

1 Background

This chapter provides the background knowledge for the applied methods. We try to carefully guide the reader through the various topics so that we build a base to understand the used methodology. Firstly, we discuss Markov Random Fields and their key properties to describe spatially connected variables in Section 1.1.1. In Section 1.1.2 we introduce the normal distribution to describe the random variables in such a field. By specifying the local connections in Section 1.1.3 we globally observe a Gaussian Markov Random Field (GMRF). Next, in Section 1.2 we relate the variables in the GMRF to a measurement process through a Bayesian Model. Given the measurement, we try to find the most likely variables through a Markov chain, as described in Section 1.2.2. Constructing this Markov chain we present the Metropolis-Hastings algorithm in Section 1.3.1, followed by the Marginal and then conditional sampler in Section 1.3.2 and the adaptive Metropolis-Hastings algorithm in Section 1.3.3.

1.1 Gaussian Markov Random Field

In this section, we guide the reader through the basic principles of Gaussian Markov Random Fields (GMRF). We discuss some of the key properties of GMRFs and

how to set up such a spatial field. The discussed topics are strongly related to our specific application and for further reading, we refer to the books [1, 2].

1.1.1 Markov Random Field

A Markov Random Field (MRF) is based on a graphical model describing spatial dependencies of n nodes, also described as vertices. We label each node with an integer i and attach a spatial position $s^{(i)}$ and a random variable $x^{(i)}$, where $i = 1, \dots, n$. On a set of all vertices $\mathcal{V}$ we define an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with a set of edges $\mathcal{E}$. If an edge describes spatial dependencies of a pair of different nodes i and j then we write $i \sim j$, where $(i, j) \in \mathcal{E}$, with no preferred direction. We limit interactions to pair-wise only.

For a subset of nodes $A = \mathcal{V}^A \subset \mathcal{V}$, we can define a subgraph $\mathcal{G}^A(\mathcal{V}^A, \mathcal{E}^A)$, including edges $\mathcal{E}^A = \{(i, j) \in \mathcal{E} \text{ and } i, j \in \mathcal{V}^A \text{ and } i \neq j\}$. A discrete example is illustrated in Figure 1.1.

Through the Markov properties of such a field, a single node i is conditionally dependent on its neighbors $\partial i = \{j : (i, j) \in \mathcal{E} \text{ and } j \neq i\}$ only. We denote $-A = \mathcal{V} - A$ to all nodes not in A but part of $\mathcal{V}$. Similarly, we can extend this to a subset $A \in \mathcal{V}$ where the neighborhood $\partial A = \{j : (i, j) \in \mathcal{E} \text{ and } j \notin A\}$ is all nodes $j \in -A$ sharing an edge to a node $i \in \mathcal{V}^A$. Conditioned on all its neighbours $\mathcal{V}^A$ is independent of all other notes $\pi(\boldsymbol{x}^{(A)}|\boldsymbol{x}^{(-A)}) = \pi(\boldsymbol{x}^{(A)}|\boldsymbol{x}^{(\partial A)})$. Here $\pi(\cdot)$ denotes the probability density for its argument. Intuitively we interpret $\boldsymbol{Q}$ in terms of conditional dependencies and $\boldsymbol{\Sigma}$ marginally, which reduces distribution to maximal n dimensional.

1.1.2 Multivariate Normal Distribution

In our case, we specifically choose a probability density that distributes each random variable $x^{(i)}$ normally and denote this with $x^{(i)} \sim \mathcal{N}(\mu^{(i)}, \sigma)$. Here $\mu_i = \mathrm{E}(x^{(i)})$ is the mean of $x^{(i)}$ and $\sigma^2 = \mathrm{Var}(x^{(i)})$ is the variance of any $x^{(i)}$.

Constructing a Gaussian Markov Random Field (GMRF), we introduce a normal probability density function π for a single node x_i, with mean μ and variance σ^2.

$$\pi(x^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[- \frac{(x^{(i)} - \mu^{(i)})^2}{2\sigma^2} \right] \tag{1.1}$$

To describe the values of all nodes in a GMRF we introduce a multivariate Gaussian

$$\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu^{(1)} \\ \vdots \\ \mu^{(n)} \end{bmatrix}, \boldsymbol{\Sigma} \right), \tag{1.2}$$

with a covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$, which is the inverse of the precision-matrix. As we deal with an undirected Markov field, $\boldsymbol{Q}$ is a symmetric positive-definite matrix and sparse. Hence, we can efficiently decompose $\boldsymbol{Q} = \mathbf{L}\mathbf{L}^T$ into its Cholesky triangle $\mathbf{L}$, which is a lower triangular matrix. Within this precision matrix, we can define the interactions and correlations between the different nodes of our GMRF.

When constructing a GMRF the precision matrix $\boldsymbol{Q}$ as well as its inverse $\boldsymbol{\Sigma}$ does a play key role. If the GMRF is fully connected them $\boldsymbol{Q}$ is completely dense, we aim to construct a GMRF so that the precision matrix is sparse. The off diagonal non-zero entries of $\boldsymbol{Q}$ provide information about conditional correlations $\text{Corr}(x^{(i)}, x^{(j)} | \boldsymbol{x}^{(-ij)}) = Q_{ij}/\sqrt{Q_{ii}Q_{jj}}$. Where as the off diagnopal entries of $\boldsymbol{\Sigma}$ give information about marginal correlations $\text{Corr}(x^{(i)}, x^{(j)}) = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$. The diagonal entries provide marginal variances $\text{Var}(x^{(i)}) = \Sigma_{ii}$ and conditional precision $\text{Prec}(x^{(i)}, x^{(j)} | \boldsymbol{x}^{(-ij)}) = Q_{ii}$.

1.1.3 Conditional distribution through Spatial Dependencies

In this subsection, we present the Conditional Auto Regressive (CAR) model and the Gibbs field to define a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and the corresponding precision matrix $\boldsymbol{Q}$. For curious readers we recommend the book [3].

The CAR method can include neighbouring nodes of numerous orders. We define the entries of the precision matrix as follows:

$$Q_{ij} = \begin{cases} \kappa_i & i = j \\ \kappa_i \beta_{ij} & i \neq j \end{cases}, \tag{1.3}$$

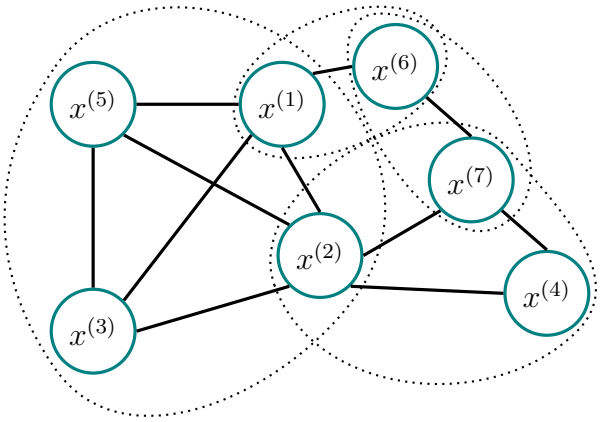


Figure 1.1: Undirected Random Field of seven nodes including maximum cliques. The random field displayed here can be described with a Graph $\mathcal{G}$. Where $\mathcal{V}$ are all nodes seen in the figure and $\mathcal{E}$ are the edges drawn with a line between nodes. We draw dotted lines around the set of maximum cliques, which are not a full subset of any other clique. Maximum cliques are a set of fully connected nodes with common neighbors. A neighborhood is defined through an edge between two nodes, such as $\{x^{(1)}, x^{(6)}\} \in \mathcal{E}$. Here $\{\{x^{(1)}, x^{(6)}\}, \{x^{(1)}, x^{(3)}, x^{(4)}, x^{(5)}\}, \{x^{(6)}, x^{(7)}\}, \{x^{(2)}, x^{(4)}, x^{(7)}\}\}$ form the set of maximum cliques. On a set of maximum cliques, we can define an energy function, Gibbs distribution or use the conditional auto regressive model to define spatial dependencies. The precision matrix $\boldsymbol{Q}$ represents those spatial dependencies weighted by the hyper parameter $\boldsymbol{\theta}$.

where $\beta_{ij} = 0$, if $x^{(i)}$ and $x^{(j)}$ are conditionally independent, and $\kappa_i > 0$ to ensure positive-definiteness of $\boldsymbol{Q}$ for all $i = 1, \ldots, n$. The full conditionals are normally distributed

$$x^{(i)}|x^{(-i)} = x^{(i)}|\partial x^{(i)} \sim \mathcal{N}\Big(\underbrace{\sum_{j:j\neq i} \beta_{ij} x^{(j)}}_{\text{mean}}, \underbrace{\kappa_i^{-1}}_{\text{variance}} \Big) \tag{1.4}$$

and depend only on neighbouring nodes $\partial x^{(i)}$ having an edge to $x^{(i)}$. We can then write the probability density over all nodes $\boldsymbol{x}$ including their means $\boldsymbol{\mu}$ as follows:

$$\pi(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \sqrt{\det(\boldsymbol{Q})} \ \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu}) \right] \tag{1.5}$$

$$= \frac{1}{(2\pi)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right]. \tag{1.6}$$

Alternatively, we can describe a set of nodes through a Gibbs field, where we define a clique $c \in \mathcal{C}$ as a set of fully connected nodes. In a clique each node $j \in c$

has mutual neighbours to any other node $i \neq j \in c$. If for any node $j \notin c$ and $j \in \mathcal{V}$, the union of $j \cup c$ is not a clique then we call $\tilde{c} \in \mathcal{C}$ a maximal clique. We conclude that maximal cliques are not a complete subset of another clique. The positive clique-potential $\phi_{\tilde{c}}(\tilde{c})$ describes the interactions of all nodes within a clique. The sum over all maximum clique-potentials is often described as the Energy:

$$E = \sum_{\tilde{c} \in \mathcal{C}} \phi_{\tilde{c}}(x^{(i)} : x^{(i)} \in c). \tag{1.7}$$

We normalize the Gibbs distribution $\pi(\boldsymbol{x})$ of a random field with a normalization constant Z.

$$\pi(\boldsymbol{x}) = \frac{1}{Z} \exp\Big[- \sum_{\tilde{c} \in \mathcal{C}} \phi_{\tilde{c}}(x^{(i)} : x^{(i)} \in \tilde{c})\Big] \tag{1.8}$$

$$Z = \sum_{i \in \mathcal{V}} \prod_{\tilde{c} \in \mathcal{C}} \exp\Big[- \phi_{\tilde{c}}(x^{(i)} : x^{(i)} \in \tilde{c})\Big] \tag{1.9}$$

The Hammersley-Clifford theorem states that if we describe any Random Field through a Gibbs distribution over maximum cliques then we deal with a MRF. First proven by Julien Besag we can show that conditional joint distribution is expressed as

$$\pi(x^{(i)}|\partial x^{(i)}) = \frac{1}{Z} \exp\Big[- \sum_{\tilde{c} \in \mathcal{C}} \phi_{\tilde{c}}(x^{(i)}, x^{(j)} : i, j \in \tilde{c} \text{ and } i \neq j)\Big] \tag{1.10}$$

$$Z = \sum_{i \in \mathcal{V}} \prod_{\tilde{c} \in \mathcal{C}} \exp\Big[- \phi_{\tilde{c}}(x^{(i)}, x^{(j)} : i, j \in \tilde{c} \text{ and } i \neq j)\Big] \tag{1.11}$$

and can be found in [4]. The potential $\phi_{\tilde{c}}$ can describe the interactions between different nodes of a Gaussian Gibbs field if set accordingly, which leads to the precision-matrix $\boldsymbol{Q}$ for a Gaussian Markov Random Field (GMRF).

1.2 linear-Gaussian hierarchical Bayesian model

Bayesian hierarchical models are a very helpful tool to describe a measurement process through different layers of complexity. It allows us to backtrack from our observations to the source of the measurement in a very precise/simplified way. Here we closely follow the terminology of [5, 6]

Usually, a measurement device delivers some data $\boldsymbol{y}$ including some noise σ and we like to model the observation to quantify some unknown parameters $\boldsymbol{x}$

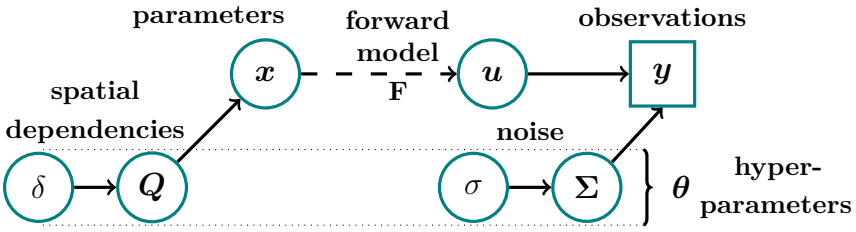


Figure 1.2: A Bayesian hierarchical model describes how we model a physical process to observe data. Through a forward model we describe a physical process depending on some parameters $\boldsymbol{x}$. The spatial dependencies of those parameters $\boldsymbol{x}$ are described through the hyper-parameters $\boldsymbol{\theta} = (\delta, \sigma)$. The precision matrix $\boldsymbol{Q}$ is defined according to our Markov Random field and the amount of interaction between certain nodes is described by δ. The forward map $\boldsymbol{F}$ models a physical process and maps the parameters $\boldsymbol{x}$ to a space of all measurable $\boldsymbol{u}$. From this space $\boldsymbol{u}$, we observe some data including some random noise γ with variance $\boldsymbol{\Sigma}(\gamma)$. A Bayesian framework allows us to find the most likely parameters and hyper-parameters $\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}$ given our measurement .

and their unknown spatial dependencies. Through the precision matrix $\boldsymbol{Q}(\delta)$, we define neighborhoods and weight pair-wise interactions of parameters $x^{(j)}$ and $x^{(i)}$ according to a hyper-prior distribution $\pi(\delta)$ in (1.12c). The inverse of the precision matrix $\boldsymbol{Q}(\delta)$ is the variance of the prior distribution $\pi(\boldsymbol{x}|\boldsymbol{Q}^{-1}(\delta))$ in (1.12b), and usually sparse. Based on this graphical model we map $\boldsymbol{x}$ through a forward model $\boldsymbol{F}$ Dimensions to space of all measurables $\boldsymbol{u}$. Next, we observe some data $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + \gamma$ and add some unknown normally distributed random noise η with variance $\boldsymbol{\Sigma}(\gamma)$. We assume that this noise vector is independent and identically distributed (i.i.d.). The likelihood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ in (1.12a) describes how likely/realistic the observations $\boldsymbol{y}$ are. In our specific case, we choose a linear forward model and a Gaussian prior, so that our likelihood function and posterior distribution in (1.12d) are normally distributed as well. In doing so we can find an analytic expression for the uncertainty of the posterior and see that it is depending on both the prior and the likelihood distribution, for more details we refer to the Appendix A.

In Equations (1.12b) - (1.12c), we summarize a hierarchical Bayesian model in a generalized form, with the hyper-parameter $\boldsymbol{\theta} = (\delta, \gamma)$. Note that in this case we define two hyper-parameters, but we can increase the dimensions of θ to an

arbitrary number suitable to describe the underlying physical process sufficiently.

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{F}\boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \qquad \text{(likelihood)} \tag{1.12a}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) \qquad \text{(prior)} \tag{1.12b}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \qquad \text{(hyper-prior)} \tag{1.12c}$$

The posterior has the following from

$$\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y} \sim \mathcal{N}(\mu_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}, (\boldsymbol{Q} + \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F})^{-1}) \quad \text{(posterior)} \tag{1.12d}$$

with the mean $\mu_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}} = \boldsymbol{\mu} + (\boldsymbol{Q} + \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu})$.

Using Bayes theorem we can find an expression for the posterior distribution $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ for the parameters $\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta}$ given the observed data $\boldsymbol{y}$.

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\,\pi(\boldsymbol{x},\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \tag{1.13}$$

If we like to compute the posterior distribution we have to compute the normalization constant $\pi(\boldsymbol{y})$, where we marginalize out $\boldsymbol{x}$ and $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{y}) = \int\int \pi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta}\,, \tag{1.14}$$

which can yield a extensive computation. Using the posterior density we can find the expected value of any function $h(\boldsymbol{x})$ though an average of the function weighted by the probability density distribution $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$.

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int\int h(\boldsymbol{x})\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta} \tag{1.15}$$

Taking a look at the posterior we can see that $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$.. and see that to solve this might be very tidious, where as the ratio is much easier...

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{1.16}$$

$$= \frac{\pi(\boldsymbol{\theta})}{\sqrt{\det(\boldsymbol{\Sigma})\,\det(\boldsymbol{Q})^{-1}}} \exp\left[\frac{(\boldsymbol{F}\boldsymbol{\mu} - \boldsymbol{y})^T(\boldsymbol{F}\boldsymbol{\mu} - \boldsymbol{y})}{\boldsymbol{\Sigma}} + \frac{(\boldsymbol{\mu} - \boldsymbol{x})^T(\boldsymbol{\mu} - \boldsymbol{x})}{\boldsymbol{Q}^{-1}}\right] \tag{1.17}$$

If we set $h(\boldsymbol{x}) = \boldsymbol{x}$ we get the mean of the parameter $\boldsymbol{x}$ or the posterior covariance if $h(\boldsymbol{x}) = (\boldsymbol{x} - \mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{x}])(\boldsymbol{x} - \mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{x}])^T$. Computing that is very expensive and usually not feasible as the parameters can too many different values, depending on the range and complexity of the problem. Instead, we can explore the parameter space for a few (e.g. 10^4) discrete values and use the Monte-Carlo method to approximate the integral.

1.2.1 Monte-Carlo Method Integration

The Monte-Carlo method was first developed in Los Alamos, United States of America, just after the II world war to simulate the flight path of neutrons. Later in 1949, N. Metropolis and S. Ulam formulated their ideas already suggesting to use 'Markoff' chains in Monte Carlo simulations to approximate continuous functions [7]. For more details on the fundamental theorems we use in this work, we refer to the books [8, 9]

Using the Monte-Carlo method we can approximate the expected value of any function $h(\boldsymbol{x})$ by the sample mean:

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int\int \boldsymbol{x}\, \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{N+1} \sum_{i=0}^{N} \boldsymbol{x}_i\,, \tag{1.18}$$

where we draw n parameter samples $\boldsymbol{x}_i$ from the posterior distribution $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$. We can argue with the central limit theorem that if $N \to \infty$ the sequence of $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_N\}$ converges the sample mean and sample variance to the true mean and true covariance of $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$. In our case, we know that the posterior is normally distributed and the task now is to draw samples from this distribution to perform such an integration. The problem is that it is not feasible to compute the posterior distribution due to the size of the parameter space. Instead, we construct a Markov chain of randomly picked parameters, which converges to the posterior distribution.

1.2.2 Markov Chains

First formulated by Andrey Markov and published in 1906, Markov chains are a very useful statistical tool to describe real-world processes [10]. In this section, we

discuss some of the properties of Markov chains to make sure that we sample from a unique equilibrium distribution, which is in our case the posterior distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. We like to show aperiodicity, irreducibility, and convergence towards a stationary distribution by proofing the detailed balance condition. If all of these properties hold we call a chain ergodic with a unique equilibrium distribution, e. g. $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. For more detailed information we refer to [3, 11].

In a finite parameter space $\Omega(\mathcal{X}, \theta)$ we draw a sequence of random variables $\mathcal{M} = \left\{\{\boldsymbol{x}_0, \boldsymbol{\theta}_0\}, \{\boldsymbol{x}_1, \boldsymbol{\theta}_1\}, \ldots, \{\boldsymbol{x}_N, \boldsymbol{\theta}_N\}\right\}$ from distributions $\pi^{(0)}, \ldots, \pi^{(N)}$. The Markov condition requires that each transition probability $\Pr(\cdot)$ of a new state $\{\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_{N+1}\}$ only depends on the current state $\{\boldsymbol{x}_N, \boldsymbol{\theta}_N\}$:

$$\Pr(\{\boldsymbol{x}_0, \boldsymbol{\theta}_0\}, \{\boldsymbol{x}_1, \boldsymbol{\theta}_1\}, \ldots, \{\boldsymbol{x}_N, \boldsymbol{\theta}_N\}) = \Pr(\{\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_{N+1}\}|\{\boldsymbol{x}_N, \boldsymbol{\theta}_N\}\,. \tag{1.19}$$

For reasons of simplicity in terminology we denote the state $\{\boldsymbol{x}_N, \boldsymbol{\theta}_N\}$ as vector $(\boldsymbol{x}_N^T, \boldsymbol{\theta}_N^T)^T = \boldsymbol{i}$ The probability for the Markov chain to be in that state $\boldsymbol{i}$ after N steps is denoted as $\pi_{\boldsymbol{i}}^{(N)} = \Pr((\boldsymbol{x}_N^T, \boldsymbol{\theta}_N^T)^T = \boldsymbol{i})$. We arrange the probability distribution of the Nth step as a row vector representing the probabilities to be in different states $\boldsymbol{i}, \boldsymbol{j} \in \Omega(\mathcal{X}, \theta)$

$$\pi^{(N)} = \left[\cdots \pi_{\boldsymbol{i}}^{(N)} \pi_{\boldsymbol{j}}^{(N)} \cdots\right] \tag{1.20}$$

A Markov chain $\mathcal{M}$ is homogeneous if the transition probability only depends on the value of the current and proposed state and not on n, the position in the chain.

$$\Pr(\{\boldsymbol{x}_{N+M+1}, \boldsymbol{\theta}_{N+M+1}\}|\{\boldsymbol{x}_{N+M}, \boldsymbol{\theta}_{N+M}\}) = \Pr(\{\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_{N+1}\}|\{\boldsymbol{x}_N, \boldsymbol{\theta}_n\}) \quad \forall M \in \mathbb{Z} \tag{1.21}$$

Then we denote the probability to transition from state $\{\boldsymbol{x}_N, \boldsymbol{\theta}_N\} = i$ to $\{\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_{N+1}\} = j$ as an element of the transition matrix $\boldsymbol{P}$:

$$P_{\boldsymbol{ij}} = \Pr((\boldsymbol{x}_{N+1}^T, \boldsymbol{\theta}_{N+1}^T)^T = \boldsymbol{j}|(\boldsymbol{x}_N^T, \boldsymbol{\theta}_N^T)^T = \boldsymbol{i}) \tag{1.22}$$

This matrix represents the transitions for one step from As we draw more and more states, some of those states are more likely to occur than others. The Chain will reach an equilibrium distribution $\pi^{(N)} \to \pi$ as $N \to \infty$. We call π the stationary

distribution for the transition matrix $\boldsymbol{P}$. Here, π is the left eigenvector of $\boldsymbol{P}$ with the eigenvalue one such that

$$\pi = \pi \boldsymbol{P} \,. \tag{1.23}$$

Once we draw values of that chain from the stationary distribution, the distribution does not change.

A Markov chain is irreducible if all states in Ω intercommunicate. That means for any two states $(\boldsymbol{x}_N^T, \boldsymbol{\theta}_N^T)^T = \boldsymbol{i}$ and $(\boldsymbol{x}_{N+1}^T, \boldsymbol{\theta}_{N+1}^T)^T = \boldsymbol{j}$ we can find a path with non-zero probability linking $\boldsymbol{i} \to \boldsymbol{j}$ and a path linking $\boldsymbol{i} \leftarrow \boldsymbol{j}$. Then the state space Ω is irreducible under $\boldsymbol{P}$. Within an irreducible chain, all states have the same period.

A period is the number of steps for a chain to revisit a set of states starting from the same set of states. An aperiodic chain has period one. If we are allowed to stay in a state i, so that the self-transition $P_{\boldsymbol{i},\boldsymbol{i}} > 0$, we can break any periodic pattern and the chain is aperiodic.

If a Markov chain is aperiodic and irreducible then this chain is ergodic. For an ergodic Markov chain on a finite state space Ω, there exists a stationary distribution π. This stationary distribution is unique if it satisfies the detailed balance condition.

$$\pi(\{\boldsymbol{x}_{N+1}, \boldsymbol{\theta}_{N+1}\} = \boldsymbol{j}) P_{\boldsymbol{j},\boldsymbol{i}} = \pi(\{\boldsymbol{x}_N, \boldsymbol{\theta}_N\} = \boldsymbol{i}) P_{\boldsymbol{i},\boldsymbol{j}} \tag{1.24}$$

$$\pi_{\boldsymbol{j}} P_{\boldsymbol{j},\boldsymbol{i}} = \pi_{\boldsymbol{i}} P_{\boldsymbol{i},\boldsymbol{j}} \tag{1.25}$$

In a large class of Markov chains the detailed balance condition is very helpful to find the stationary distribution.

As a consequence of the ergodic theorem, we observe an unique equilibrium distribution $\pi^{(N)} \to \pi$ as $N \to \infty$ independent of the initial distribution $\pi^{(0)}$. Now, we sample from the posterior distribution and can calculate the sample mean, as seen in Equation (1.18).

Next, we like to generate such a chain and define the transition probabilities using the Metropolis-Hastings algorithm.

1.3 Sampling from the posterior distribution

In this Section we present the major sampling algorithms which allow us to generate a Markov chain of posterior samples and to characterize the posterior distribution. We introduce the Metroplis-Hastings, Gibbs, MTC sampler, T-walk, adaptive MCMC.

1.3.1 Metropolis-Hastings - Markov chain Monte-Carlo

Here we introduce the Metropolis-Hastings algorithm, which is a Markov-chain Monte Carlo (MCMC) algorithm. First published in 1970 and based on the work of Metropolis et. al in 1953 this algorithm provides a framework to find samples from the posterior efficiently [12, 13]. Generating a Markov-Chain $\{\boldsymbol{x}_0, \boldsymbol{\theta}_0\}, \ldots, \{\boldsymbol{x}_j, \boldsymbol{\theta}_j\}, \ldots, \{\boldsymbol{x}_N, \boldsymbol{\theta}_N\}$ we accept and reject proposed samples to accurately calculate the sample mean.

Algorithm 1: Metropolis-Hastings step to generate a new candidate in a Markov chain

Let $\{\boldsymbol{x}_j, \boldsymbol{\theta}_j\}$ be the current state , then we **generate** $\{\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}\}$ as follows:

1 **Draw**: new state $\{\boldsymbol{x}', \boldsymbol{\theta}'\} \sim g(\boldsymbol{x}', \boldsymbol{\theta}' | \boldsymbol{x}_j, \boldsymbol{\theta}_j)$

2 **Acceptance probability**:

$$\alpha(j+1|j) \equiv \min\left\{1, \frac{\pi(\boldsymbol{x}', \boldsymbol{\theta}'|\boldsymbol{y}) g(\boldsymbol{x}_j, \boldsymbol{\theta}_j | \boldsymbol{x}', \boldsymbol{\theta}')}{\pi(\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}|\boldsymbol{y}) g(\boldsymbol{x}', \boldsymbol{\theta}' | \boldsymbol{x}_j, \boldsymbol{\theta}_j)}\right\}$$

3 **Draw**: $u \sim \mathcal{U}(0,1)$

4 **if** $u \leq \alpha(j+1|j)$ **then**

5 **Accept**: $\{\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}\} = \{\boldsymbol{x}', \boldsymbol{\theta}'\}$;

6 **else Reject**: $\{\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}\} = \{\boldsymbol{x}_j, \boldsymbol{\theta}_j\}$;

7 **end**

We generate a new state $\{\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}\}$ in our Markov-Chain by proposing a state $\{\boldsymbol{x}', \boldsymbol{\theta}'\}$ as described in Algorithm 1 we introduce the acceptance probability $\alpha(j+1|j)$ and the probability to generate a new state $g(\boldsymbol{x}', \boldsymbol{\theta}' | \boldsymbol{x}_j, \boldsymbol{\theta}_j)$ given a current state $\{\boldsymbol{x}_j, \boldsymbol{\theta}_j\}$. The transition probability $P_{\boldsymbol{j,j+1}}$ becomes:

$$\Pr(\{\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}\} | \{\boldsymbol{x}_j, \boldsymbol{\theta}_j\}) = g(\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1} | \boldsymbol{x}_j, \boldsymbol{\theta}_j) \alpha(\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1} | \boldsymbol{x}_j, \boldsymbol{\theta}_j) \tag{1.26}$$

Further, we define the acceptance probability $\alpha(j+1|j)$ as:

$$\alpha(j+1|j) \equiv \min\left\{1, \frac{\pi(\boldsymbol{x}', \boldsymbol{\theta}'|\boldsymbol{y}) g(\boldsymbol{x}_j, \boldsymbol{\theta}_j | \boldsymbol{x}', \boldsymbol{\theta}')}{\pi(\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}|\boldsymbol{y}) g(\boldsymbol{x}', \boldsymbol{\theta}' | \boldsymbol{x}_j, \boldsymbol{\theta}_j)}\right\} \tag{1.27}$$

Here we observe that the acceptance probability is constructed in such a way that we do not need to compute the full posterior distribution. It is sufficient to sample from a not normalized posterior distribution $\tilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ if we assume that the normalization constant $\pi(\boldsymbol{y}) > 0$.

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \tilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \tag{1.28}$$

Next, we draw a random number from a uniform distribution between 0 and 1. If this random number is smaller or equal then the acceptance probability of the proposed candidate, we accept this proposed state and set $\{\boldsymbol{x}_{j+1}, \boldsymbol{\theta}_{j+1}\} = \{\boldsymbol{x}', \boldsymbol{\theta}'\}$. If the uniformly drawn number is larger then we reject the proposal to stay in the current state. We can repeat this step for large enough n so that we generate an ergodic Markov chain. Sampling the hyper-parameters $\boldsymbol{\theta}$ and the parameters $\boldsymbol{x}$ is computationally very time consuming, to speed up the process we introduce the Marginal and then Conditional sample in the next Section.

Hence we like to sample from a very specific posterior distribution we show that the Markov chain, generated by the Metropolis-Hastings algorithm, fulfills the conditions for ergodicity in the Appendix B .

1.3.2 Marginal and then Conditional Sampler - MTC

Dealing with a hierarchical Bayesian model we have to sample $\boldsymbol{x}$ and $\boldsymbol{\theta}$ from the posterior distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ to find the mode of this distribution. Instead of computing full posterior samples $\{\boldsymbol{x}_i, \boldsymbol{\theta}_i\}$, we can speed up the sampling process by integration out $\boldsymbol{x}$ and independently sample hyper-parameters from the marginal posterior distribution $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ directly. Then we sample $\boldsymbol{x}_i \sim \pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_i)$ to characterize the posterior distribution $\{\boldsymbol{x}_i, \boldsymbol{\theta}_i\} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$.

Sampling from the marginal posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of the hyper-parameters $\boldsymbol{\theta}$

For the linear Bayesian hierarchical model we can eliminate $\boldsymbol{x}$, so that the marginal posterior distribution is given by: fix sqrt line

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \tag{1.29}$$

$$\propto \sqrt{\frac{\det(\boldsymbol{\Sigma}^{-1})\, \det(\boldsymbol{Q})}{\det(\boldsymbol{Q} + \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F})}} \times \exp\Big[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu})^T \boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}} (\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu})\Big] \pi(\boldsymbol{\theta})\,, \tag{1.30}$$

with the precision matrix

$$\boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F} + \boldsymbol{Q})^{-1}\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\,. \tag{1.31}$$

Note that this distribution is not Gaussian, [5].

Next, we generate a Markov chain $\{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_j, \ldots, \boldsymbol{\theta}_{K'}\}$ utilizing an MCMC algorithm on the conditional posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. At state $\boldsymbol{\theta}_j = \boldsymbol{\theta}$ we propose a new hyper-parameter sample $\boldsymbol{\theta}'$ from he proposal distribution $g(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and accept the new state with the probability according to the Metropolis-Hastings ratio:

$$1 \wedge \frac{\pi(\boldsymbol{\theta}'|\boldsymbol{y}) g(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}|\boldsymbol{y}) g(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{1.32}$$

Note that we have to compute the ratio $\pi(\boldsymbol{\theta}'|\boldsymbol{y})/\pi(\boldsymbol{\theta}|\boldsymbol{y})$. The MTC sampler is especially powerful in case of cheap evaluation of the determinants of the precision matrices, see Equation 1.30.

Once the Markov Chain $\{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_j, \ldots, \boldsymbol{\theta}_{K'}\}$ is long enough we calculate the integrated auto-correlation time τ_{int} of that chain. According to this measure and an appropriate burn in period we can refine the previous Markov chain to $\{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_j, \ldots, \boldsymbol{\theta}_K\}$, so that we have $K < K'$ independent samples of the marginal posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Sampling from the full conditional $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^{(i)})$

We can draw a sample $\boldsymbol{\theta}_i$ from the marginal by randomly choosing a state of just generated Markov chain $\{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_j, \ldots, \boldsymbol{\theta}_K\} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ Conditioned on $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_i)$ we use the Randomize Then Optimize (RTO) method by Bardsley et. al. to

draw a parameter sample $\boldsymbol{x}_i|\boldsymbol{y},\boldsymbol{\theta}_i$ [14, 15]. Here, we like to point out that the method has been introduced under various names and refer to Oliver et. al. and Oriuex et. al. for further reading [16, 17].

The conditional posterior is defined through the linear-Gaussian hierarchical Bayesian model as:

$$\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta} \sim \mathcal{N}\Big(\boldsymbol{\mu} + (\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F} + \boldsymbol{Q})^{-1}\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\mu}), (\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F} + \boldsymbol{Q})^{-1}\Big)\,, \quad (1.33)$$

for more details we refer to [2, 5, 6]. As the full conditional distribution for $\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}$ is a normal distribution we can rewrite to:

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \quad (1.34)$$

$$= \exp\|\hat{\boldsymbol{F}}\boldsymbol{x} - \hat{\boldsymbol{y}}\|^2\,, \quad (1.35)$$

where

$$\hat{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{F} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta}) \end{bmatrix}\,, \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{y} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta})\boldsymbol{\mu} \end{bmatrix}\,. \quad (1.36)$$

One sample from the posterior can be computed by minimizing the following equation with respect to $\hat{\boldsymbol{x}}$:

$$\boldsymbol{x}_i = \arg\min_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{F}}\hat{\boldsymbol{x}} - (\hat{\boldsymbol{y}} + \boldsymbol{\eta})\|^2, \quad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})\,, \quad (1.37)$$

where we add a randomized perturbation $\boldsymbol{\eta}$. Next, we substitute $-\hat{\boldsymbol{F}}^T\boldsymbol{\eta} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ we can rewrite the argument of Eq. 1.36 to

$$(\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F} + \boldsymbol{Q})\boldsymbol{x}_i = \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{v}_1 + \boldsymbol{v}_2\,, \quad (1.38)$$

where $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ are independent random variables. Finally, we can draw an independent sample from the posterior $(\boldsymbol{x}_i, \boldsymbol{\theta}_i) \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$.

1.3.3 Adaptive MCMC - GOMOS

Algorithm 2: Marginal and then Conditional (MTC) Sampler - Linear Gaussian Model

foreach $\boldsymbol{\theta} = \boldsymbol{\theta}_j \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K'}\}$ **do**
  Propose new state: $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  Acceptance probability: $\alpha(j+1|j) \equiv \min\left\{1, \frac{\pi(\boldsymbol{\theta}'|\boldsymbol{y})g(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}|\boldsymbol{y})g(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$
  Draw: $u \sim \mathcal{U}(0,1)$
  if $u \leq \alpha(j+1|j)$, **then**
    Accept: $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}'$;
    else Reject: $\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j$;
  end
end
Refine: $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K'}\}$ to $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ according to integrated auto-correlation time τ_{int} for large enough K', where $K < K'$
Draw: $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$
Draw: $\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{\theta}_i$ by solving $\boldsymbol{x}_i = \arg\min_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{F}}\hat{\boldsymbol{x}} - (\hat{\boldsymbol{y}} + \boldsymbol{\eta})\|^2$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Appendices

A

Posterior of Bayesian Hierachical model

Here we show how to obtain the posterior covariance and mean of our hierarchical Bayesian model in 1.12b - 1.12d. We do not consider the hyper-parameters and start with the joint probaility distribution of $(\boldsymbol{x}^T, \boldsymbol{y}^T)^T$, where $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ do not intersect. For more details we refer to Chapter 2 in [18] and to the book of Rue and Held [1].

The exponent of the normal Gaussian can be rewritten into:

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{\mu} + \text{const.} \tag{A.1}$$

We like to bring the joint distribution into a similar form so that we can compare the linear and second order terms and find the precision matrix and mean of the joint distribution.

In general the joint distribution to find the experssino for the postiror dostrbution

We can express this posterior through the likelihood and prior probability by Bayesian theorem, with a constant and positive normalization constant:

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}) \tag{A.2}$$

Taking the logarithmic function of this formulation we can find an expression for

the the posterior covariance, with the $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{Q}_{\boldsymbol{x}}^{-1}$ and $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Q}_{\boldsymbol{y}}^{-1}$.

$$\ln \pi(\boldsymbol{x}|\boldsymbol{y}) \propto \ln \pi(\boldsymbol{y}|\boldsymbol{x}) + \ln \pi(\boldsymbol{x}) \tag{A.3}$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{Q}_{\boldsymbol{x}}(\boldsymbol{x}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{Q}_{\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}) \tag{A.4}$$

$$= -\frac{1}{2}\Big[\boldsymbol{x}^T\Big[\boldsymbol{Q}_{\boldsymbol{x}} + \boldsymbol{A}^T\boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{A}\Big]\boldsymbol{x} + \boldsymbol{x}^T\Big[-\boldsymbol{A}^T\boldsymbol{Q}_{\boldsymbol{y}}\Big]\boldsymbol{y} \tag{A.5}$$

$$+ \boldsymbol{y}^T\Big[-\boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{A}\Big]\boldsymbol{x} + \boldsymbol{y}^T\Big[\boldsymbol{Q}_{\boldsymbol{y}}\Big]\boldsymbol{y} - 2\boldsymbol{x}^T\boldsymbol{Q}_{\boldsymbol{x}}\boldsymbol{\mu}\Big] + \text{const.} \tag{A.6}$$

Hence we deal with a Gaussian distribution, we consider second order terms only and rearrange to the precision matrix.

$$-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}^T\Big[\boldsymbol{Q}_{\boldsymbol{x}} + \boldsymbol{F}^T\boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{F}\Big] + \boldsymbol{y}^T\Big[-\boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{F}\Big] & \boldsymbol{y}^T\Big[\boldsymbol{Q}_{\boldsymbol{y}}\Big] + \boldsymbol{x}^T\Big[-\boldsymbol{F}^T\boldsymbol{Q}_{\boldsymbol{y}}\Big]\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \tag{A.7}$$

$$= \begin{bmatrix}\boldsymbol{x}^T & \boldsymbol{y}^T\end{bmatrix}\underbrace{\begin{bmatrix}\boldsymbol{Q}_x + \boldsymbol{F}^T\boldsymbol{Q}_y\boldsymbol{F} & -\boldsymbol{F}^T\boldsymbol{Q}_y\\ -\boldsymbol{Q}_y\boldsymbol{F} & \boldsymbol{Q}_y\end{bmatrix}}_{\text{precision matrix}}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \tag{A.8}$$

We denote the precision matrix of the joint field as:

$$\boldsymbol{Q}_{xy} = \begin{bmatrix}\boldsymbol{Q}_{aa} & \boldsymbol{Q}_{ab}\\ \boldsymbol{Q}_{ba} & \boldsymbol{Q}_{bb}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q}_x + \boldsymbol{F}^T\boldsymbol{Q}_y\boldsymbol{F} & -\boldsymbol{F}^T\boldsymbol{Q}_y\\ -\boldsymbol{Q}_y\boldsymbol{F} & \boldsymbol{Q}_y\end{bmatrix} \tag{A.9}$$

The mean is defined through the linear term.

$$\frac{-2\boldsymbol{x}^T\boldsymbol{Q}_{\boldsymbol{x}}\boldsymbol{\mu}}{-2} = \begin{bmatrix}\boldsymbol{x}^T & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}_{\boldsymbol{x}}\boldsymbol{\mu}\\ 0\end{bmatrix} \tag{A.10}$$

Comparing to the linear term of Equation A.1 we can formulate an expression for the joint mean:

$$\Rightarrow \boldsymbol{\mu}_{\boldsymbol{xy}} = \boldsymbol{Q}_{\boldsymbol{xy}}^{-1}\begin{bmatrix}\boldsymbol{Q}_{\boldsymbol{x}}\boldsymbol{\mu}\\ 0\end{bmatrix} \tag{A.11}$$

The mean of the conditional distribution $\boldsymbol{x}|\boldsymbol{y}$ is given by:

$$\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} = \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{Q}_{ba}^{-1}\boldsymbol{Q}_{ab}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{y}}) \tag{A.12}$$

$$\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} = \boldsymbol{\mu} + (\boldsymbol{Q}_x + \boldsymbol{F}^T\boldsymbol{Q}_y\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{Q}_y(\boldsymbol{x}-\boldsymbol{F}\boldsymbol{\mu})\,, \tag{A.13}$$

and the covariance of $\boldsymbol{x}|\boldsymbol{y}$ is given by:

$$\boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}} = \boldsymbol{Q}_{aa} = \boldsymbol{Q}_x + \boldsymbol{F}^T\boldsymbol{Q}_y\boldsymbol{F}\,, \tag{A.14}$$

as illustrated through Theorem 2.5 in [1].

B

Convergence of the Metropolis-Hastings

If we show that the detailed balance condition holds and that the state space is irreducible and aperiodic under the transition matrix $\boldsymbol{P}$, we generate a Markov chain with a unique stationary distribution proportional to $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. Since the posterior is strictly positive $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \geq 0$ on the finite state space $\Omega(\mathcal{X}, \theta)$ the generated chain is irreducable. Further, it is possible to reject any proposed state and stay in the current state, which leads to aperiodicity. The detailed balance holds for the case that $\boldsymbol{j} = \boldsymbol{i}$, but if $\boldsymbol{j} \neq \boldsymbol{i}$ it is not trivial. In case we accept $\{\boldsymbol{x}, \boldsymbol{\theta}\}^{(n+1)} = \boldsymbol{j}$ as the new state we have $\pi(\boldsymbol{j}|\boldsymbol{y})g(\boldsymbol{i}|\boldsymbol{j}) > \pi(\boldsymbol{i}|\boldsymbol{y})g(\boldsymbol{j}|\boldsymbol{i})$. This gives us $\alpha(\boldsymbol{j}|\boldsymbol{i}) = 1$ and $\alpha(\boldsymbol{i}|\boldsymbol{j}) = \frac{\pi_{\boldsymbol{i}} g(\boldsymbol{j}|\boldsymbol{i})}{\pi_{\boldsymbol{j}} g(\boldsymbol{i}|\boldsymbol{j})}$ and satisfies the detailed balance:

$$\cancel{\pi_{\boldsymbol{j}}} \frac{\pi_{\boldsymbol{i}}}{\cancel{\pi_{\boldsymbol{j}}}} g(\boldsymbol{j}|\boldsymbol{i}) = \pi_{\boldsymbol{i}} g(\boldsymbol{j}|\boldsymbol{i}) \quad .$$

If $\pi(\boldsymbol{j}|\boldsymbol{y})g(\boldsymbol{i}|\boldsymbol{j}) < \pi(\boldsymbol{i}|\boldsymbol{y})g(\boldsymbol{j}|\boldsymbol{i})$ then $\alpha(\boldsymbol{i}|\boldsymbol{j}) = 1$ and $\alpha(\boldsymbol{j}|\boldsymbol{i}) = \frac{\pi_{\boldsymbol{j}} g(\boldsymbol{i}|\boldsymbol{j})}{\pi_{\boldsymbol{i}} g(\boldsymbol{j}|\boldsymbol{i})}$, this satisfies the detailed balance as well.

In conclusion the Metropolis-Hastings algorithm samples from a unique distribution proportional to the posterior distribution.

C

Randomize then Optimize - RTO

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \tag{C.1}$$

$$\propto \exp\Big[(\boldsymbol{F}\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{F}\boldsymbol{x}-\boldsymbol{y}) + (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\Big] \tag{C.2}$$

$$= \exp\|\hat{\boldsymbol{F}}\boldsymbol{x}-\hat{\boldsymbol{y}}\|^2 \tag{C.3}$$

where

$$\hat{\boldsymbol{F}} = \begin{bmatrix}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{F}\\ \boldsymbol{Q}^{1/2}\end{bmatrix}, \quad \hat{\boldsymbol{y}} = \begin{bmatrix}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}\\ \boldsymbol{Q}^{1/2}\boldsymbol{\mu}\end{bmatrix} \tag{C.4}$$

One sample from the posterior can be computed by minimizing the following with respect to $\boldsymbol{x}$

$$\boldsymbol{x} = \arg\min_{\hat{\boldsymbol{x}}}\|\hat{\boldsymbol{F}}\hat{\boldsymbol{x}} - (\hat{\boldsymbol{y}}+\boldsymbol{\eta})\|^2, \quad \boldsymbol{\eta}\sim\mathcal{N}(\boldsymbol{0},\mathbf{I}) \tag{C.5}$$

We can solve this and rewrite to

$$\frac{\partial}{\partial\boldsymbol{x}}\Big[(\hat{\boldsymbol{F}}\boldsymbol{x} - (\hat{\boldsymbol{y}}+\boldsymbol{\eta})^T(\hat{\boldsymbol{F}}\boldsymbol{x} - (\hat{\boldsymbol{y}}+\boldsymbol{\eta})\Big] = 0 \tag{C.6}$$

$$\Leftrightarrow \boldsymbol{x}^T\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}} + \hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}\boldsymbol{x} - \hat{\boldsymbol{F}}^T(\hat{\boldsymbol{y}}+\boldsymbol{\eta}) - (\hat{\boldsymbol{y}}+\boldsymbol{\eta})^T\hat{\boldsymbol{F}}\boldsymbol{x} = 0 \tag{C.7}$$

We can argue through the symmetry of the inner product that and the symmetry of the precision matrix

$$\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}\boldsymbol{x} = \hat{\boldsymbol{F}}^T(\hat{\boldsymbol{y}}-\boldsymbol{\eta}) \tag{C.8}$$

$$\Leftrightarrow (\boldsymbol{F}^T\boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{F}+\boldsymbol{Q})\boldsymbol{x} = \boldsymbol{F}^T\boldsymbol{Q}_{\boldsymbol{y}}\boldsymbol{y} + \boldsymbol{Q}\boldsymbol{\mu} - \hat{\boldsymbol{F}}^T\boldsymbol{\eta} \tag{C.9}$$

If we substitute $-\hat{\boldsymbol{F}}^T\boldsymbol{\eta} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ we end up with

$$(\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F} + \boldsymbol{Q})\boldsymbol{x} = \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{v}_1 + \boldsymbol{v}_2 \tag{C.10}$$

where $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ are independent random variables. mayeb introduce... x^2 time nomral variubale

D

Inverting Matrices - QR factorization

E

Taylor expansion of $g(\lambda)$

We Taylor expand the function $g(\lambda)$ around $\lambda = \lambda' - \Delta\lambda$

$$g(\lambda) = \ln\det\underbrace{(\boldsymbol{F}^T\boldsymbol{F} + \lambda\boldsymbol{L})}_{\boldsymbol{B}} \tag{E.1}$$

$$\begin{aligned} g(\lambda') - g(\lambda) &= \ln\det(\boldsymbol{F}^T\boldsymbol{F} + \lambda'\boldsymbol{L}) - \ln\det(\boldsymbol{F}^T\boldsymbol{F} + \lambda\boldsymbol{L}) && \text{(E.2)}\\ &= \ln\det\left[\frac{(\boldsymbol{F}^T\boldsymbol{F} + (\lambda + \Delta\lambda)\boldsymbol{L})}{(\boldsymbol{F}^T\boldsymbol{F} + \lambda\boldsymbol{L})}\right] && \text{(E.3)}\\ &= \ln\det\left[1 + \frac{\Delta\lambda\boldsymbol{L}}{\boldsymbol{B}}\right] && \text{(E.4)}\\ &= \sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r!}\mathrm{tr}((\boldsymbol{B}^{-1}\boldsymbol{L})^r)(\Delta\lambda)^r && \text{(E.5)} \end{aligned}$$

, where we use the identity from [19] at page 29. So the derivatives of $g(\lambda)$ are:

$$\begin{aligned} g^{(r)}(\lambda) =& (-1)^{r+1}\,\mathrm{tr}\Big((\boldsymbol{B}^{-1}\boldsymbol{L})^r\Big) && \text{(E.6)}\\ \approx& (-1)^{r+1}\sum_{k=1}^{p}\boldsymbol{z_k}^T(\boldsymbol{B}^{-1}\boldsymbol{L})^r\boldsymbol{z_k} && \text{(E.7)} \end{aligned}$$

Here we use a Monte Carlo estimate and draw p vectors $\boldsymbol{z_k} \in \mathbb{R}^n$, where each vector element $z_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\{-1, 1\})$ and $i = 1, \ldots, n$.

F

Radiation transfer and absorption line shape

G

whispering gallery resonator

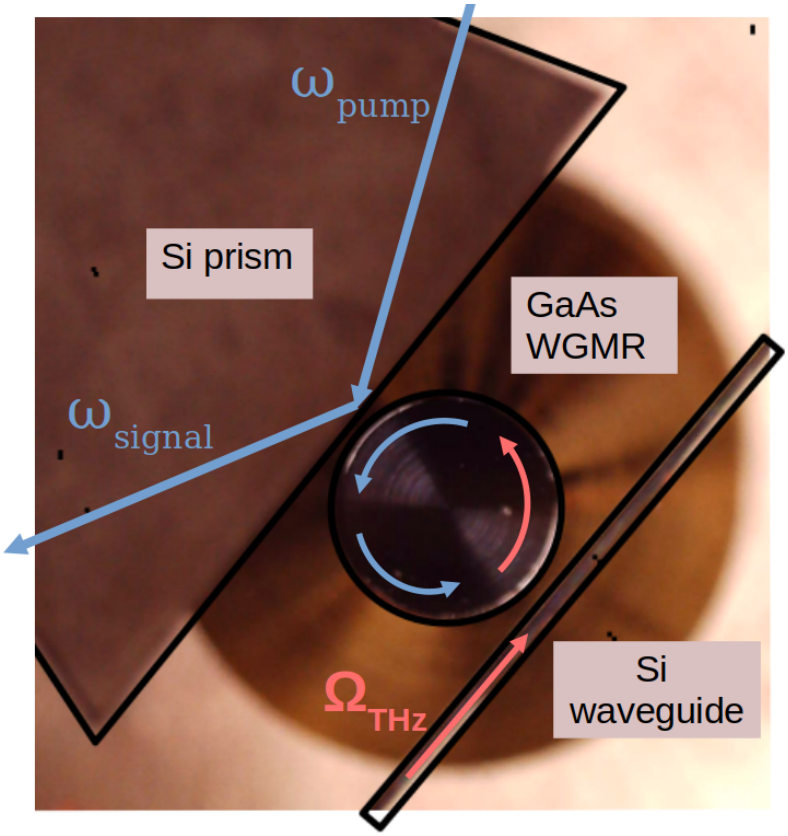


Figure G.1: whispering gallery resonator

References

[1] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[2] Dave Higdon. "A primer on space-time modeling from a Bayesian perspective". In: *Monographs on Statistics and Applied Probability* 107 (2006), p. 217.

[3] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Vol. 31. Springer Science & Business Media, 2013.

[4] Julian Besag. "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225.

[5] Colin Fox and Richard A Norton. "Fast sampling in a linear-Gaussian inverse problem". In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016), pp. 1191–1218.

[6] Daniel Simpson, Finn Lindgren, and Håvard Rue. "Think continuous: Markovian Gaussian models in spatial statistics". In: *Spatial Statistics* 1 (2012), pp. 16–29.

[7] Nicholas Metropolis and Stanislaw Ulam. "The monte carlo method". In: *Journal of the American statistical association* 44.247 (1949), pp. 335–341.

[8] John Michael Hammersley and David Christopher Handscomb. "General principles of the Monte Carlo method". In: *Monte Carlo Methods*. Springer, 1964, pp. 50–75.

[9] Paula A Whitlock and MH Kalos. *Monte Carlo Methods*. Wiley, 1986.

[10] AA Markov. "Extension of the law of large numbers to quantities, depending on each other (1906). Reprint." In: *Journal Électronique d'Histoire des Probabilités et de la Statistique [electronic only]* 2.1b (2006), Article–10.

[11] Colin Fox, Geoff K Nicholls, and Sze M Tan. "An Introduction to Inverse Problems". In: *Course notes for ELEC* 404 (2010).

[12] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: (1970).

[13] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[14] Johnathan M Bardsley. "MCMC-based image reconstruction with uncertainty quantification". In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1316–A1332.

[15] Johnathan M Bardsley et al. "Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 1136–1158.

[16] D S. Oliver, Nanqun He, and A C. Reynolds. *Conditioning permeability fields to pressure data*. 1996, cp–101.

[17] François Orieux, Olivier Féron, and J-F Giovannelli. "Sampling high-dimensional Gaussian distributions for general linear inverse problems". In: *IEEE Signal Processing Letters* 19.5 (2012), pp. 251–254.

[18] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[19] Israel Gohberg, Seymour Goldberg, and Nahum Krupnik. *Traces and determinants of linear operators*. Vol. 116. Birkhäuser, 2012.