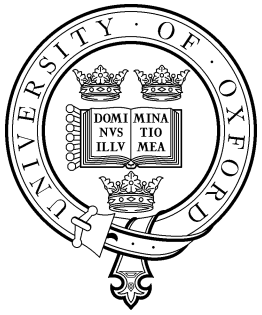
Making and Evaluating Predictions



Shilan Mistry

St Catherine's College

University of Oxford

Doctor of Philosophy

November 9, 2015

This thesis is dedicated to all my teachers. Above all, my mum, dad and brother whose enduring love and support made this possible.

Acknowledgements

I am grateful for the support of all my friends and family. I would like to thank my supervisor, Chris Farmer, for his invaluable input and stimulating conversation, and Angela Brueggemann, for providing a fixed point amongst the chaos. Most importantly I would like to thank my mum, dad and brother for it was their unconditional love and relentless optimism that sustained me throughout this endeavour.

Abstract

Given a sequence of measurements, a statistical model is a proposed solution to the inverse problem. Statistical models can be used to produce probabilistic predictions of future measurements. This thesis provides an exposition of the process of making and evaluating probabilistic predictions, with an overview of statistical models.

A review of statistical decision theory is given, and a statistical decision problem, arising in computational biology, is considered. The proposed solution to the decision problem is used to decide which genes, within a bacterial population, are labelled as *core* genes. The process of statistical modelling is understood as a decision problem, where the predictive density is the *optimal action*. The corresponding loss functions are known as *proper scoring rules*. Proper scoring rules can also be used for model parameter inference, and the large sample theory of the minimum score estimate is described. An application of proper scoring rules using the minimum score estimate is described and is shown to generalise classical model selection criteria.

A further application analyses measurement simulated from a noisy dynamical system. Four different models are applied to measurements with measurement noise simulated in the *perfect model scenario*, and in the *imperfect model scenario*. A numerical approximation to the exact Bayesian model is shown to outperform the other models in the perfect model scenario, but performance in the imperfect model scenario is dismal. This suggests that complicated models may not perform well if the measurements are from a system that deviates too far from the model. It emerges that models should always include an evaluation based on predictive performance.

Contents

1 Introduction **1**

1.1 Formulating the Inverse Problem 1

1.2 Constructing Statistical Models 2

1.3 Estimating the Core Genome 3

1.4 Proper Scoring Rules 3

1.5 Evaluating Predictive Performance 4

1.6 Noisy Dynamical Systems 5

1.7 Statement of Originality 5

1.8 Notation 6

2 Making Predictions **8**

2.1 Statistical Models 9

2.1.1 Model Construction 10

2.1.2 Model Predictions 11

2.2 Probabilistic Predictions 12

2.3 Model Inference 18

2.3.1 Maximum Likelihood 18

2.3.2 Maximum Posterior 19

2.3.3 Approximating the Posterior 23

3 Statistical Decisions Under Uncertainty **28**

3.1 Decision Theory . . . 29
3.1.1 Point Estimates . . . 31
3.2 Estimating the Core Genome . . . 35
3.2.1 Measurements . . . 36
3.2.2 The Classical Core Genome . . . 37
3.2.3 A Conservative Bayesian Decision Rule . . . 38
3.2.4 Results of the Decision Model . . . 41
3.3 Prediction as a Decision . . . 44
3.3.1 Probabilistic Predictions . . . 45

4 Proper Scoring Rules **47**
4.1 Designing a Loss Function . . . 48
4.1.1 Honesty . . . 50
4.1.2 Model Evaluation . . . 50
4.1.3 Parameter Estimation . . . 51
4.2 Some Proper Scoring Rules . . . 52
4.2.1 Logarithmic Scoring Rule . . . 52
4.2.2 Continuous Ranked Probability Score . . . 54
4.2.3 Constructing Scoring Rules . . . 55
4.3 Inference with Proper Scoring Rules . . . 57
4.3.1 Perfect Model Scenario . . . 59
4.3.2 Large Sample Theory . . . 60

5 Evaluating Predictions **66**
5.1 Bayesian Model Selection . . . 67
5.1.1 Multi-Model Probabilities . . . 68
5.1.2 Criticising Bayesian Model Selection . . . 70
5.2 Frequentist Risk Estimation . . . 73
5.2.1 Parametric Predictive Densities . . . 74

5.2.2 Criticising Frequentist Risk Estimation . . . 77
5.3 Conclusion . . . 79

6 Chaotic Dynamical Systems **81**
6.1 Simulating Measurements . . . 82
6.1.1 Perfect Model Scenario . . . 83
6.1.2 Imperfect Model Scenario . . . 84
6.2 Constructing Models . . . 84
6.2.1 Bayesian Model . . . 85
6.2.2 Sampling-Noise Model . . . 87
6.2.3 Random Analogue Prediction . . . 88
6.2.4 Climatology Model . . . 89
6.3 Evaluating Models . . . 89
6.3.1 Simulator Settings . . . 90
6.4 Two Chaotic Maps . . . 91
6.4.1 The Logistic Map . . . 91
6.4.2 Bayesian Solution . . . 92
6.4.3 The Gauss Map . . . 97
6.5 Assessing Predictive Performance . . . 100
6.5.1 Perfect Model Scenario . . . 100
6.5.2 Imperfect Model Scenario . . . 105
6.6 Conclusion . . . 106

7 Conclusion **112**
7.1 Making and Evaluating Predictions . . . 112
7.2 Predictive Modelling . . . 114

Chapter 1

Introduction

> The calculus of probabilities is an exact science, whose results are as reliable as those of arithmetic or algebra, as long as it confines itself to the numerical evaluation of probabilities. Borel [1962]

1.1 Formulating the Inverse Problem

Postulating a theory that predicts the results of numerical measurements corresponds to solving the forward model construction problem. Immediately, we notice that the solution to the forward model construction problem is ill-posed. Using measurements to infer the parameters that define a model corresponds to solving the inverse problem [Tarantola, 2006]. However, when we look for a unique solution to the inverse problem, we find that this problem is also ill-posed. That is, different models can be consistent with the same set of measurements.

Early formalisation of the inverse problem can be found in the works of Laplace, Legendre, and Gauss, where the main focus was on parameter estimation rather than model testing. The problem under consideration was the use of measurements to make inferences about

properties of the Earth and astronomical systems. This was in turn used to make inferences concerning the orbits of planets. Laplace selected model parameters by choosing values for the parameters that minimised the absolute difference between the model and the measurements [Klein, 1997]. In contrast Legendre [1805] and Gauss [1821] minimise the square of the difference. The maximum likelihood method, selecting values of the parameters which maximised the assessed probability, was employed by Bernoulli [1777] and Gauss [1821]. An in depth exposition of the history of time series analysis between 1662 and 1938 can be found in Klein [1997].

Further analysis is required to determine what model is *optimal* for the collection of measurements under study. Popper [1959] argues that measurements should be used to falsify models, indicating that this can be framed as a problem of probability. Hence we adopt a probabilistic approach by constructing a *statistical model* to describe our *uncertainty* of future measurements.

1.2 Constructing Statistical Models

The main reason for studying these models is not to elucidate the fundamental workings of a phenomenon, but to make predictions [Hoyningen-Huene, 2013]. If prediction is the main objective, a programme of model falsification is unnecessarily harsh. One might decide that if the probability of observing the sequence of empirical measurements, conditional on the statistical model, is zero, then the model should be falsified. In the numerical study of any phenomenon there is little use in employing criteria such as falsification, and, often, there are a number of models that could be considered, such that the probability of observing the measurements is not zero.

Chapter 2 reviews the fundamentals of statistical model construction. Models are constructed by relating observed variables with model variables, and by constructing conditional probabilistic relationships between them. Once a model is constructed, the measurements are used

to *fit* the model, yielding a plausible solution to the inverse problem. A statistical model yields a predictive probability distribution of the measurements that can be used to forecast a future observation.

Models are constructed by making decisions as to which probabilistic structures are used to describe our uncertainty. These decisions reflect the interests and preferences of scientists, and can lead to a plurality of statistical models whose predictions are compatible with the observed measurements.

1.3 Estimating the Core Genome

Chapter 3 reviews the fundamentals of decision theory. The classical use of statistical decision theory is for the evaluation of point estimates for parameters. The decision that we wish to undertake may depend on some parameter that is not measured directly, but has to be estimated from the available measurements. As an example of scientific decision making, a problem in computational biology is considered in Chapter 3. A biological decision is to be made, conditional on the estimate of an unobservable probability parameter. The decision problem is structured to reflect the interests and concerns of the scientists, and suggests avenues for further scientific investigation. The chapter concludes by interpreting prediction as a decision problem.

1.4 Proper Scoring Rules

If the goal is to predict, as accurately as possible, the future value of the measurement, then an *action* is to issue some point prediction of a measurement. Alternatively, our action might be to select the predictive probability density. Loss functions for which the optimal action is the predictive probability density are known as *proper scoring rules*. In Chapter 4, derivations of the two most popular proper scoring rules for continuous random variables,

the logarithmic score, and the continuous ranked probability score, are given. A general method, using convex functions, for constructing proper scoring rules is also provided.

Proper scoring rules can also be used for model inference. The value of the parameter for which the score is minimised is known as the minimum score estimate. If the measurements are modelled as independent and identical realisations of our model, and the model is *perfect*, the minimum score estimate converges, in probability, to the true value as the number of measurements increases to infinity. In the imperfect model scenario the minimum score estimate converges to a pseudo-true value as the number of measurements increases to infinity under particular model assumptions. For a large enough sample size, the limiting Normal behaviour of the estimator is derived in Chapter 4. The limiting behaviour of the minimum score estimate is used in the Chapter to estimate the predictive risk of predicting a future measurement, conditional on the choice of model.

1.5 Evaluating Predictive Performance

The Bayesian approach to model selection is to place a prior over a collection of models, and calculate the posterior probabilities. This is shown to be equivalent to selecting the model with smallest observed empirical score.

The Frequentist approach to model selection is characterised by treating the measurements as a random sample from some unknown true probability distribution, so that a statistical model is understood as an approximation of this true probability distribution. One then estimates, from the measurements, the model risk, and utilises this quantity for model selection.

In Chapter 5, we estimate the Frequentist model risk for arbitrary proper scoring rules. The natural estimator, the fitted empirical score, is shown to be biased, and the correction term is derived. The estimator generalises the classical model selection criteria, and these links are established within the chapter.

1.6 Noisy Dynamical Systems

In Chapter 6, we consider sequences of noisy measurements generated using chaotic maps. We consider the problem of forecasting future measurements, and consider a number of forecasting models. In particular, we calculate a numerical approximation to the exact Bayesian solution. We can show that the Bayesian solution outperforms all other models when the Bayesian solution faithfully reflects the measurement simulator.

We proceed to modify the sequence of measurements, by introducing a corrupting influence, and analyse the performance of the models. As the corrupting influence grows, the predictive performance of the exact Bayesian solution is increasingly dismal. This suggests that, while models that capture the exact dynamics of the underlying phenomenon may perform well in situations that agree with the model assumptions, slight deviations away from the model assumptions can result in a significant reduction of predictive performance.

1.7 Statement of Originality

After the introduction a review of standard statistical concepts in given in Chapter 2 and at the beginning of Chapter 3. Later in Chapter 3, a decision problem arising in computational biology is considered and a novel solution is proposed. This work has been published in van Tonder et al. [2014] of which the candidate is the second author solely responsible for all mathematical and statistical aspects. A novel interpretation of proper scoring rules as the loss functions associated to the predictive decision problem is given in Chapter 4. This allows for all statistical models to be compared in a principled Bayesian way. Chapter 5 is a novel application of proper scoring rules in which we demonstrate equivalence to classical model selection criteria under particular model assumptions. Two noisy dynamical systems are used to illustrate how model performance deteriorates as we deviate from the perfect model scenario into the imperfect model scenario in Chapter 6. The analysis constitutes a novel use of noisy dynamical systems in illustrating problems in data assimilation. Further-

more, a novel derivation of the exact Bayesian model is provided as well as the numerical implementation.

1.8 Notation

Consider a collection of real-valued measurements $x_{1:t} = \{x_1, , \ldots, x_t\}$, and let $x_{\neg i} = x_{1:t} \setminus \{x_i\}$, where $\setminus$ denotes the set difference. At each time i, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ is a column vector so that $x_i^T = (x_{i,1}, \ldots, x_{i,d})$, where T is the transpose operation. If A is a matrix, then A^{-1} is the inverse (if it exists), and $A^{-T} \equiv (A^{-1})^T$ is the transpose of the inverse. For an arbitrary variable $x \in \mathbb{R}^d$, let

$$\nabla_x = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d} \right) \tag{1.1}$$

denote the partial derivative with respect to the variable x so that for some vector valued function $f(x)$, $\nabla_x f(x^*)$ is the Jacobian matrix of the vector valued function evaluated at $x^* \in \mathcal{X}$, where the (i, j)-th entry of the matrix is

$$\left. \frac{\partial f_i(x)}{\partial x_j} \right|_{x=x^*}. \tag{1.2}$$

and the Hessian matrix of the function f evaluated at x^* is denoted as $\nabla_x^2 f(x^*)$, where the (i, j)-th entry of the matrix is

$$\left. \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right|_{x=x^*}. \tag{1.3}$$

Then a second-order Taylor expansion around x^* is written as

$$f(x) = f(x^*) + \nabla_x f(x^*)(x - x^*) + \frac{1}{2}(x - x^*)^T \nabla_x^2 f(x^*)(x - x^*) + \ldots \ . \tag{1.4}$$

Denote the set probability function as $\mathbb{P}\{\cdot\}$. Let $\pi(\cdot)$ denote a generic probability density function and its capital $\Pi(\cdot)$ be the associated cumulative distribution function. Bold symbols, such as $\boldsymbol{x}$, denote real valued random variables, while non-bold symbols, such as x, denote a real valued variable. Denote the value of the observed measurement as $\mathring{x}$ and

subsequently we will refer to $\mathring{x}$ as the *actualising* value. For a scalar random variable we have

$$\mathbb{P}\{\boldsymbol{x} \leq x\} = \int_{-\infty}^{x} \pi(x')dx' = \Pi(x). \tag{1.5}$$

Since we utilise $\pi(\cdot)$ to denote probability densities, we will, when necessary, utilise $\boldsymbol{\pi}$ to denote the transcendental number that is the area of the unit circle. Furthermore, understand $\boldsymbol{x} \sim \pi$ to mean that the random variable $\boldsymbol{x}$ is distributed according to the probability density π. Lastly, let $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$, and $\mathbb{C}\text{ov}[\cdot]$ denote the expectation, variance and covariance respectively.

Chapter 2

Making Predictions

In this chapter, we review a number of different probabilistic models, including Bayesian models. A summary of some classical inference techniques, such as maximum likelihood and maximum posterior estimates, is provided, and a derivation of the Normal approximation to the Bayesian posterior is also provided.

The main purpose for constructing a statistical model is to facilitate the derivation of a probability distribution, conditional on measurements that quantifies our uncertainty [Lindley, 2013]. We hope to provide a descriptive analysis of some past phenomenon, or perhaps provide some predictions about a similar phenomenon in the future. Utilising probabilistic modelling could seem to be overly complicated, as a phenomenon could be entirely deterministic. However, there are examples in which the behaviour of a deterministic model is better understood probabilistically.

Quantitative scientific analysis of some natural phenomena is a process [Cox, 2006]. Once it is decided what phenomenon is under consideration, the process is initialised by collection of relevant numerical data. These numerical data constitute a collection of measurements. Next, we design a probabilistic model, or models, of the measurements.

A probabilistic model is chosen through careful consideration of the measurements employ-

ing available knowledge of the phenomenon under study. Box [1979] famously wrote "All models are wrong but some models are more useful than others". A model is only capable of capturing particular aspects of the data, and is subject to the modeller's preferences and abilities. Even highly sophisticated models, capturing stylised facts such as volatility clustering or long memory, are confronted with issues when implemented in practice. Any model is, implicitly, a simplification or approximation of reality.

Measurements are used to infer the model parameters. When this is complete, a prediction of some future unknown measurement can be made. The goal is to make a sufficiently credible guess, and its credibility should be evaluated by comparing the model predictions with measurements.

2.1 Statistical Models

The variables of a statistical model can be classified into two types, measurements and parameters. Let $x_{1:t} = \{x_1, \ldots, x_t\}$ denote the variables that are observed, known as the measurements. Model parameters can be split into two types, latent model parameters and parameters of interest. Let θ denote the latent model parameters while y denotes the parameters of interest.

To clarify the difference between parameters of interest and latent model parameters, consider the following example. One might use a noisy dynamical system to model a sequence of measurements. In particular we might be interested in making an inference of the initial position of the system. Then one would classify the initial position as the parameter of interest and all other parameters would be the latent model parameters. Alternatively, if one is interested in making a forecast of a future measurement, the parameter of interest is the future measurement and all other parameters are latent.

2.1.1 Model Construction

A model is constructed by specifying conditional probabilistic relationships between the unobserved variables in our model and the observed measurements. This chapter reviews the construction of statistical models and methods of classical inference.

Definition 1. *Statistical Model*

Let $x_{1:t}$ be a collection of measurements. Let θ be the latent model parameters. Let y be the model parameters of interest. A statistical model M specifies the conditional joint probability density

$$\pi(y, \theta, x_{1:t}|M). \tag{2.1}$$

Different statistical models can be *constructed* by specifying how the joint density factors into conditional probability densities. The probabilistic relationships between the variables are understood as the *structures* used to construct the model. That is, statistical models are constructed from component probabilistic structures.

Definition 2. *Structure*

A structure is a collection of one or more probabilistic relationships between model variables.

An exhaustive description of all types of statistical models is not feasible, and so the discussion is limited to the statistical models referred to in this work. For an encyclopaedic account of probabilistic models for machine learning, we refer to Murphy [2012].

Once a model is selected, we use the actualising values of the measurements, $\mathring{x}_{1:t} = \{\mathring{x}_1, \ldots, \mathring{x}_t\}$ to infer the model parameters. The probability density of model parameters, conditionally dependent on the measurements, is obtained by computing the density using the rules of conditional probability as

$$\pi(y, \theta|\mathring{x}_{1:t}, M) = \frac{\pi(y, \theta, \mathring{x}_{1:t}|M)}{\pi(\mathring{x}_{1:t}|M)}. \tag{2.2}$$

Here, the denominator, $\pi(\mathring{x}_{1:t}|M)$, is a constant depending on the measurements. The probability density describing the parameter of interest is calculated by integrating out the latent

parameters,

$$\pi(y|\mathring{x}_{1:t}, M) = \int \pi(y, \theta|\mathring{x}_{1:t}, M)d\theta. \tag{2.3}$$

If further *action* is to be taken based on the parameter of interest, it is this marginal density which is required to determine the *action* [Lindley, 1965]. To avoid technical difficulties we restrict consideration to models for which the involved probability densities are well defined.

2.1.2 Model Predictions

Parameters of interest can be split into two types, parameters that are never observed, and parameters that will be observed. For example, if we are interested in predicting the next k measurements, the marginal density $\pi(x_{t+1:t+k}|\mathring{x}_{1:t}, M)$ is the k-step ahead predictive density. Given a sequence of measurements, the predictive density decomposes, by the rules of conditional probability, as

$$\pi(x_{1:t+k}|M) = \pi(x_{t+1}, \ldots, x_{t+k}|x_{1:t}, M)\pi(x_{1:t}|M). \tag{2.4}$$

Once the verifying observations $\mathring{x}_{t+1:k}$ have been made, the predictive likelihood can be written as the ratio of marginal likelihoods

$$\pi(\mathring{x}_{t+1:t+k}|\mathring{x}_{1:t}, M) = \frac{\pi(\mathring{x}_{1:t+k}|M)}{\pi(\mathring{x}_{1:t}|M)}. \tag{2.5}$$

The predictive likelihood is the factor by which the marginal likelihood is updated as more data are collected. This property holds independent of the times at which the model is updated and forecasts made, and it suffices to consider, without loss of generality, the step-ahead decomposition of the predictive likelihood.

$$\pi(\mathring{x}_{1:t}|M) = \prod_{i=1}^{t} \pi(\mathring{x}_i|\mathring{x}_{1:i-1}, M), \tag{2.6}$$

where we adopt the notation $x_{1:0} = \emptyset$. This relationship between the predictive likelihood and the marginal likelihood allows us to place the predictive likelihood as the central element in the data assimilation, forecasting, and verification processes.

2.2 Probabilistic Predictions

For some models, it is not necessary to introduce model parameters in order to make predictions. Zero-parameter models can be used to compare other forecasts against. An example of a nonparametric model is the persistence model which issues, as a prediction of the next measurement, the value of the previous measurement.

Model 1. *Persistence*

The persistence model places all probability on the value of the previous measurement. Let $\delta(\cdot)$ *be the Dirac delta function [Dirac, 1947] so that, for all* i*,*

$$\boldsymbol{x}_i \sim \delta(x_i - \mathring{x}_{i-1}). \tag{2.7}$$

While the model is non-probabilistic in nature, the use of a delta function allows point prediction models to be understood as probabilistic in that we place all probability mass on a particular value. Another example of a probabilistic model is the sample climatology model. For this model, we model the sequence of measurements as a collection of independent identically distributed (*iid*) random samples from some common cumulative probability distribution.

Model 2. *Climatology*

$$\boldsymbol{x}_i \overset{iid}{\sim} c(\cdot) \text{ for all } i. \tag{2.8}$$

where, at time i, the predictive density is an estimate of $c(\cdot)$

$$\hat{c}(x_i) = \pi(x_i|\mathring{x}_{1:i-1}) = \frac{1}{i-1}\sum_{k=1}^{i-1}\delta(x_i - \mathring{x}_k) \tag{2.9}$$

The cumulative probability distribution corresponding to $\pi(x_i|\mathring{x}_{1:i-1})$ can be estimated using the empirical distribution function. Let $\mathbb{H}(\cdot)$ denote the Heaviside function where $\mathbb{H}(x) = 0$ if $x < 0$ or $\mathbb{H}(x) = 1$ if $x \geq 0$. Then, the climatological forecast for the future measurements has the cumulative probability distribution function

$$\Pi(x_i|\mathring{x}_{1:i-1}, M) = \frac{1}{i-1}\sum_{k=1}^{i-1}\mathbb{H}(\mathring{x}_k - x_i). \tag{2.10}$$

By the introduction of additional variables and modification of the statistical relationships between the variables, different models can be defined. Perhaps the simplest parametric models are those that assume the measurements form a conditionally independent identically distributed random sample from a known probability density. These models are constructed by specifying the likelihood.

Structure 1. *Conditionally Independent Measurements*

Measurements are modelled as conditionally independent of each other if the form of the likelihood function is

$$\pi(x_{1:t}|\theta) = \prod_{i=1}^{t} \pi(x_i|\theta). \tag{2.11}$$

When the conditional densities for all the measurements are identical, then the measurements are said to be conditionally independent and identically distributed.

For instance, we might make measurements $x_i > 0$, and propose a Weibull model. This amounts to fixing the likelihood function as the Weibull density.

Model 3. *Weibull Model*

Let $x_i, \lambda, \alpha > 0$

$$\pi(x_{1:t}|\theta, M) = \prod_{i=1}^{t} Wei(x_i|\lambda, \alpha) = \prod_{i=1}^{t} \frac{\alpha}{\lambda} \frac{x_i}{\lambda}^{\alpha-1} \exp\left(-\frac{x_i}{\lambda}^{\alpha}\right) \tag{2.12}$$

The parameters of the model $\theta = (\lambda, \alpha)$ are unknown, and must be estimated using the measurements. Treating the parameters as random variables characterises the Bayesian approach to statistical inference. In particular, we adopt a prior probability density for each parameter, and compute the posterior probability after making the measurements.

Model 4. *Bayesian Model*

A Bayesian *model specifies a prior denoted* $\pi(\theta|M)$*, incorporates the measurements through the likelihood* $\pi(x_{1:t}|\theta, M)$*, and produces, as an output, the posterior density* $\pi(\theta|\mathring{x}_{1:t}, M)$*. This is achieved using Bayes' theorem,*

$$\pi(\theta|\mathring{x}_{1:t}, M) = \frac{\pi(\mathring{x}_{1:t}|\theta, M)\pi(\theta|M)}{\pi(\mathring{x}_{1:t}|M)} \propto \pi(\mathring{x}_{1:t}|\theta, M)\pi(\theta|M), \tag{2.13}$$

where $\propto$ is the symbol for proportionally. The constant of proportionality in (2.13) is

$$\pi(\mathring{x}_{1:t}|M) = \int \pi(\mathring{x}_{1:t}|\theta, M)\pi(\theta|M)d\theta. \tag{2.14}$$

The predictive density for a further measurement $\pi(y|\mathring{x}_{1:t}, M)$ is obtained by computing the integral

$$\pi(y|\mathring{x}_{1:t}, M) = \int \pi(y|\theta)\pi(\theta|\mathring{x}_{1:t}, M)d\theta. \tag{2.15}$$

For example, if we observe a sequence of t measurements indicative of a binary variable, then a Bayesian model could be the beta-binomial model.

Model 5. *Beta-Binomial Model*

Let $x_i \in \{0, 1\}$, and $n = \sum_{i=1}^{t} x_i$ and $\theta \in [0, 1]$. Let $\alpha > 0$ and $\beta > 0$ be the parameters of the Beta distribution,

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{2.16}$$

$$\pi(x_{1:t}|\theta) \propto \theta^{n}(1-\theta)^{t-n} \tag{2.17}$$

where the constant of proportionality depends on the beta function.

An application of Bayes' theorem yields a beta distribution as the posterior with parameters $\alpha^* = \alpha + n$ and $\beta^* = \beta + t - n$.

$$\pi(\theta|\mathring{x}_{1:t}) \propto \theta^{\alpha+n-1}(1-\theta)^{\beta+t-n-1}. \tag{2.18}$$

So the predictive probability density for the future measurement y is

$$\pi(y|\mathring{x}_{1:t}) = \int_0^1 \pi(y|\theta)\pi(\theta|\mathring{x}_{1:t})d\theta = \frac{B_1(y_j + \alpha^*, 1 - y_j + \beta^*)}{B_1(\alpha^*, \beta^*)}, \tag{2.19}$$

where B_1 is the beta function.

Often, a linear model of the measuring device is employed. Consider a noisy linear regression model in which the measurements are modelled as a linear combination of the regression variables corrupted with noise.

Structure 2. *Linear Measuring Device*

Employing Structure 1, let $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{\theta} \in \mathbb{R}^m$, $A \in \mathbb{R}^{d\times m}$ *and set*

$$\boldsymbol{x}_i \sim \pi(x_i|\theta) = \rho(x_i - A\theta), \tag{2.20}$$

where $\rho(\cdot)$ *is a probability density selected to model the noise in the measuring device.*

That is, for each i, $\pi(x_i|\theta) = \rho(x_i - A\theta)$, and so the joint likelihood (2.11) of the measurements is given by the product of these densities indexed by i. It follows that the i-th predictive density is given by

$$\pi(x_i|\mathring{x}_{1:i-1}) \propto \int_\Theta \rho(x_i - A\theta) \prod_{k=1}^{i-1} \rho(\mathring{x}_k - A\theta)\pi(\theta)d\theta, \tag{2.21}$$

where we have introduced $\pi(\theta)$ as the prior on the parameters. A popular choice for $\rho(\cdot)$ is the Normal density.

Probabilistic relationships between the model variables define which types of behaviour a model can exhibit. A popular model structure is to model these parameters as Normal.

Structure 3. *Model parameters exhibit Normal behaviour if the joint probability density is a multivariate Normal density*

$$\pi(\theta) = \boldsymbol{N}(\theta; \mu, \Gamma) = (2\boldsymbol{\pi})^{-\frac{\dim(\theta)}{2}} \det(\Gamma)^{-\frac{\dim(\theta)}{2}} \exp\left(-\frac{1}{2}(\theta-\mu)^T\Gamma^{-1}(\theta-\mu)\right), \tag{2.22}$$

where μ *is the mean and* Γ *is a symmetric positive definite matrix known as the covariance.*

Structure 4. *State Space*

A state space structure specifies, for each i*, the conditional relationships* $\pi(x_i|\theta_i)$ *between the model parameters and the measurements.*

Model 6. *Build a model with conditionally independent measurements (Structure 1) and a state space (Structure 4). Then, the likelihood function is*

$$\pi(x_{1:t}|\theta_{1:t}) = \prod_{i=1}^{t} \pi(x_i|\theta_i)\pi(\theta_i|\theta_{\neg i}). \tag{2.23}$$

Aside from prediction, we are often interested in *filtering* problems, which involves computing the marginal densities $\pi(\theta_i|\mathring{x}_{1:i})$, or *smoothing* problems, which involve computing the

marginal densities $\pi(\theta_i|\mathring{x}_{1:t})$, for $i \leq t$ [Särkkä, 2013]. Models with this type of structure are useful in that the modelling of the latent parameters is contained in the specification of $\pi(\theta_i|\theta_{\neg i})$. For example, state spaces are often equipped with a Markov structure.

Structure 5. *Markov*

A state space structure is Markov if

$$\pi(\theta_i|\theta_{\neg i}) = \pi(\theta_i|\theta_{i-1}) \text{ for all } i \tag{2.24}$$

where we understand $\pi(\theta_0|\theta_{0-1}) := \pi(\theta_0)$.

The predictive density for a statistical model with Markov behaviour can be computed sequentially. The model assumption implies that calculation of the posterior at time $t+1$ depends only on the posterior at time t. A Markov model of the dynamics is the linear model.

Structure 6. *Linear Dynamical System*

Consider a model with a Markov structure. Let $\boldsymbol{\theta}_i \in \mathbb{R}^m$, $B \in \mathbb{R}^{m\times m}$ *and* $\rho(\cdot)$ *be a probability density modelling the dynamical noise*

$$\boldsymbol{\theta}_0 \sim \pi(\theta_0), \tag{2.25}$$

$$\boldsymbol{\theta}_i \sim \pi(\theta_i|\theta_{i-1}) = \rho(\theta_i - B\theta_{i-1}). \tag{2.26}$$

Alternatively, one might employ a deterministic model of the trajectory of the state space parameters. The evolution of the state can be written using $\delta(\cdot)$, the Dirac delta function.

Structure 7. *Deterministic Dynamics*

Let $f : \mathbb{R}^m \to \mathbb{R}^m$ *and consider a model with a Markov structure (Structure 5). A deterministic model of the state space is of the form*

$$\pi(\theta_i|\theta_{i-1}) = \delta(f(\theta_{i-1}) - \theta_i). \tag{2.27}$$

Combining a conditionally independent structure with a Markov behaviour builds a Hidden Markov model.

Model 7. *Hidden Markov Model.*

Build a model with Structures 1, 4, 5 with parts

$$\boldsymbol{\theta}_0 \sim \pi(\theta_0), \tag{2.28}$$

$$\boldsymbol{\theta}_i \sim \pi(\theta_i|\theta_{i-1}), \tag{2.29}$$

$$\boldsymbol{x}_i \sim \pi(x_i|\theta_i). \tag{2.30}$$

For the Hidden Markov model, the i-th predictive density is given by

$$\pi(x_i|\mathring{x}_{1:i-1}) = \int\int \pi(x_i|\theta_i)\pi(\theta_i|\theta_{i-1})\pi(\theta_{i-1}|\mathring{x}_{1:i-1})d\theta_{i-1}d\theta_i, \tag{2.31}$$

where $\pi(\theta_{i-1}|\mathring{x}_{1:i-1})$ is the posterior density of the parameter θ_{i-1}, after observing the measurements $\mathring{x}_{1:i-1}$, obtained using Bayes' theorem. If a more complicated model structure is desired, we can introduce additional model parameters and specify probabilistic relationships between all the parameters.

Model 8. *Hierarchical State Space Model*

Build a model with Structures 1 and 4. Augment it with additional parameters ψ so that its parts are

$$\boldsymbol{\psi} \sim \pi(\psi), \tag{2.32}$$

$$\boldsymbol{\theta}_0 \sim \pi(\theta_0|\psi), \tag{2.33}$$

$$\boldsymbol{\theta}_i \sim \pi(\theta_i|\theta_{\neg i}, \psi), \tag{2.34}$$

$$\boldsymbol{x}_i \sim \pi(x_i|\theta_i). \tag{2.35}$$

So the i-th predictive density is found by computing the marginal density

$$\pi(x_i|\mathring{x}_{1:i-1}) = \int\int \pi(x_i, \theta_i, \psi|\mathring{x}_{1:i-1})d\theta_i d\psi \tag{2.36}$$

$$= \int\int\int \pi(x_i|\theta_i)\pi(\theta_i|\theta_{\neg i}, \psi)\pi(\theta_{\neg i}, \psi|\mathring{x}_{1:i-1})d\theta_i d\psi d\theta_{\neg i}, \tag{2.37}$$

where $\pi(\theta_{\neg i}, \psi|\mathring{x}_{1:i-1})$ is the posterior. A popular, flexible model is a Hierarchical State Space model combined with a Markov model of the state dynamics (Structure 5), assuming conditionally independent measurements (Structure 1).

2.3 Model Inference

For high-dimensional models, an evaluation of (2.15), the marginal density, could prove to be difficult, or even impossible. The uncertainty in the parameters of interest is computed as

$$\pi(y|\mathring{x}_{1:t}, M) = \int \pi(y|\theta)\pi(\theta|\mathring{x}_{1:t}, M)d\theta, \tag{2.38}$$

where θ denotes the latent model parameters. Instead, one might estimate the parameter $\hat{\theta}$, conditional on the measurements, and make a prediction using the point estimate for the parameter so that the predictive density is replaced, using the parameter estimate, as

$$\pi(y|\mathring{x}_{1:t}, M) \cong \pi(y|\hat{\theta}). \tag{2.39}$$

where $\cong$ is understood as using $\pi(y|\hat{\theta})$ in place of $\pi(y|\mathring{x}_{1:t}, M)$. Forecasts made in this way are called *plug-in* forecasts, as opposed to the full probabilistic forecast obtained through direct marginalisation.

2.3.1 Maximum Likelihood

The most well-known parameter estimate is the maximum likelihood estimate. The estimate is found by maximising the likelihood function, or, equivalently, the log likelihood function

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \pi(\mathring{x}_{1:t}|\theta) = \underset{\theta}{\operatorname{argmax}} \log \pi(\mathring{x}_{1:t}|\theta) = \underset{\theta}{\operatorname{argmin}} \left(-\log \pi(\mathring{x}_{1:t}|\theta)\right). \tag{2.40}$$

For example, consider a sequence of measurements analysed using Structure 2. Then the log likelihood function is

$$\log \pi(\theta|\mathring{x}_{1:t}) = \sum_{i=1}^{t} \log \rho(A_i\theta - \mathring{x}_i). \tag{2.41}$$

Let $v = A_i\theta - \mathring{x}_i$, so if $\rho(v) = \boldsymbol{N}(v|0, \sigma^2)$, the maximum likelihood estimate is given by

$$-\log \pi(\mathring{x}_{1:t}|\theta) = \frac{t}{2}\log(2\boldsymbol{\pi}\sigma^2) + \frac{1}{2\sigma^2}||A\theta - \mathring{x}_{1:t}||_2^2 \text{ where } A = \begin{bmatrix} A_1^T \\ \vdots \\ A_t^T \end{bmatrix} \tag{2.42}$$

$$\implies \hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmin}} ||A\theta - \mathring{x}_{1:t}||_2^2. \tag{2.43}$$

Alternatively, one might model the noise using a Laplacian density, for $a > 0$

$$\rho(v) = \frac{1}{2a}\exp\left(-\frac{|v|}{a}\right) \implies \hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmin}} ||A\theta - \mathring{x}_{1:t}||_1. \tag{2.44}$$

Or, perhaps, employ a uniform model of the noise so that, for $v \in [-a, a]$

$$\rho(v) = \frac{1}{2a} \implies \hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmin}} ||A\theta - \mathring{x}_{1:t}||_\infty. \tag{2.45}$$

In many cases, one can write the density of the noise as $\rho(v) \propto \exp(-\zeta(v))$, for some function ζ, so that finding the maximum likelihood estimate is equivalent to the minimisation program

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} \zeta(x_i - A^T\theta). \tag{2.46}$$

The Bayesian maximum posterior estimate is closely related to the maximum likelihood estimate, differing with the addition of a prior density for the model parameters.

2.3.2 Maximum Posterior

The maximum posterior estimate is found by solving the following optimisation problem:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, \pi(\theta|\mathring{x}_{1:t}) = \underset{\theta}{\operatorname{argmax}}\, \pi(\mathring{x}_{1:t}|\theta)\pi(\theta) = \underset{\theta}{\operatorname{argmax}} \left(\log(\pi(\mathring{x}_{1:t}|\theta)\pi(\theta))\right). \tag{2.47}$$

It follows that if a flat prior, $\pi(\theta) \propto 1$, is used to model the parameters, the optimisation programs for the maximum likelihood (2.40) and the maximum posterior (2.47) estimates are equivalent. Since $\pi(\theta|x)$ is a probability density, it is a non-negative integrable function.

So, we can write the posterior as $\pi(\theta|\mathring{x}_{1:t}) \propto \exp(-\Phi(\theta))$, where Φ is a non-negative function. The maximum posterior estimate is found by minimising Φ where Φ is sometimes called the *misfit* function. When a model with conditionally independent and identically distributed measurements is employed, the misfit function can be written as a sum of the individual log likelihoods for each observed measurement.

In the case of the noisy linear model, each minimisation program is feasible, and has a global stationary point, because the noise $\rho(v)$ is log-concave. If the misfit function is convex, and the prior density is log-concave, the optimisation problem remains convex, and a global stationary point can be found. For more complicated models, the misfit function is not necessarily convex, and a global optimisation program can be difficult to carry out.

In the previous example, the parameters are the same for each measurement. In the analysis of measurements made from a dynamical system, a more complex model is often required. If we construct a state space model (Structure 4) and model the measuring device using a Normal density, $\rho(\cdot) = \boldsymbol{N}(\cdot|0, \Gamma)$, the log likelihood function is

$$\log \pi(\mathring{x}_{1:t}|\theta) = \log \prod_{i=1}^{t} (2\boldsymbol{\pi})^{-\frac{k}{2}} \det(\Gamma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(A_i\theta_i - \mathring{x}_i)^T\Gamma^{-1}(A_i\theta_i - \mathring{x}_i)\right) \tag{2.48}$$

$$= -\frac{tk}{2}\log(2\boldsymbol{\pi}) - \frac{t}{2}\log\det(\Gamma) - \frac{t}{2}\mathbf{tr}(\Gamma^{-1}V), \tag{2.49}$$

where V is the sample covariance.

$$V = \frac{1}{t}\sum_{i=1}^{t}(A_i\theta_i - \mathring{x}_i)(A_i\theta_i - \mathring{x}_i)^T. \tag{2.50}$$

Let the prior of the state space be $\pi(\theta_{1:t}) = \boldsymbol{N}(\theta_{1:t}|\theta_{\text{prior}}, \Gamma_{\text{prior}})$, then the maximum posterior estimate is found by solving

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmin}} \left(\frac{1}{2}\sum_{i=1}^{t}||A_i\theta_i - \mathring{x}_i||^2_{R^{-1}} + \frac{1}{2}||\theta_{1:t} - \theta_{\text{prior}}||^2_{\Gamma^{-1}_{\text{prior}}}\right), \tag{2.51}$$

where we adopt the notation $||x||_A = x^TAx$.

In the case that the state parameters are modelled as a linearly evolving dynamical system, we consider a Hidden Markov Model (Model 7) in which the measurement device is linear

(Structure 2) and the Markov behaviour of the dynamical model is also linear (Structure 6). Assuming that all model probability densities are Normal yields the Normal Linear Hidden Markov model.

Model 9. *Normal Linear Hidden Markov Model*

Construct a Hidden Markov Model (Model 7). Let $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{\theta}_i \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times m}$, $\Gamma \in \mathbb{S}_+^m$, $\theta_{prior} \in \mathbb{R}^m$, $\Gamma_{prior} \in \mathbb{S}_+^m$, $B \in \mathbb{R}^{d\times m}$, $R \in \mathbb{S}_+^d$, *where* $\mathbb{S}_+^n$ *denotes the set of symmetric positive definite matrices of dimension* n. *Set*

$$\boldsymbol{\theta}_0 \sim \boldsymbol{N}(\theta_0|\theta_{prior}, \Gamma_{prior}), \tag{2.52}$$

$$\boldsymbol{\theta}_i \sim \boldsymbol{N}(\theta_i|A\theta_{i-1}, \Gamma), \tag{2.53}$$

$$\boldsymbol{x}_i \sim \boldsymbol{N}(x_i|B\theta_i, R). \tag{2.54}$$

where Γ *is the dynamical covariance,* $(\Gamma_{prior}, \theta_{prior})$ *are the prior mean and prior covariance of the initial state, and lastly* R *is the measurement covariance.*

The solutions to the filtering, smoothing, and prediction problems for a Normal Linear Hidden Markov model are collectively known as the Kalman [1960] filter. In particular, because all the densities involved are Normal, the posterior density is also Normal due to the conjugacy properties of the Normal distribution. While the assumption of a linear Normal model of the measuring device may be reasonable in many situations, the same is not true of a linear dynamical model. Often, a nonlinear model of the state space is desired.

Consider an instance in which the state is modelled to evolve according to a system of autonomous ordinary differential equations. Let, for $s \in \mathbb{R}$, $\theta(s) \in \mathbb{R}^m$ and $F : \mathbb{R}^m \to \mathbb{R}^m$ satisfy the differential equation $d\theta(s)/ds = F(\theta(s))$. The sequence of states is the system state at the sequence of points s_i that correspond to a discretisation of the ordinary differential equation. One such discretisation method is the backward difference method. We will assume that the discretisation steps are equal. Let $s_i - s_{i-1} = \Delta \in \mathbb{R}$, for all i, be the discretisation step. Approximate the differential equation, using a backward difference method, as

$$\theta_i - \theta_{i-1} = \Delta F(\theta_i). \tag{2.55}$$

Note that this equation implies the existence of a function f such that $\theta_i = f(\theta_{i-1})$. Given θ_{i-1}, we solve the approximate implicit equation for θ_i using Newton's method. Let $z \in \mathbb{R}^m$ denote some initial guess of the root of (2.55), then a Taylor expansion around $z + \varepsilon$, where $\varepsilon \in \mathbb{R}^m$ is small and ignoring all but the first order term, is given by

$$0 = \Delta F(z) + \theta_{i-1} - z + (\Delta \nabla_z F(z) - I)\,\varepsilon + \cdots, \tag{2.56}$$

where $I \in \mathbb{S}_+^m$ is the identity matrix. Ignoring higher-order terms and solving the above equation yields $\varepsilon = (I - \Delta \nabla_z F(z))^{-1}(\theta_{i-1} - z + \Delta F(z))$. Updating z to $z + \varepsilon$, the process is repeated with the new z until a specified convergence criterion is fulfilled, or the maximum number of allowed iterations are completed. Then, set $\theta_i = z$ and repeat the process until the desired duration of the trajectory is achieved. The backward difference method is a single step method, as it only depends on the θ_i and the previous state, θ_{i-1}.

Structure 8. *Single Step Discretised Differential Equation*

Consider a Markov type Model (Structure 5). Let $\nabla_s \theta(s) = F(\theta(s))$ be a differential equation modelling the dynamics of the state space, and let $f(\theta_{i-1}) = \theta_i$ be the function implied from the single-step discretisation of the differential equation, so that

$$\pi(\theta_i|\theta_{i-1}) = \delta(\theta_i - \theta_{i-1} + \Delta \nabla_\theta F(\theta_i)) \tag{2.57}$$

$$= \delta(\theta_i - f(\theta_{i-1})). \tag{2.58}$$

If we wish to model a time series using a differential equation, and we believe the measurement-noise to be Normal, we could employ the following model.

Model 10. *Deterministic Nonlinear State Space with a Normal Linear Measuring Device. Build a Hidden Markov model (Model 7) with a Normal Linear Measuring Device (Structure 2) and a Single Step Discretised Differential Equation (Structure 8) modelling the trajectory of the state parameters. Let*

$$\boldsymbol{\theta}_0 \sim \boldsymbol{N}(\theta_{prior}, \Gamma_{prior}), \tag{2.59}$$

$$\boldsymbol{\theta}_i \sim \delta(\theta_i - f(\theta_{i-1})), \tag{2.60}$$

$$\boldsymbol{x}_i \sim \boldsymbol{N}(x_i|\theta_i, R). \tag{2.61}$$

Generally, an analytic form of the posterior density is not available. If the maximum posterior estimate can be found, a Normal approximation to the posterior density is possible.

2.3.3 Approximating the Posterior

Using only the maximum posterior estimate is equivalent to approximating the posterior density of the parameters using a delta function

$$\pi(\theta|\mathring{x}_{1:t}) \approx \delta(\theta - \hat{\theta}), \tag{2.62}$$

so that the predictive density is $\pi(y|\mathring{x}_{1:t}) \cong \pi(y|\hat{\theta})$. A better approximation of the posterior density can be obtained by using a Taylor approximation near the maximum posterior estimate, which leads to an approximating Normal posterior with mean equal to the maximum posterior estimate and covariance given by the inverse of the Hessian evaluated at the maximum posterior estimate [Särkkä, 2013].

Proposition 1. *Build a Bayesian model (Model 4). A second order approximation near a maximum of the posterior density yields the Normal approximation*

$$\pi(\theta|\mathring{x}_{1:t}) \approx \boldsymbol{N}(\theta|\hat{\theta}, t^{-1}\nabla_\theta^2 \Phi_t(\theta)^{-1}), \tag{2.63}$$

where $\Phi_t(\theta) = -t^{-1} \log \pi(\mathring{x}_{1:t}|\theta)\pi(\theta)$ *is the misfit function [Särkkä, 2013].*

Proof. The log posterior density is proportional to the negative of the misfit function

$$\log \pi(\theta|\mathring{x}_{1:t}) \propto \log \pi(\mathring{x}_{1:t}|\theta)\pi(\theta) = -t\Phi_t(\theta), \tag{2.64}$$

where $\hat{\theta}$, the maximum posterior estimate, satisfies $\nabla_\theta \Phi_t(\hat{\theta}) = 0$. A second-order Taylor expansion of the misfit function, around the maximum posterior estimate, is

$$t\Phi_t(\theta) \approx t\Phi_t(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T t\nabla_\theta \Phi_t(\theta)(\theta - \hat{\theta}) + \ldots, \tag{2.65}$$

where the first order term is zero. The posterior density is approximately proportional to a Normal density.

$$\pi(\theta|\mathring{x}_{1:t}) \propto \exp\left(-t\Phi_t(\hat{\theta})\right) \exp\left(-\frac{1}{2}(\theta-\hat{\theta})^T t\nabla^2_\theta \Phi_t(\hat{\theta})(\theta-\hat{\theta})\right) \tag{2.66}$$

$$\implies \pi(\theta|\mathring{x}_{1:t}) \approx \boldsymbol{N}\left(\theta|\hat{\theta}, t^{-1}\nabla^2_\theta \Phi_t(\hat{\theta})^{-1}\right). \tag{2.67}$$

□

An application of this approximation is to Bayesian inverse problems characterised by use of differential equations. This approximation is particularly good when the posterior density is unimodal. For a dynamic sequence of measurements, consider models with conditionally independent measurements and a state space trajectory.

Model 11. *Normal Noisy Measurements*

Let $h : \mathbb{R}^m \to \mathbb{R}^n$ be some differentiable nonlinear function modelling the measuring device. Let

$$\boldsymbol{\theta} \sim \boldsymbol{N}(\theta_{prior}, \Gamma_{prior}), \tag{2.68}$$

$$\boldsymbol{x} \sim \boldsymbol{N}(x|h(\theta), \Gamma_{noise}). \tag{2.69}$$

Even though the densities involved are Normal, the introduction of the nonlinear function results in a non-Normal posterior. The next proposition derives a Normal approximation to the posterior in the case of one measurement.

Proposition 2. *The Normal approximation to the posterior density of Model 11 is*

$$\pi(\theta|\mathring{x}) \approx \boldsymbol{N}(\theta|\hat{\theta}, (A^T\Gamma^{-1}_{noise}A + \Gamma^{-1}_{prior})^{-1}), \tag{2.70}$$

where $\hat{\theta}$ is the maximum posterior estimate, and A is the Jacobian of h evaluated at the maximum posterior estimate.

Proof. By Bayes' theorem, the posterior is proportional to the product of the two Normal densities. The misfit function is given by

$$\Phi_1(\theta) = \frac{1}{2}||\mathring{x} - h(\theta)||^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}||\theta - \theta_{\text{prior}}||^2_{\Gamma^{-1}_{\text{prior}}}. \tag{2.71}$$

The maximum posterior estimate, $\hat{\theta}$, is found by minimising the misfit function, or equivalently solving

$$0 = \nabla_\theta \Phi_1(\theta) = -\Gamma^{-1}_{\text{noise}}(\mathring{x} - h(\theta))\nabla_\theta h(\theta) + \Gamma^{-1}_{\text{prior}}(\theta - \theta_{\text{prior}}). \tag{2.72}$$

Notice that this equation requires computation of the Jacobian of the nonlinear function h. To calculate the posterior variance, we evaluate the second derivative at the maximum posterior estimate,

$$\nabla^2_\theta \Phi_1(\theta) = \nabla_\theta h(\theta)^T \Gamma^{-1}_{\text{noise}} \nabla_\theta h(\theta) + \Gamma^{-1}_{\text{noise}} \nabla^2_\theta h(\theta)(\mathring{x} - h(\theta)) + \Gamma^{-1}_{\text{prior}}. \tag{2.73}$$

Often, the Jacobian is difficult enough to compute, and computation of the Hessian is extremely difficult. To avoid computing the Hessian, we linearly approximate the measuring device near to the maximum posterior estimate as $h(\theta) \approx h(\hat{\theta}) + A(\hat{\theta})(\theta - \hat{\theta})$, where A is the Jacobian of f evaluated at the maximum posterior estimate. Calculation of the posterior covariance simplifies to

$$\nabla^2 \Phi_1(\hat{\theta}) \approx A^T \Gamma^{-1}_{\text{noise}} A + \Gamma^{-1}_{\text{prior}} \implies \Gamma_{\text{post}} \approx (A^T \Gamma^{-1}_{\text{noise}} A + \Gamma^{-1}_{\text{prior}})^{-1}. \tag{2.74}$$

□

Model 12. *Normal Noisy Linearised Differential Equations*
Let $h(\cdot)$ be a model of the measurement device, and f be derived from the single step discretisation of a differential equation (Structure 8), modelling the evolution of the state. Let

$$\boldsymbol{\theta}_0 \sim \boldsymbol{N}(\theta_0 | \theta_{prior}, \Gamma_{prior}), \tag{2.75}$$

$$\boldsymbol{\theta}_i \sim \delta(\theta_i - f(\theta_{i-1})), \tag{2.76}$$

$$\boldsymbol{x}_i \sim \boldsymbol{N}(x_i | h(\theta_i), R). \tag{2.77}$$

This model is a generalisation of Model 9, in which the evolution of the state vector is given by the discretisation of a differential equation, and the measurements are nonlinear functions of the model parameters. The misfit function is given by

$$t\Phi_t(\theta_0) = \sum_{i=1}^{t} \frac{1}{2} ||\mathring{x}_i - h(\theta_i)||^2_{R^{-1}} + \frac{1}{2} ||\theta_0 - \theta_{\text{prior}}||^2_{\Gamma^{-1}_{\text{prior}}}. \tag{2.78}$$

Once the maximum posterior estimate of the state is found, we wish to characterise the posterior variance. Let $A(\hat{\theta}_i) = \nabla_\theta h(\hat{\theta}_i)$ be the Jacobian. The Hessian of the misfit function evaluated at the maximum posterior estimate is

$$t\nabla^2_\theta \Phi_t(\hat{\theta}_0) = \sum_{i=1}^{t} A(\hat{\theta}_i)^T R^{-1} A(\hat{\theta}_i) + \Gamma^{-1}_{\text{prior}}. \tag{2.79}$$

Therefore, the Normal approximation to the posterior of the initial state is written, applying Proposition 1, as

$$\pi(\theta_0|x_{1:t}) \approx \boldsymbol{N}\left(\theta_0 \middle| \hat{\theta}, t^{-1}\left(\sum_{i=1}^{t} A(\hat{\theta}_i)^T R^{-1} A(\hat{\theta}_i) + \Gamma^{-1}_{\text{prior}}\right)^{-1}\right). \tag{2.80}$$

It must be stressed that by no means is the Normal approximation to the posterior accurate. This must be checked on a model-by-model basis by investigating the behaviour of the misfit function around the maximum posterior estimate. These types of approximate Normal models are often utilised for large scale dynamical systems governed by differential equations. Stuart [2010] discuss the conditions under which the optimisation problem is well posed. Cotter et al. [2010] uses the Hellinger distance to quantify the approximation of the posterior. More recently, Bui-Thanh et al. [2012] exploit matrix sparsity to construct a low rank approximation of the Hessian. Schillings and Schwab [2013] consider forward models that are parametric deterministic elliptic partial differential equations.

In this chapter, we discussed the building of statistical models and the derivation of the predictive density. For a Bayesian model, this inference process and prediction is summarised as

$$\pi(y|\mathring{x}_{1:t}) = \int \pi(y|\theta)\pi(\theta|\mathring{x}_{1:t})d\theta. \tag{2.81}$$

Once the measurements have been assimilated, we are sometimes interested in the posterior predictive densities, otherwise known as the smoothed measurement densities,

$$\pi(x_i|\mathring{x}_{1:t}) = \int \pi(x_i|\theta)\pi(\theta|\mathring{x}_{1:t})d\theta, \tag{2.82}$$

where we have replaced the prior with the posterior and computed predictive densities as usual. In the case of a Frequentist model, the smoothed measurement densities are $\pi(x_i|\hat{\theta})$, the likelihood function evaluated at the maximum likelihood estimate. Once the posterior density is obtained, more work yields the predictive density. If this is possible, then the posterior predictive densities can be obtained.

When employing nonlinear models, there arises a problem with estimation of parameters for which there is no general solution. Studying a sequence of generated measurements allows us to perform experiments under controlled conditions. Nonlinear processes can be sensitive to initial conditions and parameter specifications. Even if the uncertainty in the model and measurements are Normal, classical methods, such as least-squares, are generally inapplicable because of the possibility of many local minima.

Before the next measurement is made, any decision made will depend on the assessed predictive density. The decisions we are interested in making are those pertaining to the usefulness of the model under consideration. In particular, two questions arise. Firstly, given a number of models, which model is the best? Secondly, once we have identified the best model, how do we evaluate its performance?

The next chapter introduces the theory of statistical decisions made under uncertainty. Prediction is understood as a decision problem where the actions can be either points or probabilities. This leads to loss functions known as proper scoring rules.

Chapter 3

Statistical Decisions Under Uncertainty

A statistical decision is a decision made on the basis of the value of a variable, conditional on measurements. When the decision variable is an unknown model parameter, the uncertainty in its value is given by a probability distribution, furnished by a probabilistic model. The resulting decision, depending on both the actualising value of the measurements, and the uncertainty in the decision parameter, is a statistical decision made under uncertainty.

The chapter begins with a description of the elements of a coherent decision problem and a review of some classical decision problems. The classical approach is characterised by fixing a loss function and deriving the associated optimal action. In scientific experimentation decisions are often ad hoc but made in an informed manner. Instead of fixing a loss function and finding the optimal action we instead consider the action undertaken to be fixed and describe the behaviour of the associated loss function. To illustrate this we consider a decision made in the analysis of genomic sequences and describe the behaviour of the loss function for which that decision is optimal. To conclude the chapter we extend the analysis by framing probabilistic prediction as a decision problem in which the predictive probability is the optimal action.

In this chapter we consider a specific problem that arises in the subject of computational biology. This problem has been published collaboratively in van Tonder et al. [2014] where the entire statistical analysis is contributed by the author of this thesis. The presentation of the problem in this thesis expands on the statistical aspects of the problem and summarises the resulting biological findings of the published article. The solution to this problem is one of the novel contributions of this thesis.

In particular we seek to estimate a set of *core genes* (see Section 3.2) from a large collection of bacterial genomes. While a Bayesian would be satisfied by just computing the posterior density the biologist is interested in determining a set of genes that can be then be further studied in the laboratory. Limited resources and time demand that the biologist decide which genes should be investigated further. We formulate this intention as a decision problem and derive a decision rule to decide whether or not a gene is *core*. The chapter concludes by framing probabilistic prediction as a decision problem in which the loss function is designed so that the predictive density is the optimal action.

3.1 Decision Theory

For a complete treatment of decision theory reasoning, we refer to Raiffa and Schlaifer [1951]. Rather than analyse the decision problem in terms of utility, we instead analyse it in terms of opportunity loss, and assume that the terminal and sampling utilities are additive [Raiffa and Schlaifer, 1951, DeGroot, 1970]. This way we speak of *loss* as opposed to *utility*.

Within an inference context, we require particular elements to make a coherent, quantitative decision [Bernardo and Smith, 2001]. We begin by making a sequence of measurements, $\mathring{x}_{1:t}$, and constructing a statistical model $\pi(y, \theta, x_{1:t}|M)$ of the parameters of interest, the within-model parameters, and the measurements. If the model is Bayesian, then $\pi(\theta|M)$ is the within-model prior.

The goal of decision theory is, based on the measurements, to decide on an action amongst

a set of allowable actions, $a \in \mathcal{A}$. A statistical decision is made by specifying which action is chosen, conditional on a particular sequence of measurements obtained. The statistical decision is made under uncertainty if the decision depends on a parameter that has not been observed. Letting $\boldsymbol{y}$ denote the unknown model parameter of interest, the uncertainty for a Bayesian model is

$$\pi(y|\mathring{x}_{1:t}, M) = \int \pi(y|\theta)\pi(\theta|\mathring{x}_{1:t}, M)d\theta. \tag{3.1}$$

When the whole posterior density is not available, a plug-in estimate $\hat{\theta}$, estimated from the observations, can be used, and the forecast is $\pi(y|\mathring{x}_{1:t}, M) \cong \pi(y|\hat{\theta})$, where the symbol $\cong$ means "in place of". In this chapter, we are interested in a single arbitrary model, and so M is dropped from the notation.

The consequence of the decision is expressed using a real-valued *loss* function, $\mathcal{L} : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, which specifies the cost that is incurred when the action is $a \in \mathcal{A}$ and the parameter is $y \in \mathcal{Y}$. A decision is made such that the action chosen is optimal with respect to the loss function. If $\boldsymbol{y}$ has already been observed there is no uncertainty in the actualising value $\mathring{y}$. Then, we would choose the action $\tilde{a}$ that minimises the loss

$$\tilde{a} = \underset{a}{\operatorname{argmin}}\, \mathcal{L}(\mathring{y}, a). \tag{3.2}$$

In most scenarios, $\boldsymbol{y}$ has not yet been observed, or may never be observed, and so we compute the expected loss in which the loss function is averaged against the computed uncertainty of $\boldsymbol{y}$, the predictive density $\pi(y|\mathring{x}_{1:t})$. The expected loss is known as the risk.

Definition 3. *Given a parameter $y \in \mathcal{Y}$, with uncertainty described by $\boldsymbol{y} \sim \pi(y|\mathring{x}_{1:t})$, the expected loss, or risk, is*

$$\mathbb{E}[\mathcal{L}(\boldsymbol{y}, a)] = \int_{\mathcal{Y}} \mathcal{L}(y, a)\pi(y|\mathring{x}_{1:t})dy. \tag{3.3}$$

Definition 4. *The action that minimises the risk is known as the Bayes action,*

$$\tilde{a} = \underset{a}{\operatorname{argmin}}\, \mathbb{E}[\mathcal{L}(\boldsymbol{y}, a)] \tag{3.4}$$

Loss functions that lead to equivalent decisions are said to be equivalent. For instance, if $\tilde{a}$ is the Bayes action associated to the loss function $\mathcal{L}$, then the loss function $\mathcal{L}^*$ defined in (3.5) below is equivalent, where C is a positive constant and $k(\mathring{x}_{1:t})$ is some arbitrary function of the measurements alone.

$$\mathcal{L}^*(y, a) = C\mathcal{L}(y, a) + k(\mathring{x}_{1:t}). \tag{3.5}$$

Proof. If $\tilde{a}$ is the Bayes action associated to the loss function $\mathcal{L}$, then taking a derivative of both sides evaluated at $\tilde{a}$ shows that it also minimises $\mathcal{L}^*$. □

Given a set of measurements and a statistical model we can make decisions based on marginals from the posterior density. There are many types of decisions that could be made, depending on the phenomenon under study and the goals of the experiment. The classic application of statistical decision under uncertainty is that of point parameter estimation. For particular fixed loss functions, the optimal choice for a parameter estimate is given as a function of the marginal posterior density. We consider the examples of the squared loss function and the absolute loss function for which the optimal decision is the predictive mean and predictive median respectively.

3.1.1 Point Estimates

We are interested in making point estimates of the parameter of interest. For particular loss functions the Bayes action can be found analytically as a function of the posterior density. Within the context of prediction, we provide three examples of loss functions, and their related Bayes actions. These are standard results that can be found, for example, in O'Hagan and Forster [2004] We consider loss functions for which the Bayes action is the maximum posterior estimate, the expected value, and the q-quantile. The maximum posterior estimate can be understood as solving the decision problem under the 0/1 loss function.

Proposition 3. *If we use the* 0/1 *loss function, which assigns zero cost to the correct deci-*

sion, and unit cost to a wrong decision

$$\mathcal{L}(y,a) = \begin{cases} 0 & \text{if } y = a \\ 1 & \text{if } y \neq a \end{cases} \tag{3.6}$$

then the Bayes action is the maximum posterior estimate.

Proof. To show this, consider, for $\epsilon > 0$, the set $B_\epsilon(a) = \{y \in \mathcal{Y}; ||a - y|| < \epsilon\}$ which converges to the singleton $\{a\}$ as $\epsilon \to 0$ and where $||\cdot||$ denotes the Euclidean distance. Then we can write the loss function (3.6), in the limit as $\epsilon \to 0$ as

$$\mathcal{L}(y,a) = \lim_{\epsilon \to 0} \mathbb{1}\{y \notin B_\epsilon(a)\}. \tag{3.7}$$

The value of a which minimises expected loss can be written, using (3.7), as

$$\tilde{a} = \operatorname*{argmin}_a \mathbb{E}[\mathcal{L}(\boldsymbol{y}, a)] = \operatorname*{argmin}_a \mathbb{E}\left[\lim_{\epsilon \to 0} \mathbb{1}\{\boldsymbol{y} \notin B_\epsilon(a)\}\right] \tag{3.8}$$

where $\boldsymbol{y} \sim \pi(y|\mathring{x}_{1:t})$. Notice that for any fixed ϵ, $\mathbb{1}\{y \notin B_\epsilon(a)\}$ is an integrable function. Then interchanging the limit and expectation is possible using the dominated convergence theorem [Rudin, 1964], and noticing that the expectation of the indicator function is the probability of the indicated event, the right-hand side of (3.8) can be written as

$$\operatorname*{argmin}_a \lim_{\epsilon \to 0} \mathbb{E}\left[\mathbb{1}\{\boldsymbol{y} \notin B_\epsilon(a)\}\right] = \operatorname*{argmin}_a \lim_{\epsilon \to 0} \mathbb{P}\{\boldsymbol{y} \notin B_\epsilon(a)\} \tag{3.9}$$

$$= \operatorname*{argmin}_a \lim_{\epsilon \to 0} (1 - \mathbb{P}\{\boldsymbol{y} \in B_\epsilon(a)\}) \tag{3.10}$$

$$= \operatorname*{argmax}_a \lim_{\epsilon \to 0} \mathbb{P}\{\boldsymbol{y} \in B_\epsilon(a)\} \tag{3.11}$$

$$= \operatorname*{argmax}_a \mathbb{P}\{\boldsymbol{y} = a\} \tag{3.12}$$

where for real valued random variables $\mathbb{P}\{\boldsymbol{y} = a\}$ is the probability density function $\pi(a|\mathring{x}_{1:t})$ so that $\tilde{a}$ is the maximum posterior estimate. □

Alternative measures of centrality are attractive. For instance, consider the quadratic loss for which the Bayes action is the mean of the predictive density.

Proposition 4. *The Bayes action $\tilde{a}$ associated to the quadratic loss function*

$$\mathcal{L}(y,a) = (a-y)^T Q(a-y), \tag{3.13}$$

is the predictive mean, where Q is a symmetric positive definite matrix.

Proof. Let $\boldsymbol{y} \sim \pi(y|\mathring{x}_{1:t})$ so that the Bayes action $\tilde{a}$ satisfies

$$0 = \left(\frac{\partial}{\partial a}\mathbb{E}[(a-\boldsymbol{y})^T Q(a-\boldsymbol{y})]\right)\Big|_{a=\tilde{a}} = \mathbb{E}[2Q(a-\boldsymbol{y})]|_{a=\tilde{a}} = 2Q(\tilde{a}-\mathbb{E}[\boldsymbol{y}]). \tag{3.14}$$

From which it follows that $\tilde{a} = \mathbb{E}[\boldsymbol{y}]$. □

A third example is, for $y \in \mathbb{R}$, the absolute loss function

$$\mathcal{L}(y,a) = |y-a|. \tag{3.15}$$

Instead of considering the absolute loss function we analyse a generalisation of it,

$$\mathcal{L}(y,a;q_1,q_2) = q_1(a-y)\mathbb{H}(a-y) + q_2(y-a)\mathbb{H}(y-a), \tag{3.16}$$

where $q_1, q_2 > 0$ are known positive numbers and $\mathbb{H}(\cdot)$ is the Heaviside function. Equation (3.15) is recovered from (3.16) by setting $q_1 = q_2$.

Proposition 5. *The Bayes action associated to the loss function $\mathcal{L}(y,a;q_1,q_2)$ is the q-quantile of $\pi(y|\mathring{x}_{1:t})$ where $q = \frac{q_2}{q_1+q_2}$.*

Proof. To prove this we first write down the risk as an integral.

$$\mathbb{E}[\mathcal{L}(\boldsymbol{y},a;q_1,q_2)] = q_1\int_{-\infty}^{a}(a-y)\pi(y|\mathring{x}_{1:t})dy + q_2\int_{a}^{\infty}(y-a)\pi(y|\mathring{x}_{1:t})dy. \tag{3.17}$$

Write, using $\frac{\partial}{\partial y}\Pi(y) = \pi(y|\mathring{x}_{1:t})$, the first integral on the right-hand side of (3.17) as

$$q_1\int_{-\infty}^{a}(a-y)\pi(y|\mathring{x}_{1:t})dy = q_1\int_{-\infty}^{a}(a-y)\frac{\partial}{\partial y}\Pi(y)dy \tag{3.18}$$

$$= q_1\left((a-y)\Pi(y)\right)|_{-\infty}^{a} + q_1\int_{-\infty}^{a}\Pi(y)dy. \tag{3.19}$$

The upper part of the first term vanishes at $y = a$. The lower part vanishes since $\lim_{y\to-\infty} \Pi(y) = 0$. We can use similar arguments to evaluate the second integral in (3.17) with some differences. To evaluate this integral, we write it in terms of the complementary distribution function $\overline{\Pi}(y) = 1 - \Pi(y)$. Then, the second integral in (3.17) can be written in a manner similar to (3.18) as

$$q_2 \int_a^\infty (y-a)\pi(y|\mathring{x}_{1:t})dy = q_2 \left((y-a)\overline{\Pi}(y)\right)\Big|_a^\infty + q_2 \int_a^\infty \overline{\Pi}(y)dy. \tag{3.20}$$

Then, substituting the simplified expressions (3.18) and (3.20) into (3.17) yields

$$\mathbb{E}[\mathcal{L}(\boldsymbol{y}, a; q_1, q_2)] = q_1 \int_{-\infty}^a \Pi(y)dy - q_2 \int_\infty^a \overline{\Pi}(y)dy. \tag{3.21}$$

Then, take a derivative with respect to a, set it equal to zero, and solve for the Bayes action

$$0 = q_1\Pi(a) - q_2\overline{\Pi}(a) \implies \Pi(a) = \frac{q_2}{q_1 + q_2} = q. \tag{3.22}$$

It follows that $\tilde{a} = \Pi^{-1}(q)$ is the q-quantile and so completes the proof. □

Consider the following loss function, which depends on some fixed constants c and l.

$$\mathcal{L}(y, a; c, l) = ca + l(y-a)\mathbb{H}(y-a). \tag{3.23}$$

While this loss function can be shown to be equivalent to (3.16), written in this way we can give an economic interpretation to the constants c and l. We can understand ca as the cost associated with employing action a while $l(y-a)$ can be interpreted as the expense incurred after predicting a, and observing y.

Proposition 6. *The Bayes action associated with the loss function $\mathcal{L}(y, a; c, l)$ is the q-quantile of $\pi(y|\mathring{x}_{1:t})$ where $q = 1 - c/l$ and c and l are nonnegative constants such that $c \leq l$.*

Proof. The proof follows by substitution. □

This loss function penalises the underestimation of y, but the loss function can be reversed so as to penalise overestimation. In practice, the loss function should be selected to reflect

the decision problem that we wish to be solved. As such an example, we consider a decision problem that arises in computational biology of estimating, given a number of genomes, a *core* set of genes.

3.2 Estimating the Core Genome

Within a laboratory, cultures of bacteria known as *isolates* are isolated for study. The DNA of each of these isolates can be sequenced, and then analysed using statistical techniques. More recently, advances in DNA sequencing have resulted in a vast increase in the number of bacterial genomes that have been sequenced and made available for study, in public databases such as GenBank, the Sequence Read Archive, and European Nucleotide Archive. Increasing computational power allows for comparative genomics studies involving hundreds or even thousands of sequences.

A bacterial species can be defined by its pan-genome, which consists of a core genome conventionally defined as those genes present in all isolates, and an accessory genome, which includes the genes absent from one or more isolates or unique to a given isolate [Tettelin et al., 2005]. Identifying the core complement of genes in a bacterial species is often the first step in population genomics studies, and the core genome can be defined in different ways. The most conservative and most frequently employed method is only to include genes present in all of the isolates within the study population. However, this presents problems related to both biological sampling and the sequencing technology. This classical approach to estimating the core genome depends on the Frequentist estimate of the gene frequency. The Frequentist estimate is computed as the frequency of the gene occurrence within the sample.

Any collection of isolates is a subset of the entire population for the species of interest, and if the subset of isolates has limited genetic diversity, then the number of core genes shared by all isolates in that sample will be higher than in a dataset which is genetically more diverse.

This is not necessarily a problem, unless the intention is to extrapolate the findings to the wider bacterial population. Another biological limitation to using the Frequentist estimate is that there may be rare variant strains which are missing genes that would otherwise be considered core genes. These variant strains may survive long enough to be sampled, potentially skewing the analysis.

Due to imprecision in the collection and processing of the genomic sequences, we are prone to underestimating the probability of a gene occurring. To reflect the nature of this problem, we introduce a loss function that harshly penalises underestimation. We proceed to make a decision, based on the posterior density and the loss function, on the action of whether or not to include a gene in the core. The action is shown to be equivalent to checking whether the posterior density evaluated at one is zero.

3.2.1 Measurements

For each of the collections of bacterial genomes, we have a chosen reference genome, consisting of a finite set of genes. Each gene is considered independently during the analysis. Let t be the number of isolates under consideration. Let $\boldsymbol{x}_i \in \{0, 1\}$ be a binary random variable so that $\sum_{i=1}^{t} \mathring{x}_i$ is the number of times we observe the gene in the set of t isolates. We model the occurrence of any gene using a beta-binomial model (Model 5) and find, for each gene in the reference, the posterior density $\pi(\theta|\mathring{x}_{1:t})$.

This decision model is applied to a number of large genomic databases. Collections of the bacteria *H. pylori* ($t = 107$), *S. aureus* ($t = 534$), *S. pneumoniae* ($t = 336$), *N. meningitidis* ($t = 514$), and *C. jejune* ($t = 601$), were analysed using this model. The resulting lists of core genes were further investigated for biological significance. Further particulars of the datasets are described in van Tonder et al. [2014], where the Bayesian decision model developed in this section is published.

3.2.2 The Classical Core Genome

The classical approach to defining the core genome is only to consider genes which are present in all strains. This is equivalent to computing the frequency of the gene occurrence within the sample, and only selecting genes for which the frequency is one. This is an estimate of $\boldsymbol{\theta}$, a random variable modelling the probability parameter. We decide that a gene is core if $\hat{\theta}$, the parameter estimate, is equal to one. Let $a(\mathring{x}_{1:t}) \in \{0, 1\}$ be a binary action depending on the observed measurements. For each gene, if $a(\mathring{x}_{1:t}) = 1$, the gene is included in the estimated core genome. If $a(\mathring{x}_{1:t}) = 0$, the gene is excluded.

The classical decision rule, including only the genes that are observed to be present in all strains, is given by

$$a(\mathring{x}_{1:t}) = 1 \iff \hat{\theta} = 1 \text{ where } \hat{\theta} = \frac{1}{t}\sum_{i=1}^{t} \mathring{x}_i. \tag{3.24}$$

The decision rule is conservative in the sense that only genes present in all strains are identified as core genes. The estimate of the probability parameter is sensitive to underestimation. If an error in the sequencing process occurs, and a gene is not identified, the estimate of the probability parameter is smaller than it should be. To reflect this behaviour, we employ a conservative loss function in which we penalise underestimation of the probability harshly. The harsher the penalty, the more conservative the estimate of the probability parameter.

The estimator, $\hat{\theta} = \frac{1}{t}\sum_{i=1}^{t} \mathring{x}_i$ is the maximum likelihood estimator which, as noted in Section 2.3.2 is equivalent to the maximum posterior estimate when the prior is proportional to one. As such we consider an extension of this estimator to Bayesian models in which the prior is selected, either proportional to one or otherwise.

Beta-Binomial Model

The Frequentist approach computes, as the estimate of the probability parameter, the maximum likelihood estimate. The measurements are modelled as conditionally independent, and

identically distributed Bernoulli random variables. Rather than employing a Frequentist approach to estimate the probability parameter, we instead propose a Bayesian model of the probability parameter, and formulate a decision rule dependent on the posterior. Perhaps the most basic model of the probability parameter is the beta-binomial model, in which the prior is a beta distribution, the likelihood is a binomial distribution, and the posterior is a beta distribution.

For each gene, we compute the posterior density given the observations. Let $\theta \in [0, 1]$ be the probability of observing the gene. The parameter has not been, nor ever will be, observed. It is only known through inference using the available measurements. Let the prior $\pi(\theta)$ be a beta density with parameters α and β. Let the likelihood function $\pi(x_i|\theta)$ be a binomial density. Then the posterior density is given by Bayes' as

$$\pi(\theta|\mathring{x}_{1:t}) \propto \theta^{\alpha+n-1}(1-\theta)^{\beta+t-n-1}, \tag{3.25}$$

where $n = \sum_{i=1}^{t} \mathring{x}_i$. The posterior distribution is a beta distribution with parameters $\alpha^* = \alpha + n$ and $\beta^* = \beta + t - n$.

3.2.3 A Conservative Bayesian Decision Rule

In this section, we describe the decision problem by selecting a prior and fixing a loss function. The prior and loss function are selected to achieve the most conservative estimate of the estimated core genome. The resulting decision is shown to be equivalent to checking whether the posterior density is non zero when evaluated at one,

$$a(\mathring{x}_{1:t}) = 1 \iff \pi(\theta|\mathring{x}_{1:t})|_{\theta=1} > 0. \tag{3.26}$$

Conservative Loss Function

To avoid the problem of underestimation, we wish to make the *most* conservative estimate possible by penalising underestimates as harshly as possible. For some fixed $c > 0$, let $c\hat{\theta}$ be

the minimum fixed cost of making any estimate. For $l \geq c$, let $l(\theta - \hat{\theta})$ be the additional loss incurred when we underestimate the parameter. Thus the loss incurred when estimating $\hat{\theta}$, when the actual value is θ, given by (3.23), is

$$\mathcal{L}(\theta, \hat{\theta}; c, l) = c\hat{\theta} + l(\theta - \hat{\theta})\mathbb{H}(\theta - \hat{\theta}). \tag{3.27}$$

The parameters of the loss function, c and l, are understood using the economic interpretation of cost and loss, but the equivalence of the loss functions in Propositions 5 and 6 showing that fixing c and l is analogous to analysing the posterior density at a fixed quantile q, where the relationship between c, l and q are given by the previously mentioned propositions.

That is, employing an economic interpretation, if we were able to observe $\boldsymbol{\theta} = \theta$ directly, we would incur a cost, $l(\theta - \hat{\theta})$ when our estimate $\hat{\theta}$ is smaller than θ on top of the fixed cost $c\hat{\theta}$. The larger the choice of l, the more averse we become to underestimating $\hat{\theta}$, the probability parameter. Analogously, choosing l larger, while c remains fixed implies that we are interested in the behaviour of the left tail of the posterior density. In the analysis of the core genomes the non-economic interpretation of the loss function seems appropriate in that, for c fixed and as $l \to \infty$ we focus on the behaviour of the posterior density near $\theta = 1$, the value of the parameter relevant to deciding whether a gene is core or not.

In reality, we never observe $\boldsymbol{\theta}$, and can only estimate the expected loss dependent on $\mathring{x}_{1:t}$, the binary measurements. The decision is made under our uncertainty of $\boldsymbol{\theta}$, where the uncertainty is the posterior density conditional on the choice of model. We then select as our estimate

$$\hat{\theta} = \underset{\theta' \in [0,1]}{\operatorname{argmin}} \int_0^1 \mathcal{L}(\theta, \theta'; c, l)\pi(\theta|\mathring{x}_{1:t})d\theta, \tag{3.28}$$

which minimises the expected loss. From Proposition 6, the Bayes action is, for $q \in [0, 1]$, the q-quantile of the posterior density which is related to c and l. The estimate, for fixed l and arbitrary c, is

$$\hat{\theta}_l = \Pi^{-1}(q) = \Pi^{-1}\left(1 - \frac{c}{l}\right), \tag{3.29}$$

where $\Pi(\cdot)$ is the cumulative distribution function of the posterior density. As $l \to \infty$ the penalty for underestimation becomes harsher. Write

$$\hat{\theta} = \lim_{l\to\infty} \hat{\theta}_l = \lim_{l\to\infty} \Pi^{-1}\left(1 - \frac{c}{l}\right) = \lim_{q\to 1} \Pi^{-1}(q) = \inf\left\{\theta; \Pi(\theta) \geq 1\right\}. \tag{3.30}$$

Recall that the decision rule for whether or not a gene is core, is given by $a(\mathring{x}_{1:t}) = 1 \iff \hat{\theta} = 1$ (3.24). Since $\hat{\theta} \in [0, 1]$ we can rewrite the decision rule by noticing that, as $q \to 1$ and using (3.30), the estimate $\hat{\theta}$ must satisfy

$$\int_0^{\hat{\theta}} \pi(\theta|\mathring{x}_{1:t})d\theta = 1 \tag{3.31}$$

It follows that $\hat{\theta} < 1 \iff \pi(\theta|\mathring{x}_{1:t})|_{\theta=1} = 0$. Therefore

$$a(\mathring{x}_{1:t}) = 1 \iff \pi(\theta|\mathring{x}_{1:t})|_{\theta=1} > 0. \tag{3.32}$$

Conservative Prior

The selection of a prior amounts to specifying our prior belief as to whether a gene is, or is not, in the core genome. We might also adopt the belief that we are equally unsure of whether a gene is present in the core or not. Priors that reflect this type of belief are known as *near ignorance* priors. Rather than selecting a near ignorance prior, we argue that a prior should be selected to reflect the nature of the decision process. Prior to analysing any of the isolates, we have no reason to believe that any particular gene is or is not in the core. As each strain is analysed, we accumulate evidence to suggest that a gene is not a core gene. To reflect this process, we adopt the prior belief that every gene is a core gene and then employ the posterior distribution $\pi(\theta|\mathring{x}_{1:t})$ to make the decision whether or not to remove the gene from the estimated core. Formally, this corresponds to setting the prior to a beta density with parameters $(\alpha, \beta) = (1, 0)$.

3.2.4 Results of the Decision Model

The decision problem can be applied to any Bayesian model of the probability parameter. The conservative choice of prior and the conservative loss function are utilised for a number of genes, to determine the estimated core genome for a number of bacterial strains.

Recently, Croucher et al. [2013] approached the problem of estimating the core genome by identifying clusters of orthologous genes (COGs) using computationally intensive gene prediction techniques (*Orthologous* genes are those genes that are thought to share a common ancestor so that clusters of orthologous genes are related genes from different species that appear to share a common ancestor). COGs present in 100% of genomes in a single copy were defined as core genes. While the methodology can be quite computationally expensive, this approach has the benefit of not being dependent on a choice of reference.

Overall, the COGs-based method estimated the size of the core genome in the Massachusetts *S. pneumoniae* dataset to include 1194 genes, whereas our Bayesian analysis of the same dataset estimated 948 core genes. In total, 840 genes were common to both estimations, and 108 and 354 genes were unique to each methodology respectively. An explanation for the difference in core genome estimates was that we set the Genome Comparator analysis to consider only full-length genes, whereas partial sequences were included in the COGs-based methodology.

Running Genome Comparator, lowering the threshold to 90% sequence alignment, so as to allow up to 10% of the alignment sequence to be missing, results in similar estimated core genes to the COGs methodology. Specifically, the Bayesian decision model estimated 1206 core genes, compared to the 1194 genes estimated using the COGs methodology, of which 1027 genes were common to both methodologies. Lowering the threshold to 90% is an ad hoc decision similar to the classical core gene decision. The comparison with the COGs based method suggests that the sequence alignment threshold is a decision parameter. This alludes to another decision problem that could be analysed, suggesting an avenue for further investigation.

Visualising the Core Genome

Once the core genome has been estimated, the biologist is interested in analysing the behaviour of core genes with the greatest diversity. The evolutionary distance parameter is not directly observed and so the definition of the evolutionary distance depends on the choice of a statistical model describing the rates of substitution for each nucleotide. Once a model is fixed, the evolutionary distance for each gene is estimated from the sample, and is used as a proxy of diversity. The evolutionary distance is a function of the computed estimates of the substitution rates for each nucleotide.

The evolutionary distance is modelled using a Markov chain, in which the (a, b)-th entry in the transition matrix is the probability of substituting base a for base b. Here, (a, b) ranges over the possible choices, $\{A, C, T, G\}$, for the nucleotide base. Different models, defining different evolutionary distances, are derived by defining additional structure on the transition matrix. Jukes and Cantor [1969] assume all the substitution rates are equal. Further generalisations of the transition matrix with varying rates are derived by various authors and are summarised in Yang [2006].

For the purposes of this work, we utilise a Markov chain model of the evolutionary distance where the rates of nucleotide substitution are assumed to be identical. This is just one proxy of diversity, and others can be used. One can estimate the evolutionary distance, for a particular gene, between strain i and strain j, as

$$d_{i,j} = -\frac{3}{4} \log \left(1 - \frac{4}{3} p_{i,j} \right) \tag{3.33}$$

where $p_{i,j}$ is the number of points in the genome at which the nucleotide bases differ between the two strains, known as the Hamming [1950] distance between two codes. Since bacterial genomes are circular, we plot the median estimated evolutionary distance for each core gene on a circle in Figure 3.1 for each bacterium. We suggest that genes with large estimated evolutionary distances maybe under evolutionary selection, and/or are undergoing recombination. Genes that are highly conserved, those with an estimated evolutionary distance of

zero, are likely to be essential to the species, and possible variants that could occur may be lethal, and so are rarely sampled.

Some Biological Conclusions

Using the evolutionary distance between nucleotide sequences as a proxy for diversity, we identify genes for which the evolutionary distance is in the tail of the distribution of evolutionary distances for all core genes, and summarise some of the biological findings. An in-depth exposition of the biological significance can be found in the publication [van Tonder et al., 2014]. Since we are interested in identifying core genes exhibiting the greatest diversity, as measured by the estimated evolutionary distance, we select, for further study, genes for which the median distance is in the top 5% across all genes.

In *C. jejuni*, there are 43 estimated core genes whose evolutionary distance is above the 95% percentile. Of these, 11 are hypothetical proteins of unknown function. A hypothetical protein is a protein whose existence has been predicted using statistical techniques, but for which no experimental evidence exists yet. In addition 8 further genes are associated with coenzyme metabolism.

In particular, the gene with the significantly larger evolutionary distance is annotated in the reference genome as a putative hydrolase. Searches against a number of databases returned hits to sequences annotated as beta-lactamase, metallo-beta-lactamase, and thioredoxin reductase in other *C. jejuni* genomes. Beta-lactamases and metallo-beta-lactamases are enzymes that provide resistance to a wide range of broad-spectrum, beta-lactam antibiotics, while Thioredoxin reductase is essential for DNA replication, and has recently been proposed as a possible novel antimicrobial target site [Meyer et al., 2009, Lu et al., 2013]. The high level of sequence diversity observed implies that this gene is likely to occur under selective pressure, and/or is a site of frequent recombination.

In *N. meningitidis*, there are 37 estimated core genes whose evolutionary distance is above the 95% percentile where 9 encode unknown hypothetical proteins. The two core genes with

the largest evolutionary distances also encode putative proteins with unknown functions. One is adjacent to a gene proposed as a novel antimicrobial target site [Ren et al., 2005]. The other is a hypothetical protein that may play a role in stationary phase survival [Poulsen and Jensen, 1991].

Clusters of red lines in the *N. meningitidis* genome indicate regions of the genome with higher sequence diversity among the estimated core genes. One gene cluster encodes proteins in the NADH dehydrogenase complex. Authors have speculated that sequence differences in this gene region may be related to differences in fitness, and possibly virulence [Joseph et al., 2011, Kong et al., 2013].

This study exploited the large volume of publicly-available whole genome sequence data to outline a method for analysing bacterial genomes. In particular, a Bayesian decision model is proposed to estimate the core genome. The decision rule can be seen as a generalisation of the classical Frequentist decision rule used to estimate the core genome, to a Bayesian decision using the most conservative loss function that harshly penalises underestimation of the parameter. The Bayesian model utilised was basic, and more complicated models, such as the hierarchical Dirichlet models found in Hjort et al. [2010], could be employed. The resulting list of core genes can be explored in a biologically relevant manner.

3.3 Prediction as a Decision

Unobserved model parameters of interest can be divided into two categories, those that will be observed and those that will never be observed. Statistical decisions under uncertainty can be made in both cases. In the previous decision analysis, we made a decision as to whether a gene should be considered core, and thus warranting further study. We will never actually verify that a gene is a core gene in the strict sense (present in *all* strains).

When making predictions of parameters that will be observed, there are two types of predictions that can be made, point predictions and probabilistic predictions. We are interested

in determining which loss functions we should use to evaluate probabilistic predictions. The loss functions associated with probabilistic predictions are known as proper scoring rules. We provide a means to construct any number of proper scoring rules using convex functions. Proper scoring rules can be used to arrive at point estimates for unknown model parameters known as minimum score estimates. Proper scoring rules generalise classical estimates when the scoring rule is logarithmic. If the proper scoring rule is logarithmic, the minimum score estimate is equivalent to the maximum likelihood estimate.

3.3.1 Probabilistic Predictions

In the case of probabilistic prediction, the action space is a set of probability densities. As such, we understand $a \in \mathcal{A} \subset \mathcal{P}(\mathcal{Y})$ to be a probability density. A set of possible actions then forms a subset of the set of all probability densities on the measurement space. While the discussion is centred around the use of probability densities, we can, without any effect, replace probability densities with their associated probability distribution function.

When we considered the problem of point prediction, we began with a loss function and derived, as a function of the predictive density, the optimal point prediction. In that case, we began with a loss function and then found the optimal action. If we choose to fix a loss function, we should adopt the strategy of summarising our posterior distribution as the appropriate point estimate corresponding to the Bayes act.

If, however, we are interested in the entire predictive probability density, the loss functions described above do not suffice. We may, for example, be interested in particular density characteristics, such as expectation, the variance, and the median, just to name a few. In the case of nonlinear systems, the evolving predictive density could bear little or no resemblance to previous probabilistic predictions. In such cases, static distributional characteristics, such as the mean or variance, may not be useful. This motivates investigation into loss functions that are designed for probabilistic prediction. These loss functions are investigated in the next chapter and are known as *proper scoring rules*.

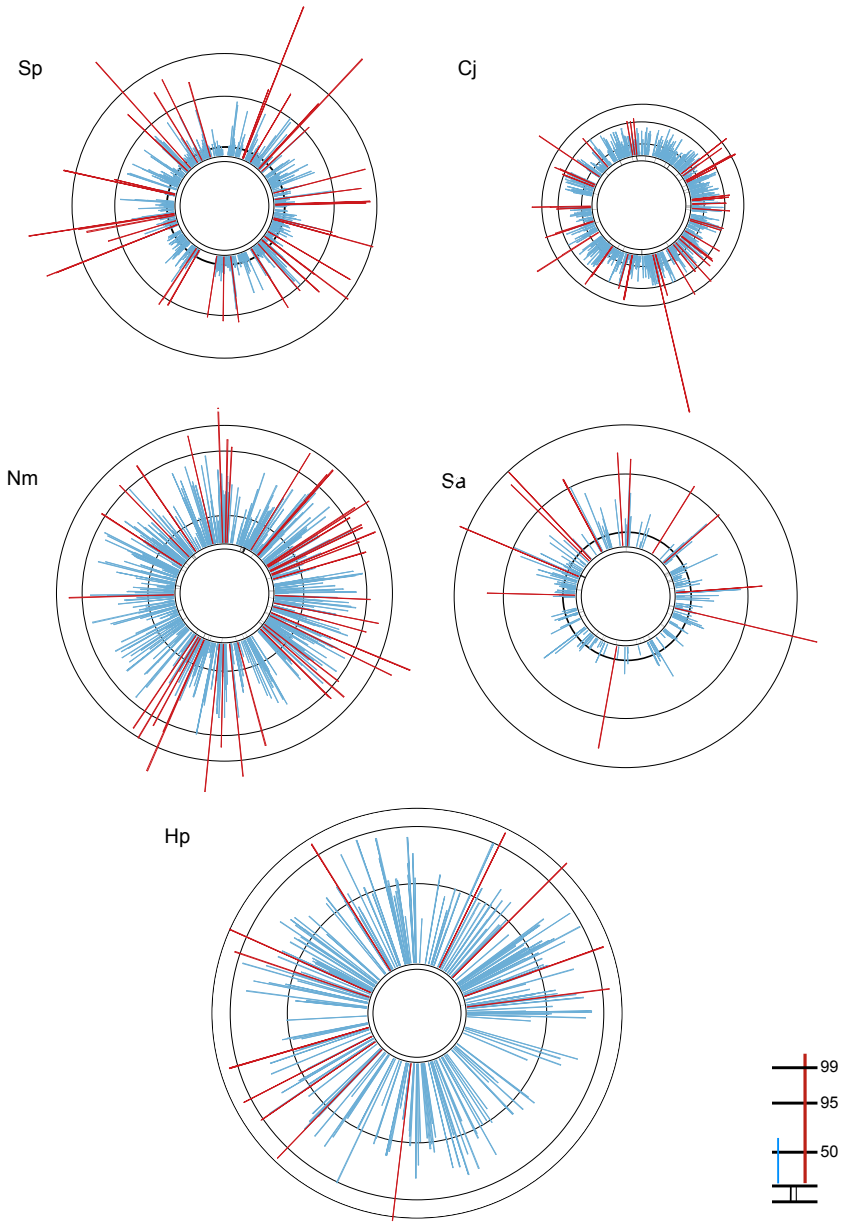


Figure 3.1: Median p-distance plotted against genome position for Sp) *S. pneumoniae*, Cj) *C. jejuni*, Nm) *N. meningitidis*, Sa) *S. aureus*, and Hp) *H. pylori*. The length of each line on a genome diagram was normalised using the maximum p-distance value for each bacterium. The outer rings mark the 99th, 95th, and 50th percentiles. Median p-distances above the 95th percentile are red, while all others are blue. The inner pair of rings marks the position of genes with a median p-distance of zero. Reproduced from van Tonder et al. [2014].

Chapter 4

Proper Scoring Rules

Proper scoring rules have been discussed by many authors. The logarithmic score was first introduced by Good [1952]. The quadratic score is considered in Brier [1950] and De Finetti [1970]. Use of proper scoring rules in weather forecasting was pioneered in Winkler [1969], Murphy [1973] and Murphy and Winkler [1984]. More recently, Gneiting and Raftery [2007] provided an overview of a large number of different proper scoring rules, and Jolliffe and Stephenson [2012] discuss practical applications.

In the previous chapter we investigated a problem in computational biology by analysing the behaviour of the loss function when the action undertaken is known and fixed. In this chapter we derive proper scoring rules in a similar manner in that proper scoring rules are loss functions associated with a decision problem in which the optimal action is fixed as the predictive probability. Framing prediction as a decision problem constitutes a novel approach to deriving proper scoring rules. While the results given in this chapter are all previously known, framing prediction as a decision problem allows for these results to be collected within the framework of decision theory offering an original perspective on a unified theory of model inference and evaluation.

We do not attempt to provide a comprehensive overview of all proper scoring rules. Instead, we begin by deriving proper scoring rules as probabilistic loss functions. We derive,

using the calculus of variations, the logarithmic score as the unique local proper score, where locality is a property of the proper scoring rule precisely defined later. We further utilise the calculus of variations to show that the continuous ranked probability score is a proper scoring rule. A general method of constructing proper scoring rules from convex functions is described. Proper scoring rules can also be used for parameter inference. The parameter estimate is chosen to minimise the observed empirical score. When the proper scoring rule is the logarithmic score and the statistical model employed models the measurements as conditionally independent and identically distributed, the classical maximum likelihood estimate is recovered. In the remainder of the chapter, we describe the limiting behaviour of the minimum score estimate for models with conditionally independent and identically distributed measurements..

4.1 Designing a Loss Function

A scientist, making a number of informed decisions, proposes a statistical model $\pi(y, \theta, x_{1:t}|M)$ which is applied to $\mathring{x}_{1:t}$, a sequence of measurements. For the purpose of model evaluation the scientist utilises the statistical model to make a prediction of $\boldsymbol{y}$, an unknown measurement. The fixed action is to assume that $\boldsymbol{y} \sim \pi(y|\mathring{x}_{1:t})$ and the problem remains to find a loss function for which this action is optimal.

In the case of probabilistic prediction the set of allowable actions is then the set of probabilities and so $a(\cdot)$ the action, is understood as a probability density. Let $\mathcal{S} : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ denote a loss function, where $\mathcal{A} = \mathcal{P}(\mathcal{Y})$ is the space of probabilities, and for which the optimal action is $\pi(\cdot|\mathring{x}_{1:t})$. That is, the expected loss is minimised at $a(\cdot) = \pi(\cdot|\mathring{x}_{1:t})$, satisfies

$$\frac{\partial}{\partial a}\mathbb{E}[\mathcal{S}(\boldsymbol{y}, a)]\bigg|_{a=\pi} = 0 \tag{4.1}$$

where $\pi := \pi(\cdot|\mathring{x}_{1:t})$ is the predictive density with which the expectation is taken.

After obtaining $\mathring{y}$, the actualising value of the measurement, we can perform an evaluation of $\pi(\cdot|\mathring{x}_{1:t})$ with $\mathring{y}$, the predictive probability density by evaluating $\mathcal{S}(\mathring{y}, \pi(\cdot|\mathring{x}_{1:t}))$. A loss

function for which the optimal action is the predictive density is known as a *proper scoring rule*. Proper scoring rules, and only proper scoring rules, are internally consistent in that the expected score assigns an optimal expected value to $\pi(\cdot|\mathring{x}_{1:t})$ if, and only if, $\boldsymbol{y}$ is distributed according to $\pi(\cdot|\mathring{x}_{1:t})$ [Bröcker and Smith, 2007].

Definition 5. *Let $\mathcal{A}$ be a set of probability densities. A proper scoring rule $\mathcal{S}$ is a loss function for which the Bayes action is the predictive distribution $\pi(\cdot|\mathring{x}_{1:t}, M)$. That is, for any probability density $a(\cdot) \in \mathcal{A}$, a proper scoring rule satisfies*

$$\int \mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t}, M))\pi(y|\mathring{x}_{1:t}, M)dy \leq \int \mathcal{S}(y, a(\cdot))\pi(y|\mathring{x}_{1:t}, M)dy. \tag{4.2}$$

or written equivalently as

$$\mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\mathring{x}_{1:t}))] \leq \mathbb{E}[\mathcal{S}(\boldsymbol{y}, a(\cdot))] \tag{4.3}$$

where $\boldsymbol{y} \sim \pi(y|\mathring{x}_{1:t})$.

The left-hand side of (4.2) depends only on the Bayes action, that is, the predictive density. As such the quantity is the Bayes risk, and it depends on the choice of proper scoring rule. Let

$$\mathcal{E}(\pi(\cdot|\mathring{x}_{1:t}, M)) = \int \mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t}))\pi(y|\mathring{x}_{1:t}, M)dy = \mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\mathring{x}_{1:t}))] \tag{4.4}$$

be known as the *prediction entropy*. Then a third related quantity, the *prediction divergence*, is related to the score and entropy by

$$\mathcal{D}(\pi(\cdot|\mathring{x}_{1:t}, M), a(\cdot)) = \mathbb{E}[\mathcal{S}(\boldsymbol{y}, a(\cdot))] - \mathcal{E}(\pi(\cdot|\mathring{x}_{1:t}, M)). \tag{4.5}$$

This is the change in risk when switching from the predictive density to an alternative density. The divergence function, by construction, is non-negative, and $a(y) = \pi(y|\mathring{x}_{1:t}, M)$ implies that the divergence is zero. The use of the terms *entropy* and *divergence* becomes clear when the proper scoring rule used is the logarithmic score, in that there is a relationship between *entropy* and *utility* [Jaynes and Bretthorst, 2003].

4.1.1 Honesty

Often, authors explain that a scoring rule must be *proper* in order to ensure that the forecaster provides an *honest* assessment of their subjective probability. In this manner, the forecaster is separated from the assessment process, and is asked to provide a subjective probability to describe the outcome of future measurements. Since the scoring rule is *proper*, the forecaster will incur a greater *risk* if they issue any probability other than their *honest* assessment.

We reject this explanation, since it forces us to consider situations in which the forecaster is purposefully deceitful. While this may be a problem in specific cases, in this work we do not distinguish between the forecaster and the assessor. Often, the forecaster and assessor are one and the same. By ensuring that our scoring rules are proper, we are not ensuring that we are honest, but instead ensuring that the assessment procedure is fixed, regardless of the predictive probability density and independently of the forecasting model. Hence, scoring rules are proper to ensure that we employ an internally consistent evaluation of predictions and measurements, rather than compelling the forecaster to be honest. As such there is no reason to distinguish between scoring rules that are proper and those that are not. All scoring rules are by construction proper.

4.1.2 Model Evaluation

When the actualising value of the measurement is available we can compute the real-valued score by evaluating a proper scoring rule at the actualising value of the measurement.

Definition 6. *Suppose a future measurement* $\boldsymbol{y}$*, forecasted using the predictive probability* $\pi(y|\mathring{x}_{1:t}, M)$*, is made. Then the score is*

$$\mathcal{S}(\mathring{y}, \pi(y|\mathring{x}_{1:t})) \tag{4.6}$$

where $\mathring{y}$ *is the actualising value of the measurement and* $\mathcal{S}$ *is some proper scoring rule.*

For a sequence of measurements we can evaluate the model as we assimilate the measurements. Given a sequence of measurements $\mathring{x}_{1:t}$, we can, at an intermediate time i, fit the model using the previous $i-1$ measurements and evaluate the score using $\mathring{x}_i$, the observed value of the measurement. Summing over all i yields a quantity known as the *prequential score* [Dawid, 1984].

Definition 7. *Let S be a proper scoring rule and consider a sequence of predictions such that, for each i, $\boldsymbol{x}_i \sim \pi(x_i|\mathring{x}_{1:i-1}, M)$. Then*

$$\sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i|\mathring{x}_{1:i-1}, M)), \tag{4.7}$$

is known as the prequential score.

Computing the prequential score has a number of immediate benefits. For example, it does not use the same measurements for both model fitting and scoring. Additionally the prequential score can be updated sequentially as more measurements are made available.

4.1.3 Parameter Estimation

Proper scoring rules can be utilised for inferences in which estimates are made by minimising the prequential score. Such an estimate is known as the minimum score estimate.

Definition 8. *Let the predictive density be in some parametric family so that $\pi(x_i|\mathring{x}_{1:i-1}, M) \cong \pi(x_i|\theta)$. Then the value of θ that minimises the prequential score,*

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i|\theta)), \tag{4.8}$$

is known as the minimum score estimate.

This definition allows for the possibility that for each i, the densities $\pi(x_i|\theta)$ are not necessarily identical. When, for every i, the densities $\pi(x_i|\theta)$ are identical, it will be shown that the minimum score estimate is asymptotically Normal, and the limiting distribution is derived.

4.2 Some Proper Scoring Rules

A proper scoring rule is a probabilistic loss function that evaluates the predictive density using the actualising value of the measurement. In this chapter, we consider two popular continuous proper scoring rules: the *logarithmic scoring rule* and the *continuous ranked probability score*. To derive the logarithmic scoring rule, we seek proper scoring rules that are *local* meaning that the score only involves the predictive density in the neighbourhood of the actualising value of the measurement. Furthermore, we use the calculus of variations to show that the continuous ranked probability score is a proper scoring rule.

4.2.1 Logarithmic Scoring Rule

Bernardo [1979] and Dawid [2007] derive the general structure of the decision problem whose proper scoring rule is determined, up to some constant, by only the value of the density π at the measurement y. Bernardo [1979] shows that the logarithmic score is the only local score of this type. Furthermore, Dawid [2007] derives the associated entropy and divergence. The logarithmic score is

$$\mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t})) = -\log \pi(y|\mathring{x}_{1:t}). \tag{4.9}$$

In the case of the logarithmic score, the prediction entropy is

$$\mathcal{E}(\pi(\cdot|\mathring{x}_{1:t})) = -\int \pi(y|\mathring{x}_{1:t}) \log \pi(y|\mathring{x}_{1:t}) dy. \tag{4.10}$$

This is known as the Shannon [1948] Entropy. The associated prediction divergence is

$$\mathcal{D}(\pi(\cdot|\mathring{x}_{1:t}), a(\cdot)) = \int \log\left(\frac{\pi(y|\mathring{x}_{1:t})}{a(y)}\right) \pi(y|\mathring{x}_{1:t}) dy. \tag{4.11}$$

This is the Kullback and Leibler [1951] divergence between the densities π and a. Notice that the logarithmic score is a function only of the probability density $\pi(\cdot|\mathring{x}_{1:t})$, evaluated at the point y. When this is the case a proper scoring rule is known as local.

Definition 9. *A proper scoring rule is local if it only depends on the density evaluated at the actualising value of the measurement*

$$\mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t})) \equiv \mathcal{S}(\pi(y|\mathring{x}_{1:t})). \tag{4.12}$$

This allows us to employ a variational argument to solve for local proper scoring rules. It turns out that the logarithmic scoring rule is the only local proper scoring rule up to an equivalent decision problem.

Theorem 1. *If $\mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t}))$ is a local proper scoring rule, and $\pi(y|\mathring{x}_{1:t})$ is sufficiently differentiable, then the proper scoring rule is the logarithmic scoring rule.*

Proof. The Bayes action is the predictive density and so, by definition, is a stationary (critical) point of the expected loss, $\mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\mathring{x}_{1:t}))] = \mathbb{E}[\mathcal{S}(\pi(\boldsymbol{y}|\mathring{x}_{1:t}))]$. We now arrive at a minimisation problem. That is, we find

$$\operatorname*{argmin}_{a(\cdot)} \mathbb{E}[\mathcal{S}(a(\boldsymbol{y}))] = \operatorname*{argmin}_{a(\cdot)} \int \mathcal{S}(a(y))\pi(y|\mathring{x}_{1:t})dy \tag{4.13}$$

subject to the condition that $a(\cdot)$ is a probability density. Since $a(\cdot)$ is a probability density, we have the integral condition $\int a(y)dy = 1$. Then we use Lagrange multipliers to find the stationary points of the Lagrangian

$$\mathcal{J}(a(\cdot)) = \int \mathcal{S}(a(y))\pi(y|\mathring{x}_{1:t})dy + C\left(\int a(y)dy - 1\right) \tag{4.14}$$

where C is the Lagrange multiplier. If $a(\cdot)$ is a stationary point of (4.14) then we must have, for $\epsilon > 0$ and $\rho(\cdot)$, an arbitrary nonzero function $\partial_\epsilon \mathcal{J}(a + \epsilon\rho)|_{\epsilon=0} = 0$. Now write

$$\mathcal{J}(a + \epsilon\rho) = \int \mathcal{S}(a + \epsilon\rho)\pi(y|\mathring{x}_{1:t})dy + C\left(\int (a(y) + \epsilon\rho(y))dy - 1\right). \tag{4.15}$$

Then, an application of the chain rule evaluated at $\epsilon = 0$ yields

$$0 = \frac{\partial}{\partial\epsilon}\mathcal{J}(a + \epsilon\rho)\bigg|_{\epsilon=0} = \int \left(\frac{\partial\mathcal{S}(a(y))}{\partial a}\pi(y|\mathring{x}_{1:t})dy + C\right)\rho(y)dy. \tag{4.16}$$

For the right-hand side to be zero, the integrand must vanish as $\rho(\cdot)$ is arbitrary. Furthermore, since a is assumed to be a stationary point, we must solve

$$\frac{\partial\mathcal{S}(\pi(y|\mathring{x}_{1:t}))}{\partial\pi}\pi(y|\mathring{x}_{1:t}) + C = 0. \tag{4.17}$$

Rearranging and integrating yields

$$\mathcal{S}(\pi(y|\mathring{x}_{1:t})) = -C \log \pi(y|\mathring{x}_{1:t}) + k(y) \tag{4.18}$$

where $k(y)$ is an arbitrary function depending only on y, the future measurement. Since any choice of C and $k(y)$ leads to the same decision problem, we can choose $C = 1$ and $k(y) \equiv 0$ and define a set of equivalent decision problems $\mathcal{S}^*$ by varying C and $k(y)$. □

The logarithmic scoring rule is, by far, the most employed proper scoring rule. If one uses the logarithmic scoring rule, classical quantities used for model selection can be recovered.

4.2.2 Continuous Ranked Probability Score

If we look for proper scoring rules that depend on more than just the value of the density, we can arrive at a large number of possible choices. An example of such a proper scoring rule is the continuous ranked probability score. Let

$$\Pi(y|\mathring{x}_{1:t}) = \int_{-\infty}^{y} \pi(y'|\mathring{x}_{1:t}) dy' \tag{4.19}$$

be the predictive cumulative distribution function for the predictive density. Then the continuous ranked probability score is

$$\mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t})) = \int \left(\Pi(y'|\mathring{x}_{1:t}) - \mathbb{H}(y' - y)\right)^2 dy'. \tag{4.20}$$

If $a(\cdot)$ is an arbitrary probability density, denote the corresponding cumulative distribution function by $A(\cdot)$. The corresponding prediction entropy and divergence functions are given by

$$\mathcal{E}(\pi(\cdot|\mathring{x}_{1:t})) = -\frac{1}{2} \int \Pi(y|\mathring{x}_{1:t})(1 - \Pi(y|\mathring{x}_{1:t})) dy \tag{4.21}$$

and

$$\mathcal{D}(\pi(\cdot|\mathring{x}_{1:t}), a(\cdot)) = \frac{1}{2} \int (\Pi(y|\mathring{x}_{1:t}) - A(y))^2 dy. \tag{4.22}$$

The next theorem establishes that (4.20) does define a proper scoring rule.

Theorem 2. *The continuous ranked probability score (4.20) is a proper scoring rule.*

Proof. Let $\pi(\cdot|\mathring{x}_{1:t})$ be the predictive density. The expected score, or risk, associated to using the cumulative distribution $A(\cdot)$ as the predictive distribution is

$$\mathbb{E}[\mathcal{S}(\boldsymbol{y}, A(\cdot))] = \int \left(\frac{1}{2} \int (A(y') - \mathbb{H}(y' - y))^2 dy' \right) \pi(y|\mathring{x}_{1:t}) dy. \tag{4.23}$$

So, for an arbitrary variation P and $\epsilon > 0$ the risk is

$$\mathbb{E}[\mathcal{S}(\boldsymbol{y}, A + \epsilon P)] = \frac{1}{2} \int \int (A(y') + \epsilon P(y') - \mathbb{H}(y' - y))^2 \pi(y|\mathring{x}_{1:t}) dy' dy. \tag{4.24}$$

The derivative of the expected score with respect to ϵ, setting $\epsilon = 0$, and equating it to zero, is given by

$$\left. \frac{\partial \mathbb{E}[\mathcal{S}(\boldsymbol{y}, A + \epsilon P)]}{\partial \epsilon} \right|_{\epsilon=0} = \int \left(\int (A(y') - \mathbb{H}(y' - y)) \pi(y|\mathring{x}_{1:t}) dy \right) P(y') dy' = 0. \tag{4.25}$$

Solving the above equation implies that

$$\int (A(y') - \mathbb{H}(y' - y)) \pi(y|\mathring{x}_{1:t}) dy = 0, \tag{4.26}$$

as the variation $P(\cdot)$ is arbitrary. It follows that

$$A(y) = \int \mathbb{H}(y - z) \pi(z) dz = \int_{-\infty}^{y} \pi(y|\mathring{x}_{1:t}) dz = \Pi(y|\mathring{x}_{1:t}). \tag{4.27}$$

Therefore, the expected score is minimised by the cumulative distribution function associated to the predictive density. □

4.2.3 Constructing Scoring Rules

While the logarithmic score can be derived by looking for all local proper scoring rules, a general construction of arbitrary proper scoring rules is possible. An arbitrary proper scoring rule is found by proposing a general divergence function. In particular, the general divergence function used is known as the Bregman divergence. Bregman [1967] shows that the proposed divergence function satisfies the general properties of a prediction divergence function. Specifically, $\mathcal{D}(a, \pi) \geq 0$ and $\mathcal{D}(a, \pi) = 0 \iff a = \pi$.

In this section, we use an arbitrary convex function to construct a proper scoring rule. The proof takes advantage of the properties of convex functions on convex sets. A more general version of this proof, given by Dawid [2007], allows for general topological considerations. When working with real-valued variables, such generality is not required.

Theorem 3. *Let $\mathcal{P}$ be the (convex) space of all probability densities. Let ψ be a (sufficiently differentiable) convex function on $\mathcal{P}$. Then, we can construct a proper scoring rule*

$$\mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t})) = -\nabla\psi\big(\pi(y)\big) - \int \big(\psi\big(\pi(z)\big) - \pi(z)\nabla\psi\big(\pi(z)\big)\big)\, dz. \tag{4.28}$$

Proof. Since the space of probability densities is a convex set, and, by assumption, ψ is a convex function, first-order conditions imply, for every $a, \pi \in \mathcal{P}$

$$\psi(a) - \psi(\pi) - \nabla\psi(\pi)(a - \pi) \geq 0, \tag{4.29}$$

where we understand

$$\nabla\psi(\pi) = \left.\frac{\partial\psi(a)}{\partial a}\right|_{a=\pi}. \tag{4.30}$$

Since (4.29) holds for the functions a and π, it must also hold for each point. We can then define a divergence function as

$$\mathcal{D}(a(\cdot), \pi(\cdot)) := \int \psi\big(a(z)\big) - \psi\big(\pi(z)\big) - \nabla\psi\big(\pi(z)\big)\big(a(z) - \pi(z)\big)dz \geq 0. \tag{4.31}$$

This function is known as the Bregman divergence. Note that this function satisfies the properties of the prediction divergence, $\mathcal{D}(a(\cdot), \pi(\cdot)) \geq 0$ and $\mathcal{D}(\pi(\cdot), \pi(\cdot)) = 0$, and so the proposed divergence function satisfies the properties of a prediction divergence function. To find the associated proper scoring rule, we evaluate the divergence at the actualising value of the measurement by setting $a = \delta(z - y)$, so that

$$\begin{aligned}
\mathcal{S}(y, \pi(\cdot)) &= \mathcal{D}(\delta(z - y), \pi(z)) && (4.32)\\
&= -\int \nabla\psi(\pi(z))(\delta(z - y) - \pi(z)) + \psi(\delta(z - y)) - \psi(\pi(z))dz && (4.33)\\
&= -\nabla\psi\big(\pi(y)\big) - \int \big(\psi\big(\pi(z)\big) - \pi(z)\nabla\psi\big(\pi(z)\big)\big)\, dz + \int \psi(\delta(z - y))dz. && (4.34)
\end{aligned}$$

The last integral on the right-hand side is a function depending only on y, the measurement. As such, we can assume, without loss of generality that this function is zero. If this is not the case, we can modify the proper scoring rule by the non-zero function and arrive at an equivalent proper scoring rule. □

Note that the same proof can also apply to distribution functions, which also form a convex set. The logarithmic score can be recovered by setting $\psi(a) = a \log a$, while the Brier score is recovered by setting $\psi(a) = a^2$, assuming the underlying space consists of two elements. Using this construction, any number of proper scoring rules can be obtained by considering different convex functions. As an interesting example set $\psi(a) = a^\alpha$. Then, we find that

$$\mathcal{S}(y, \pi(\cdot)) = (\alpha - 1) \int \pi(z)^\alpha dz - \alpha \pi(y)^{\alpha - 1}, \tag{4.35}$$

which is known as the Tsallis score, since the associated entropy function is known as the Tsallis et al. [2009] entropy. This construction also yields another motivation for looking for local proper scoring rules by requiring the that integral term in (4.28) is zero. Solving $\nabla \psi(a) a - \psi(a) = 0$ yields $\psi(a) = a \log a$.

4.3 Inference with Proper Scoring Rules

In this section, we investigate the behaviour of minimum score estimates. The asymptotic behaviour of such estimates is described by a limiting Normal distribution. While the purpose of developing proper scoring rules is for evaluation, they are also suitable for estimation. Proper scoring rules can be used to estimate parameters by choosing, as the estimate, the value of θ that minimises the observed prequential score. Using a parametric family of probability densities as our predictive density, we set $\pi(x_i | \mathring{x}_{1:i-1}, M) \cong \pi(x_i | \theta)$. Then, recalling Definition 8,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i | \theta)) \tag{4.36}$$

is the minimum score estimate. Minimum score estimates (4.36) are a type of M-estimators. An M-estimate is the solution to the problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} f(x_i, \theta). \tag{4.37}$$

for some function f. The conditions under which a limiting Normal behaviour of such estimates is possible has been described by Huber [2011]. Under these conditions, the details of convergence can be found in standard statistical texts, such as Rohatgi and Saleh [2011], Casella and Berger [2002] and Dudley [2002]. Under particular conditions, the minimum score estimate is equivalent to the classical maximum likelihood estimate. These conditions are discussed below.

The Logarithmic Score and the Maximum Likelihood Estimate

When employing the logarithmic score the minimum score estimate is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left(-\sum_{i=1}^{t} \log \pi(\mathring{x}_i|\theta) \right) = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^{t} \pi(\mathring{x}_i|\theta) \tag{4.38}$$

If the model is formulated with conditionally independent and identically distributed measurements (Structure 1) the likelihood function decomposes as

$$\pi(x_{1:t}|\theta) = \prod_{i=1}^{t} \pi(x_i|\theta). \tag{4.39}$$

where, for each i, the densities $\pi(x_i|\theta)$ are identical. Then the right-hand side of (4.38) is (2.40), the maximum likelihood estimate.

This class of models is by no means exhaustive. For example, one might consider a model in which measurements are made on a grid. Each measurement might be modelled conditional on the model parameters as well as the surrounding measurements by specifying $\pi(x_{i,j}|\theta, x_{k,m})$ where the pair (k, m) ranges across the neighbours of (i, j) on the grid.

4.3.1 Perfect Model Scenario

We accept that the model we construct is an approximation of reality. We can, however, analyse the behaviour of our model in the *perfect model scenario*.

Definition 10. *In the* perfect model scenario*, the measurements are known (by construction) to be generated using a set of equations, either probabilistic, deterministic, or a combination.*

When we apply our model to a set of measurements obtained from reality, we say that we are in the *imperfect model scenario*.

For models with conditionally independent and identically distributed measurements the behaviour of the minimum score estimate in the perfect model scenario can be analysed using the law of large numbers. Models of this form are families of probability densities $\pi(\cdot|\theta)$ indexed by θ.

In the first instance, consider the case in which the perfect model is, for some fixed $\tilde{\theta}$, $\pi(\cdot|\tilde{\theta})$. That is, within the perfect model scenario we have, for all i, $\boldsymbol{x}_i \sim \pi(\cdot|\tilde{\theta})$ and the law of large numbers can then be applied.

Proposition 7.

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} \mathcal{S}(\boldsymbol{x}_i, \pi(\cdot|\theta)) \to \tilde{\theta} \text{ as } t \to \infty \tag{4.40}$$

where convergence is in probability.

Proof. Under particular regularity conditions on both S and π, the minimum score estimate is an M-estimator and so $\hat{\theta}$ delivers an unbiased and consistent estimate of $\tilde{\theta}$ [Huber, 2011, Hampel et al., 2011]. □

In the second case consider the case where the perfect model is a probability density function $\tau(\cdot)$ from which the measurements form an independent and identically distributed random sample. In contrast to the first case, we assume that $\tau(\cdot) \neq \pi(\cdot|\theta)$ for any θ. Due to the model assumption of conditionally independent and identically distributed measurements the law of large numbers still applies but the notion of a true parameter value no longer

makes sense. Instead we say that the minimum score estimate converges to the pseudo-true parameter state in the limit of an infinite number of measurements. The limiting value of the minimum score estimate, or the pseudo-true parameter will vary with the choice of proper scoring rule.

Many statistical models do not make the assumption of conditionally independent and identically distributed measurements. For example, consider a state space model (Model 6) where, for each i the densities $\pi(x_i|\theta)$ are not identical. The law of large numbers does not apply and the estimate does not converge. Even with a perfect model of a deterministic chaotic system with an infinity of measurements the measurement noise makes the identification of the exact state impossible [Judd and Smith, 2001]. The value of the minimum score estimate will depend on decisions made with respect to the experimental design such as choice of score rule, or forecast lead time.

For statistical models with conditionally independent and identically distributed measurements the large sample behaviour of the minimum score estimate can be described. The conditions under which a limiting Normal behaviour of such estimates is possible has been described by Huber [2011]. In the next section the large sample behaviour of the minimum score estimate for models with conditionally independent and identically distributed measurements is described. This result is necessary for the discussion in Chapter 5 on classical quantities of model selection.

4.3.2 Large Sample Theory

In this section, we derive the limiting Normal behaviour of the minimum score estimate for statistical models exhibiting conditionally independent and identically distributed measurements. If, for each i, the predictive density is given by the parametric family, $\pi(x_i|\theta)$, then the minimum score estimate satisfies

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(\cdot|\theta)). \tag{4.41}$$

The minimum score estimate is defined by the choice of proper scoring rule. The estimator (4.41) is of the type of an M-estimator and, hence, a limiting Normal distribution of the minimum score estimate can be derived analogously. Since the minimum score estimator is a type of M-estimator, the estimate is consistent and unbiased [Huber, 2011]. If the M-estimator is suitably differentiable we can define the estimating equation as

$$\Psi_t(\theta) = \sum_{i=1}^{t} \frac{\partial \mathcal{S}(\mathring{x}_i, \pi(\cdot|\theta))}{\partial \theta}. \tag{4.42}$$

The minimum score estimate $\hat{\theta}$ is a solution to the equation $\Psi_t(\theta) = 0$. Define the following associated quantities as $J_t(\theta) = \mathbb{V}[\Psi_t(\theta)]$ and $H_t(\theta) = -\mathbb{E}\left[\nabla_\theta \Psi_t(\theta)\right]$, where the expectation is taken with respect to the true density of the measurements.

For some arbitrary measurement, say y, the risk of using $\pi(\cdot|\theta)$, for some θ, is given by

$$\mathcal{R}(\theta) = \int \mathcal{S}(y, \pi(\cdot|\theta))\tau(y)dy. \tag{4.43}$$

The risk, as a function of θ, is assumed to be minimised a some point $\tilde{\theta}$ known as the pseudo true parameter. Define, for y, an arbitrary measurement,

$$\Psi(\theta) = \frac{\partial}{\partial \theta}\mathcal{S}(y, \pi(\cdot|\theta)). \tag{4.44}$$

Further define $J(\theta) = \mathbb{V}[\Psi(\theta)]$ and $H(\theta) = -\mathbb{E}[\nabla_\theta \Psi(\theta)]$. This allows for a description of the limiting Normal distribution of the minimum score estimate. The precise details of convergence can be found in statistical texts such as Rohatgi and Saleh [2011], Casella and Berger [2002] and Dudley [2002]. Before we derive the limiting Normal distribution we require the following lemma.

Lemma 1.

$$\frac{1}{\sqrt{t}}\Psi_t(\tilde{\theta}) \to \boldsymbol{N}(0, J(\tilde{\theta})). \tag{4.45}$$

where convergence is in distribution.

Proof. By the law of large numbers, for fixed $\tilde{\theta}$, we write

$$\frac{1}{t}\Psi_t(\tilde{\theta}) \to \mathbb{E}[\Psi(\tilde{\theta})] \tag{4.46}$$

where, assuming that the interchange of differentiation and integration is allowed we can write

$$\mathbb{E}[\Psi(\tilde{\theta})] = \int \left(\frac{\partial}{\partial\theta}\mathcal{S}(y,\pi(\cdot|\theta))\right)\Big|_{\theta=\tilde{\theta}} \tau(y)dy = \left(\frac{\partial}{\partial\theta}\int \mathcal{S}(y,\pi(\cdot|\theta))\tau(y)dy\right)\Big|_{\theta=\tilde{\theta}} = 0 \quad (4.47)$$

So, $\Psi(\tilde{\theta})$ is a zero mean random variable with variance $\mathbb{V}[\Psi(\tilde{\theta})] = J(\tilde{\theta})$ and the result follows from the central limit theorem. □

Theorem 4. *Under suitable regularity conditions, the minimum score estimate $\hat{\boldsymbol{\theta}}$, is asymptotically Normal.*

$$\sqrt{t}(\hat{\boldsymbol{\theta}} - \tilde{\theta}) \sim \boldsymbol{N}(0, H^{-1}JH^{-T}) \quad (4.48)$$

where $J = \mathbb{V}[\Psi(\tilde{\theta})]$ and $H = -\mathbb{E}[\nabla_\theta\Psi(\tilde{\theta})]$ and $\boldsymbol{N}$ is the Normal distribution.

Proof. Let $\hat{\boldsymbol{\theta}}$ be the minimum score estimate. Write the Taylor expansion of $\Psi_t(\hat{\boldsymbol{\theta}})$ around $\tilde{\theta}$, the estimate, as

$$\Psi_t(\hat{\boldsymbol{\theta}}) = \Psi_t(\tilde{\theta}) + \nabla_\theta\Psi_t(\tilde{\theta})(\hat{\boldsymbol{\theta}} - \tilde{\theta}) + \dots \ . \quad (4.49)$$

Rearranging and noting that, $\Psi_t(\hat{\boldsymbol{\theta}}) = 0$, so that we can write

$$\sqrt{t}(\hat{\boldsymbol{\theta}} - \tilde{\theta}) \approx -\left(\frac{1}{t}\nabla_\theta\Psi(\tilde{\theta})\right)^{-1}\left(\frac{1}{\sqrt{t}}\Psi(\tilde{\theta})\right). \quad (4.50)$$

Using Lemma 1, the last term on the right-hand side converges to the Normal distribution $\boldsymbol{N}(0, J(\tilde{\theta}))$. The following average, using the law of large numbers, converges, in probability, to its expectation

$$-\frac{1}{t}\nabla_\theta\Psi_t(\tilde{\theta}) \to H(\tilde{\theta}). \quad (4.51)$$

Combining these results yields

$$H(\tilde{\theta})^{-1}\mathbf{z} \text{ where } \mathbf{z} \sim \boldsymbol{N}(0, J(\tilde{\theta})), \quad (4.52)$$

from which is follows that

$$\sqrt{t}(\hat{\boldsymbol{\theta}} - \tilde{\theta}) \to \boldsymbol{N}(0, H^{-1}JH^{-T}). \quad (4.53)$$

□

In the case of the logarithmic score, for conditionally independent measurements, we have

$$\Psi(\theta) = \sum_{i=1}^{t} \frac{\partial \log \pi(\mathring{x}_i|\theta)}{\partial \theta} \text{ and } \nabla_\theta \Psi(\theta) = -\sum_{i=1}^{t} \frac{\partial^2 \log \pi(\mathring{x}_i|\theta)}{\partial \theta^2}. \tag{4.54}$$

If the measurement noise is modelled as independent and identically distributed, then $\Psi(\theta)$ is the score function, $t^{-1}\nabla_\theta \Psi(\tilde{\theta})$ is the empirical Fisher information, and the minimum score estimate is equivalent to the maximum likelihood estimate. The following identity (4.55) holds in the perfect model scenario.

Proposition 8. *In the perfect model scenario, the true density $\tau(\cdot) := \pi(\cdot|\tilde{\theta})$, is known to be a member of the parametric family of approximating densities. Then, the identity*

$$0 = \mathbb{E}[\Psi(\tilde{\theta})\Psi(\tilde{\theta})^T] + \mathbb{E}[\nabla_\theta \Psi(\tilde{\theta})] \tag{4.55}$$

holds for the logarithmic score.

Proof. The following relationship, for an arbitrary measurement x, follows by the chain rule

$$\frac{\partial \log \pi(x|\theta)}{\partial \theta} = \frac{1}{\pi(x|\theta)} \frac{\partial \pi(x|\theta)}{\partial \theta} \implies \frac{\partial \pi(x|\theta)}{\partial \theta} = \pi(x|\theta) \frac{\partial \log \pi(x|\theta)}{\partial \theta}. \tag{4.56}$$

Then, since $\pi(x|\theta)$ is a probability density, we have

$$\int \pi(x|\theta) dx = 1 \implies \int \frac{\partial \pi(x|\theta)}{\partial \theta} dx = 0. \tag{4.57}$$

Using (4.56), substitute to obtain

$$0 = \int \frac{\partial \pi(x|\theta)}{\partial \theta} dx = \int \frac{\partial \log \pi(x|\theta)}{\partial \theta} \pi(x|\theta) dx. \tag{4.58}$$

Taking a derivative of the right-hand side yields

$$0 = \int \frac{\partial^2 \log \pi(x|\theta)}{\partial \theta^2} \pi(x|\theta) dx + \int \frac{\partial \pi(x|\theta)}{\partial \theta} \frac{\partial \log \pi(x|\theta)}{\partial \theta} dx \tag{4.59}$$

$$= \int \frac{\partial^2 \log \pi(x|\theta)}{\partial \theta^2} \pi(x|\theta) dx + \int \left(\frac{\partial \log \pi(x|\theta)}{\partial \theta} \right) \left(\frac{\partial \log \pi(x|\theta)}{\partial \theta} \right)^T \pi(x|\theta) dx. \tag{4.60}$$

Evaluating the above expression at $\theta = \tilde{\theta}$, in the perfect model scenario $\boldsymbol{x}_i \sim \pi(x|\tilde{\theta})$, we have

$$0 = \mathbb{E}\left[\left(\frac{\partial \log \pi(x|\tilde{\theta})}{\partial \theta} \right) \left(\frac{\partial \log \pi(x|\tilde{\theta})}{\partial \theta} \right)^T \right] + \mathbb{E}\left[\frac{\partial^2 \log \pi(x|\tilde{\theta})}{\partial \theta^2} \right]. \tag{4.61}$$

Forming the average of the above quantity over the t measurements, and taking advantage of the linearity of the expectation operator, yields (4.56), the result. □

It follows that the asymptotic variance of the minimum score estimate for the logarithmic score reduces to the Fisher information, when the true density is a member of the parametric family of approximating densities. This identity does not hold for an arbitrary proper scoring rule. One might be inclined to search for proper scoring rules that do satisfy this identity but outside of the perfect model scenario the identity will never hold, and so in practice finding such proper scoring rules may not be a useful endeavour. In the imperfect model scenario, the matrix $H^{-1}JH^{-T}$ is the Godambe [1976] information matrix, generalised to an arbitrary proper scoring rule.

Estimation Problems

Corollary 5. *Assuming that we have consistent estimates of the matrices so that*

$$\hat{J}_t \to J \text{ and } \hat{H}_t \to H, \tag{4.62}$$

where convergence is in distribution. Then, we have, for large t, $\hat{\boldsymbol{\theta}} \sim \boldsymbol{N}(\cdot|\tilde{\theta}, H_t^{-1}J_tH_t^{-T})$.

Proof. Using the asymptotic distribution of the minimum score estimate, assuming the estimates of the matrices to be consistent, and applying Slutsky's theorem [Rohatgi and Saleh, 2009], yields the result

$$\sqrt{t}(\hat{H}_t^{-1}\hat{J}_t\hat{H}_t^{-T})^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \tilde{\theta}) \to \boldsymbol{N}(0, I). \tag{4.63}$$

□

Obtaining an unbiased estimate of the risk can be difficult in practice. It relies on the computation of the matrices $\hat{J}_t$ and $\hat{H}_t^{-1}$. The estimate $\hat{H}_t^{-1} = \nabla_\theta^2\Psi_t(\hat{\theta})^{-1}$ is easier to compute. A natural estimate of J is the outer product $\hat{J}_t = \Psi_t(\hat{\theta})\Psi_t(\hat{\theta})^T$. Unfortunately, by definition, $\Psi_t(\hat{\theta}) = 0$. If independent repetitions of the measurements are available, J

can be estimated by the sample variance of the associated individual contributions to the score. In more realistic conditions, we observe only a single replicate of the measurements, as in the case of time series. To avoid this problem, Heagerty and Lumley [2000] suggest window subsampling methods that introduce the assumption that measurements, sufficiently far apart, either spatially or temporally, are uncorrelated. This, however, introduces an ad hoc manner in which the measurements are separated.

In this chapter we employ a decision theoretic argument to introduce proper scoring rules as the loss functions required for evaluating predictive performance. The logarithmic and continuous ranked probability score are analysed using the calculus of variations and a method for constructing an arbitrary number of proper scoring rules is given. The chapter continues with a discussion of the use of proper scoring rules for parameter inference, particularly the behaviour of the minimum score estimate for models with conditionally independent and identically distributed measurements for large sample sizes. In the next chapter we analyse classical quantities of model selection within the proper scoring rule framework.

Chapter 5

Evaluating Predictions

The purpose of this chapter is to explore the use of the prequential score (Definition 7) for model selection, and describe its relationship with classical model selection criteria. In particular the Bayesian method of model selection (Section 5.1) and the Frequentist method of risk estimation (Section 5.2) are analysed within the proper scoring rule framework. The assumptions under which the Bayesian and Frequentist approaches to model selection are criticised within the proper scoring rule framework in Section 5.1.2 and Section 5.2.2. This chapter constitutes a novel application of proper scoring rules and a novel interpretation of classical model selection quantities (Theorem 8).

In this chapter, we consider the problem of evaluating the predictive performance of a statistical model. For a collection of measurements $\mathring{x}_{1:t}$, consider a model $M \in \mathcal{M}$, where $\mathcal{M}$ is a finite collection of statistical models. Each statistical model M is written using conditional probability densities to formulate a within-model data likelihood function $\pi(x_{1:t}|\theta, M)$, where the $x_{1:t}$ are the observations, and θ are the latent model parameters. Bayesian methods are characterised by the explicit use of a within-model prior density $\pi(\theta|M)$, so that the within-model posterior $\pi(\theta|\mathring{x}_{1:t}, M) \propto \pi(\mathring{x}_{1:t}|\theta, M)\pi(\theta|M)$ is obtained using Bayes' theorem. For different models the corresponding parameters need not be the same.

Definition 11. *The model*

$$\hat{M} = \underset{M}{\operatorname{argmin}} \sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i|\mathring{x}_{1:i-1}, M)), \tag{5.1}$$

with the smallest prequential score, is defined to be the optimal model.

If we employ, for every i, a parametric predictive density $\pi(x_i|\theta)$ then for each i, the predictive densities are

$$\pi(x_i|\mathring{x}_{1:i-1}, M) \cong \pi(x_i|\hat{\theta}_i) \text{ where } \hat{\theta}_i = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{i-1} \mathcal{S}(\mathring{x}_i, \pi(x_i|\theta)). \tag{5.2}$$

That is, at every time step i, we employ all previous measurements to estimate $\hat{\theta}_i$. As such, the parameter must be estimated again at every time i, in order to compute the prequential score.

This model selection criterion is consistent, in the sense that if the perfect model is contained within the collection of models, the prequential model selection criterion will, asymptotically, select the correct model [Dawid and Vovk, 1999]. This behaviour follows from the definition of proper scoring rules.

Classical model selection criteria are divided into two types, *Bayesian* and *Frequentist*. The Frequentist criteria estimate the risk, or expected score, of a future, unknown measurement. The Bayesian criteria are characterised by placing a prior on the set of models and selecting the model with the largest posterior probability. We proceed by first considering the Bayesian approach to model selection, characterised by use of the marginal likelihood for the measurements. Afterwards, we consider the Frequentist approach to model selection, characterised by use of the prequential score to estimate the risk in predicting a future measurement.

5.1 Bayesian Model Selection

The Bayesian approach to model selection is characterised by placement of a prior on the collection of models. An application of Bayes' theorem allows one to compute the posterior

probability of each model. The model with the largest posterior probability is said to be optimal. Pairs of models can be compared using the ratio of the marginal likelihoods for the measurements, known as the Bayes factor. The Bayesian Information Criterion does this, but with an approximation to the marginal likelihood for the measurements. This model selection criterion is seen to be equivalent to comparing prequential logarithmic scores for each model.

5.1.1 Multi-Model Probabilities

Consider a model from a finite collection of models, $M \in \mathcal{M}$, for which the within-model prior is $\pi(\theta|M)$ and, the within-model likelihood is $\pi(x_i|\theta, M)$. For each model, the posterior $\pi(\theta|\mathring{x}_{1:t}, M)$ is computed using Bayes' theorem. Letting $\pi(M|\mathcal{M})$ denote the prior multi-model density, the posterior multi-model density is

$$\pi(M|\mathring{x}_{1:t}, \mathcal{M}) \propto \pi(\mathring{x}_{1:t}|M)\pi(M|\mathcal{M}), \tag{5.3}$$

obtained using Bayes' theorem, where

$$\pi(\mathring{x}_{1:t}|M) = \int \pi(\mathring{x}_{1:t}|\theta, M)\pi(\theta|M)d\theta \tag{5.4}$$

is the marginal density of the measurements. A popular choice of prior is to set $\pi(M|\mathcal{M}) \propto 1$ indicating *a priori* no preference between the models. The implications of this choice for the multi-model prior are discussed later in the chapter.

Bayes' Factor

When comparing two models, $M, M^* \in \mathcal{M}$, one considers the Bayes factor between two models. The Bayes factor

$$B(M, M^*) = \frac{\pi(\mathring{x}_{1:t}|M)}{\pi(\mathring{x}_{1:t}|M^*)} \frac{\pi(M|\mathcal{M})}{\pi(M^*|\mathcal{M})} \tag{5.5}$$

is the ratio of the posterior probabilities of each model. Selecting the model M when $B > 1$ over M^* is equivalent to selecting the model with the largest posterior probability. If we assume a flat multi-model prior, this is equivalent to selecting the model with the largest marginal likelihood of the measurements.

An approach to computing the Bayes' factor is, for each model, to approximate the marginal likelihood of the measurements. An approximation of the marginal likelihood, under particular model behaviour, leads to the classical Bayesian Information Criterion [Schwarz et al., 1978].

Theorem 6. *An approximation of $\pi(\mathring{x}_{1:t}|M)$ for each model is*

$$\pi(\mathring{x}_{1:t}|M) \approx \pi(\mathring{x}_{1:t}|\hat{\theta})\pi(\hat{\theta}) \left(\frac{2\boldsymbol{\pi}}{t}\right)^{\frac{\dim(\theta)}{2}} \det[\nabla^2_\theta \Phi_t(\hat{\theta})]^{-\frac{1}{2}}, \tag{5.6}$$

where $\Phi_t(\theta) = -t^{-1}\log(\pi(\mathring{x}_{1:t}|\theta)\pi(\theta))$ and $\hat{\theta}$ is the maximum posterior estimate.

Proof. Consider the marginal likelihood of the measurements

$$\pi(\mathring{x}_{1:t}|M) = \int \pi(\mathring{x}_{1:t}|\theta)\pi(\theta)d\theta. \tag{5.7}$$

Write the Taylor expansion of $\Phi_t(\theta) = -t^{-1}\log\pi(\mathring{x}_{1:t}|\theta)\pi(\theta)$ in θ around the maximum posterior estimate $\hat{\theta}$ as

$$\Phi_t(\theta) \approx \Phi_t(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \nabla^2_\theta \Phi_t(\hat{\theta})(\theta - \hat{\theta}). \tag{5.8}$$

Substituting (5.8) into (5.7), we have

$$\int \pi(\mathring{x}_{1:t}|\theta)\pi(\theta)d\theta = \int \exp(-t\Phi_t(\theta))d\theta \tag{5.9}$$

$$\approx \pi(\mathring{x}_{1:t}|\hat{\theta})\pi(\hat{\theta}) \int \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T t\nabla^2_\theta \Phi_t(\hat{\theta})(\theta - \hat{\theta})\right) d\theta. \tag{5.10}$$

The integral on the right-hand side is proportional to the density of a zero mean random variable with covariance $[t\nabla^2_\theta \Phi_t(\hat{\theta})]^{-1}$, and so we have

$$\pi(\mathring{x}_{1:t}|M) \approx \pi(\mathring{x}_{1:t}|\hat{\theta})\pi(\hat{\theta}) \left(\frac{2\boldsymbol{\pi}}{t}\right)^{\frac{\dim(\theta)}{2}} \det[\nabla^2_\theta \Phi_t(\hat{\theta})]^{-\frac{1}{2}}. \tag{5.11}$$

□

The approximation (5.6) is known as the generalised Bayesian Information Criterion [Konishi and Kitagawa, 2008]. Under further conditions, the generalised Bayesian Information Criterion reduces to the classical Bayesian Information Criterion (BIC) [Schwarz et al., 1978].

Corollary 7. *Assume that the measurements are conditionally independent and identically distributed. Further, assume a flat prior, so that* $\pi(\theta|M) \propto 1$*. Then, for large enough* t*, the Laplace approximation* (5.6) *is the Bayesian Information Criterion.*

Proof. If the prior is flat, and the measurements are conditionally independent and identically distributed, the maximum posterior estimate is equivalent to the maximum likelihood estimate. Furthermore,

$$\nabla^2_\theta \Phi_t(\hat{\theta}) = -\frac{1}{t}\sum_{i=1}^{t} \frac{\partial^2 \log \pi(x_i|\theta)}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \tag{5.12}$$

is the observed Fisher information matrix. For large enough t, we can ignore all terms that do not involve t and so we have

$$\log \pi(\mathring{x}_{1:t}|M) \approx \log \pi(\mathring{x}_{1:t}|\hat{\theta}) - \dim(\theta) \log t. \tag{5.13}$$

Multiplying by -2 gives the same formula found in the (only) proposition derived in Schwarz et al. [1978]. □

5.1.2 Criticising Bayesian Model Selection

Prequential Logarithmic Score

Within the proper scoring rules framework, the Bayes factor can be given a clear interpretation. In fact, model selection using a Bayes factor is equivalent to comparison of models using the logarithmic score. Equivalent to (5.5), we select the optimal model, $\hat{M}$, by solving

$$\hat{M} = \underset{M}{\operatorname{argmax}} \left(\log \pi(\mathring{x}_{1:t}|M) \right). \tag{5.14}$$

Then, the optimal model chosen by selecting the model with smallest prequential logarithmic score is

$$\begin{aligned}\hat{M} = \underset{M}{\operatorname{argmin}} \left(\sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i|\mathring{x}_{1:i-1}, M)) \right) &= \underset{M}{\operatorname{argmin}} \left(-\sum_{i=1}^{t} \log \pi(\mathring{x}_i|\mathring{x}_{1:i-1}, M) \right) \\ &= \underset{M}{\operatorname{argmax}} \left(\log \prod_{i=1}^{t} \pi(\mathring{x}_i|\mathring{x}_{1:i-1}, M) \right) \\ &= \underset{M}{\operatorname{argmax}} \left(\log \pi(\mathring{x}_{1:t}|M) \right). \end{aligned} \tag{5.15}$$

Therefore, selecting the model with the largest multi-model posterior probability is equivalent to selecting the model with smallest logarithmic score. Understanding the negative logarithm of the marginal likelihood of the measurements as a proper scoring rule, one might consider a generalisation to arbitrary proper scoring rules, where, in the case of the logarithmic score, we have $\mathcal{S}(\mathring{x}_{1:t}, \pi(x_{1:t}|M)) = -\log \pi(\mathring{x}_{1:t}|M)$. Using (5.13), or more generally, the logarithm of (5.11), as an approximation to $\pi(x_{1:t}|M)$, using a proper scoring rule different from the logarithmic score could be understood as a generalisation of the Bayesian Information Criterion, or its generalised version respectively.

Rather than approximating the marginal likelihood of all the measurements, we suggest that the predictive densities be evaluated directly. The prequential logarithmic score should be used in place of the Bayesian Information Criterion. Within the proper scoring rule paradigm, we treat predictive densities as the objects of interest, as opposed to models. For the logarithmic score, the score of the joint marginal likelihood of the measurements is equivalent to the prequential score

$$-\log \pi(\mathring{x}_{1:t}|M) = -\sum_{i=1} \log \pi(\mathring{x}_i|\mathring{x}_{1:i-1}). \tag{5.16}$$

This does not always hold for an arbitrary proper scoring rule.

$$\mathcal{S}(\mathring{x}_{1:t}, \pi(\cdot|M)) \neq \sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i|\mathring{x}_{1:i-1})). \tag{5.17}$$

Rather than applying a proper scoring rule to the joint likelihood of all the measurements, we suggest that we utilise the prequential score, along with an arbitrary proper scoring rule. The

prequential score is consistent in the sense that it asymptotically selects the correct model, if it is within our collection of models [Dawid, 1997]. This is also one of the attractive aspects of the Bayes factor . Rather than employing approximations, such as BIC, to the marginal likelihood of the measurements, we should focus on producing the predictive densities, and evaluate the model using the proper scoring rules.

Model Averaging

If we are to compute the multi-model posterior probability then it is reasonable that we could consider the multi-model predictive density

$$\pi(y|\mathring{x}_{1:t}, \mathcal{M}) = \sum_{M \in \mathcal{M}} \pi(y|\mathring{x}_{1:t}, M)\pi(M|\mathring{x}_{1:t}, \mathcal{M}), \tag{5.18}$$

where $\pi(M|\mathring{x}_{1:t}, \mathcal{M}) \propto \pi(\mathring{x}_{1:t}|M)\pi(M|\mathcal{M})$ and $\pi(M|\mathcal{M})$ is the prior of model M. That is, the collection of models yields a single predictive density. One can then treat this density as if from another model, and evaluate the forecasts made by the collection of models. This leads to methods of model averaging, outlined in detail in [Claeskens and Hjort, 2008], where the weights are proportional to the multi-model posterior probabilities.

Different predictive densities, based on a single model or a collection of models, can be evaluated separately using proper scoring rules. With this approach, one might be able to determine, for a collection of models, whether or not inclusion of a model increases the predictive performance.

Formulating the Multi-Model Prior

For the Bayesian, probabilities quantify a subjective degree of belief. Within the perfect model scenario consider the instance in which the perfect model is a member of the collection of models. Before we make a sequence of measurements we have no evidence *a priori* to suggest which model in our collection is the perfect model. This belief is formalised by choosing $\pi(M|\mathcal{M}) \propto 1$, a flat multi-model prior. The multi-model posterior quantifies

the belief that a model is the perfect model. As the number of measurements tends to infinity the multi-model posterior will place all probability mass on the perfect model. For a finite number of measurements, the model which has the maximum multi-model posterior probability serves as an estimate of the perfect model.

In the imperfect model scenario we know that none of the models in the collection are perfect. Therefore, we have *a priori* knowledge that the probability of any model being perfect is zero. To avoid this pitfall one might alter one's *a priori* belief by stating that *a priori* we have no evidence to suggest which is the *best* model. A best model must be in some respect optimal, and hence some criterion of optimality must be introduced.

Whatever that optimality criterion might be, models should be compared by evaluating that optimality criterion and so determine the optimal model directly. The central argument throughout this thesis is that a statistical model is considered optimal if it yields superior predictive performance and encourages the use of proper scoring rules for the purposes of the determining the optimal model.

5.2 Frequentist Risk Estimation

The Frequentist approach to risk estimation is characterised by introduction of the notion of a true density, from which the measurements are assumed to form a independent and identically distributed random sample. A statistical model is understood, in the Frequentist sense, as an approximation of the true density. The risk of using a statistical model $\pi(\cdot|\mathring{x}_{1:t}, M)$, when the true generating density is $\tau(\cdot)$, is

$$\mathcal{R} = \int \mathcal{S}(y, \pi(\cdot|\mathring{x}_{1:t}))\tau(y)dy = \mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\mathring{x}_{1:t}))], \tag{5.19}$$

where $\mathcal{S}$ is a proper scoring rule and $\boldsymbol{y} \sim \tau(\cdot)$ is a future unknown measurement. The true density $\tau(y)$ is known only through the observed sample, and we seek an estimator of the predictive risk. We can use *all* the measurements to fit the model and *all* the measurements

to estimate the risk. A proposed estimator of the risk (5.19) is

$$\hat{\mathcal{R}} = \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(\mathring{x}_i, \pi(\cdot|\mathring{x}_{1:t})). \tag{5.20}$$

This estimator utilises the measurements twice. Once, to fit the model, and, secondly, to estimate the risk. As such, the estimator is biased, and a correction term is required. So, assuming that, for all i, $\boldsymbol{x}_i \sim \tau(\cdot)$, the expected bias is

$$b = \int \left(\frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|x_{1:t})) - \int \mathcal{S}(y, \pi(\cdot|x_{1:t}))\tau(y)dy \right) \prod_{i=1}^{t} \tau(x_i)dx_i. \tag{5.21}$$

We consider the instance in which the predictive densities are from some parametric family of probability densities.

5.2.1 Parametric Predictive Densities

Formally, we write, for each i, $\boldsymbol{x}_i \sim \tau(\cdot)$. A future unknown measurement $\boldsymbol{y}$ is also assumed to be a realisation of the 'true' density so that the risk is

$$\mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\mathring{x}_{1:t}))] = \int \mathcal{S}(y, \pi(\cdot|\hat{\theta}))\tau(y)dy, \tag{5.22}$$

where $\pi(y|\mathring{x}_{1:t}) \cong \pi(y|\theta)$ and $\hat{\theta}$ is the minimum score estimate of the parameter made from the measurements. A natural estimator of the risk is (5.19). Since the measurements are used both to estimate the parameter, and then to estimate the risk, the natural estimator is biased.

Theorem 8. *An unbiased estimate of the risk is*

$$\hat{\mathcal{R}} = \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\hat{\theta})) - \mathbf{tr}[\hat{J}_t \hat{H}_t^{-1}(\hat{\theta})], \tag{5.23}$$

where $\hat{J}_t$ and $\hat{H}_t$ are estimates of J and H respectively.

Proof. The bias is computed by treating the measurements as a sample from the true distribution as

$$b = \int \left(\frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\hat{\theta})) - \int \mathcal{S}(y, \pi(\cdot|\hat{\theta}))\tau(y)dy \right) \tau(x_{1:t})dx_{1:t}. \tag{5.24}$$

Introducing the pseudo true state $\tilde{\theta}$, write the above integrand split into three parts as

$$P_1 = \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\hat{\theta})) - \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\tilde{\theta})), \tag{5.25}$$

$$P_2 = \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\tilde{\theta})) - \int \mathcal{S}(y, \pi(\cdot|\tilde{\theta}))\tau(y)dy, \tag{5.26}$$

$$P_3 = \int \mathcal{S}(y, \pi(\cdot|\tilde{\theta}))\tau(y)dy - \int \mathcal{S}(y, \pi(\cdot|\hat{\theta}))\tau(y)dy, \tag{5.27}$$

so that $b = \mathbb{E}\left[P_1 + P_2 + P_3\right]$, where the expectation is with respect to measurements, assumed to be distributed according to $\tau(\cdot)$, the true density. Let $J = \mathbb{E}[\Psi(\tilde{\theta})\Psi(\tilde{\theta})^T]$ and $H = \mathbb{E}[-\nabla_\theta \Psi(\tilde{\theta})]$ be the matrices that define the limiting distribution of the minimum score estimate as in Theorem 4. Consider (5.25) a second order Taylor expansion of the pseudo-true state around the estimate is

$$\frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\tilde{\theta})) \approx \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\hat{\theta})) + \frac{1}{2}(\tilde{\theta} - \hat{\theta})^T \frac{1}{t}\sum_{i=1}^{t} \frac{\partial^2 \mathcal{S}(x_i, \pi(\cdot|\theta))}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} (\tilde{\theta} - \hat{\theta}). \tag{5.28}$$

Substituting this into (5.25), and taking an expectation with respect to the measurements, results in

$$\mathbb{E}[P_1] = \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\hat{\theta})) - \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\tilde{\theta}))\right] \approx \frac{1}{2}\mathbf{tr}[JH^{-1}]. \tag{5.29}$$

Consider (5.27), a second order Taylor expansion of the estimate around the pseudo-true state is

$$\int \mathcal{S}(y, \pi(\cdot|\hat{\theta}))\tau(y)dy \approx \int \mathcal{S}(y, \pi(\cdot|\tilde{\theta}))\tau(y)dy + \frac{1}{2}(\hat{\theta} - \tilde{\theta})^T \left(\int \frac{\partial^2 \mathcal{S}(y, \pi(\cdot|\tilde{\theta}))}{\partial \theta^2}\tau(y)dy\right)(\hat{\theta} - \tilde{\theta}). \tag{5.30}$$

Substituting this into (5.27), and taking an expectation with respect to the measurements, results in

$$\mathbb{E}[P_3] = \mathbb{E}\left[\int \mathcal{S}(y, \pi(\cdot|\tilde{\theta}))\tau(y)dy - \int \mathcal{S}(y, \pi(\cdot|\hat{\theta}))\tau(y)dy\right] \approx \frac{1}{2}\mathbf{tr}[JH^{-1}]. \tag{5.31}$$

Lastly, consider (5.26), which, by assumption

$$\mathbb{E}[P_2] = \frac{1}{t}\sum_{i=1}^{t} \mathcal{S}(x_i, \pi(\cdot|\tilde{\theta})) - \int \mathcal{S}(y, \pi(\cdot|\tilde{\theta}))\tau(y)dy \to 0. \tag{5.32}$$

Adding $\mathbb{E}[P_1] + \mathbb{E}[P_2] + \mathbb{E}[P_3]$ yields $b \approx \mathbf{tr}[JH^{-1}]$. Replacing J and H by their estimates yields the result. □

This theorem generalises the case of composite likelihoods [Varin et al., 2011], where composite likelihoods are a particular type of proper scoring rule [Dawid et al., 2014]. Composite likelihoods take advantage of how the log likelihood functions decompose. When utilising the logarithmic score classical model selection criteria can be recovered.

Classical Model Selection Criteria

Most classical information criteria can be understood as scoring particular types of models using the logarithmic score [Burnham and Anderson, 2002]. For example, consider models with conditionally independent and identically distributed measurements with likelihood function $\pi(x_i|\theta)$, so that the minimum score estimate is the maximum likelihood estimate. Then the bias-corrected risk estimate is the Takeuchi [1976] Information Criterion, divided by 2, written as

$$\hat{\mathcal{R}} = -\frac{1}{t}\sum_{i=1}^{t} \log \pi(\mathring{x}_i|\hat{\theta}) + \mathbf{tr}[JH^{-1}]. \tag{5.33}$$

Furthermore, if we assume that the true model is a member of the parametric family of predictive densities, then the unbiased estimate of the risk is the Akaike [1973] Information Criterion, as Proposition 8 implies $\mathbf{tr}[JH^{-1}] = \dim(\theta)$.

Theorem 8 can be understood as a generalisation of classical model selection criteria. Therefore the proper scoring rule approach, when modelling measurements as conditionally independent and identically distributed and employing the logarithmic score, recovers classical model selection criteria. This constitutes a novel application of the proper scoring rule framework as well as a novel generalisation of the classical model selection criteria and is the primary result of this chapter. As such the approach to model selection using Frequentist risk estimation can be understood as the solution to a Bayesian decision problem where a proper scoring rule is the loss function and the optimal action is the predictive density. What

characterises the Frequentist Risk estimation as 'Frequentist' is the application of the law of large numbers to allow for a description of the limiting behaviour of the estimate as the number of measurements tends to infinity. This assumption is criticised and discussed in the following section.

5.2.2 Criticising Frequentist Risk Estimation

The assumption necessary for the calculation of the Frequentist risk is that of a parametric model where measurements are modelled as conditionally independent and identically distributed. In particular, the assumption of conditionally independent and identically distributed measurements is necessary in order to apply the law of large numbers. In particular, assuming a parametric model, we have for any fixed θ

$$\frac{1}{t}\sum_{i=1}^{t}\mathcal{S}(\mathring{x}_i, \pi(\cdot|\theta)) \to \int \mathcal{S}(y, \pi(y|\theta))\tau(y)dy = \mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\theta))] \tag{5.34}$$

where $\boldsymbol{y} \sim \tau(\cdot)$ is an unknown future measurement. The Frequentist risk estimate allows for the use of all measurements in model inference and then all measurements again to estimate the risk by approximating the bias in the estimator. An alternative manner in which an unbiased estimator of the Frequentist risk can be obtained is to split the measurements into training and testing sets.

Training and Testing

Let $s < t$, so that $\mathring{x}_{1:s}$ is the training set and $\mathring{x}_{s+1:t}$ is the testing set. Statistical models are fitted using the training set, and then evaluated using the testing set. The minimum score estimate

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{s}\mathcal{S}(\mathring{x}_i, \pi(x_i|\theta, M)) \tag{5.35}$$

is found using the training set. For each model the average fitted score is evaluated using the testing set and the optimal model is

$$\hat{M} = \underset{M}{\operatorname{argmin}} \left(\frac{1}{t-(s+1)} \sum_{i=s+1}^{t} \mathcal{S}(\mathring{x}_i, \pi(x_i|\hat{\theta}, M)) \right). \tag{5.36}$$

In recent applications of proper scoring rules this approach of separating measurements into training and testing sets is popular. Examples of such analysis can be found, for instance, in statistical models of seismic hazard estimation [Albarello and D'Amico, 2008], rainfall [Little et al., 2009], wind power [Lau et al., 2010, Pinson and Kariniotakis, 2010, Jeon and Taylor, 2012, Pinson, 2012], windspeed [Sloughter et al., 2010], call centre traffic [Taylor, 2012], and load forecasting [Arora and Taylor, 2013], to name a few areas of application.

If the measurements are modelled as conditionally independent and identically distributed, then as both s and t tend to infinity the average prequential score will tend to the expected score $\mathbb{E}[\mathcal{S}(\boldsymbol{y}, \pi(\cdot|\tilde{\theta}))]$ where $\tilde{\theta}$ is the pseudo-true state. When this model assumption is not met the law of large numbers does not apply and the average prequential score is not an estimate of the Frequentist risk for a future measurement. In fact, the fitted score

$$\sum_{i=s+1}^{t} \mathcal{S}(\mathring{x}_i, \pi(\cdot|\hat{\theta})) \text{ where } \hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{s} \mathcal{S}(\mathring{x}_i, \pi(x_i|\hat{\theta})) \tag{5.37}$$

is the sum of the scores for the k-step ahead predictions, where $k = 1, \ldots, t-(s+1)$. Note that we collect one score for each k. In contrast the prequential approach requires that we estimate $\hat{\theta}$ again (5.2) once the next measurement has been observed. The prequential score then reflects the predictive performance across a sequence of finite measurements rather than relying on large sample convergence.

If we wish to assess model performance across a number of forecast horizons then an appropriate modification of the prequential score is possible. For some fixed horizon k, the prequential score is

$$\sum_{i=1}^{t-k} \left(\sum_{j=0}^{k} \mathcal{S}(\mathring{x}_{i+j}, \pi(x_{i+j}|\hat{\theta}_i)) \right) \text{ where } \hat{\theta}_i = \underset{\theta}{\operatorname{argmin}} \sum_{l=1}^{i-1} \mathcal{S}(\mathring{x}_l, \pi(x_l|\theta)) \tag{5.38}$$

An optimal model is then selected, reflecting our preferences in experimental design.

Existence of the True Probability

The Frequentist position is characterised by the introduction of the existence of a 'true' probability from which the measurements form an independent and identically distributed random sample. An estimate of $\tau(\cdot)$ can be obtained from the sample of measurements and this estimate is the climatological model (Model 2). This model can be employed and scored in the usual manner by fixing a proper scoring rule, computing the prequential score and comparing it to alternative models. The introduction of a true probability, for the purpose of designing a model selection criterion can be avoided by utilising the prequential score. Since models are evaluated on their predictive performance there is no need to explicitly introduce a true probability. Neither the existence or nonexistence of a true probability is required for an application of proper scoring rule analysis.

5.3 Conclusion

In this chapter, we considered the classical model selection criteria, the Akaike Information Criterion and the Bayesian Information Criterion. The criteria are interpreted within the proper scoring rules framework. The classical approaches to model selection, using the Bayesian method of model selection (Section 5.1) and Frequentist method of risk estimation (Section 5.2), are understood as the solutions to a Bayesian decision problem , where a proper scoring rule is the loss function and the optimal action is the predictive density. Within this unifying framework an critical analysis of the classical model selection criteria is provided (Section 5.1.2 and Section 5.2.2). A novel generalisation of classical model selection is given by Theorem 8.

The Bayesian Information Criterion is computed by placing a flat, multi-model prior on the collection of models, and estimating the posterior multi-model probability for each model (Theorem 6). The model with the largest posterior probability is selected as the optimal model. Selecting an optimal model using the Bayes factor is seen to be equivalent to select-

ing the model with smallest prequential logarithmic score (5.15). Rather than considering approximations to the Bayes factor, we suggest that the prequential score be used, reducing to the Bayes factor in the case of the logarithmic score. Lastly, The use of the multi-model prior is criticised.

The Akaike Information Criterion is understood as an estimate of the predictive risk of a future unknown measurement. All the measurements are used to fit the model and to estimate the risk. As such the risk estimator is biased and a correction term is required. A generalisation to arbitrary proper scoring rules is given (Theorem 8). Estimation of the risk in this manner requires that the average score, for some fixed value of the parameter, approximates the risk for a single future measurement. This will not be true for all models. Once again the prequential score is preferred.

The prequential score is also useful for models in which the average score does not approach the risk of a single future measurement and so the law of large numbers cannot be applied. The prequential score avoids the problem of overfitting, as measurements are never used simultaneously for model fitting and evaluation. In the next chapter, we analyse simulations from a noisy chaotic dynamical system, apply various statistical models and the optimal model is selected using the prequential score.

Chapter 6

Chaotic Dynamical Systems

In this chapter we devise a series of experiments using low-dimensional dynamical systems to illustrate how model performance deteriorates outside of the perfect model scenario. Classically, noisy dynamical systems have been used to illustrate problems in parameter estimation. For example see McSharry and Smith [1999], Smirnov et al. [2005], Aguirre and Letellier [2009], and more recently Du and Smith [2012]. Rather than focusing on parameter estimation these analyses focus on the predictive performance of the statistical models. By analysing sequences of measurements generated from low-dimensional dynamical systems we are able to clearly define experiments in the perfect model scenario and design further experiments in the imperfect model scenario.

For some low-dimensional dynamical systems a numerical approximation of the Bayesian formulation is possible. We extend the Bayesian formulation of Judd [2003] by deriving the exact Bayesian solution for two one-dimensional chaotic maps. Using a low-dimensional grid, a numerical approximation to the Bayesian posterior is calculated. This is a numerical approximation to the probability density function rather than a sampling approach based on MCMC or a Gibbs sampler, such as the approach found in Bremer and Kaplan [2001]. The derivation of the Bayesian solution and the numerical implementation constitute a novel contribution.

Experiments are performed by first simulating sequences of noisy measurements using a one-dimensional chaotic map. Experiments in the imperfect model scenario are performed by considering measurements generated and corrupted with another one-dimensional dynamical system. The sequences of noisy measurements are analysed using a collection of statistical models. Each model, for a fixed proper scoring rule is evaluated and the density of expected score is approximated. The use of proper scoring rules allows us to illustrate the performance of the forecasting models in both the perfect and the imperfect model scenarios. By varying a 'mixing' parameter, we investigate predictive performance for various levels of corruption.

We shall see that the models that explicitly take into account the model dynamics perform better in the perfect model scenario. In the perfect model scenario the numerical Bayesian solution is shown to be optimal. When the corruption parameter is increased, the numerical Bayesian solution significantly under-performs compared with the other forecasting strategies. This suggests a danger in utilising precise dynamical models for forecasting when analysing measurements in the imperfect model scenario, and why forecast evaluation is an essential procedure.

The chapter proceeds by first defining the simulation process for experiments performed in the perfect model scenario and the imperfect model scenario. Next the four statistical models used are introduced and the predictive densities are computed. Lastly, fixing the logarithmic score as the proper scoring rule, the predictive performance of the models is evaluated and the chapter concludes with a discussion of the results.

6.1 Simulating Measurements

In this chapter we are interested in performing experiments in the perfect model scenario and comparing them to experiments performed in the imperfect model scenario. We wish to illustrate slight deviations away from the perfect model scenario. In this section we describe

exactly what is meant by experiments performed in either the perfect or imperfect model scenarios.

For each sequence of measurements we apply a collection of statistical models to $\mathring{x}_{1:t}$, a sequence of simulated measurements. The following defines the simulation procedure.

Definition 12. *A sequence $\mathring{x}_{1:t}$ of measurements is simulated in the following manner: Let $\mathcal{F}(\theta, \alpha)$ be some fixed nonlinear function known as the 'simulator'. Here θ is the state parameter, and α are the settings of the simulator. For some initial state θ_0 define, for each $i \leq t$,*

$$\theta_i = \mathcal{F}(\theta_{i-1}, \alpha), \tag{6.1}$$

$$\mathring{x}_i = \theta_i + \epsilon_i. \tag{6.2}$$

where the measurement noise $\epsilon_{1:t}$ is a realisation of an independent and identically distributed random sample with distribution $\boldsymbol{N}(0, \sigma^2)$.

6.1.1 Perfect Model Scenario

Within the perfect model scenario, recalling Definition 10, we have knowledge of the equations, both probabilistic and deterministic, that simulate the measurements. We will focus on changes to the simulator as opposed to mis-specification of the measurement noise. We shall assume probabilistic knowledge of the measurement noise throughout these experiments. The use of this knowledge is reflected when we construct some of the statistical models. Two of the four models under consideration are constructed using a function $f(\theta, \alpha)$ known as the forward map. We shall define experiments in the perfect model scenario as the analysis of simulated measurements calculated using a simulator which is equal to the forward map.

Definition 13. *An experiment in the perfect model scenario is performed by analysing $\mathring{x}_{1:t}$, a sequence of measurements simulated according to Definition 12 where the simulator is*

equivalent to the forward map.

$$\mathcal{F}(\theta, \alpha) := f(\theta, \alpha) \tag{6.3}$$

where θ is the state parameter and the variable defining the simulator.

6.1.2 Imperfect Model Scenario

We investigate deviations from the perfect model scenario by modifying the simulator to include the contribution of a function q, we shall refer to as the 'corrupting' map. When measurements are simulated in this manner we are in the imperfect model scenario.

Definition 14. *An experiment in the imperfect model scenario is performed by analysing $\mathring{x}_{1:t}$, a sequence of measurements simulated according to Definition 12 where the simulator is written in the form*

$$\mathcal{F}(\theta, \alpha, \beta, p) := pf(\theta, \alpha) + (1-p)q(\theta, \beta) \tag{6.4}$$

where f is the forward map, q is the corrupting map. The variables defining the simulator consist of α the parameters of the forward map, β the parameter defining the corruption map, and $p \in [0, 1]$ is the corruption mixing parameter.

Regardless of the exact settings of the simulator, any sequence of simulated measurements can be analysed using the following statistical models.

6.2 Constructing Models

We describe the four statistical models used to analyse $\mathring{x}_{1:t}$, a sequence of simulated measurements. The four models under consideration are called the Bayesian model, the sampling noise model, the random analogue prediction model, and the climatology model. In this section each model is defined and the predictive density is calculated.

6.2.1 Bayesian Model

For fully Bayesian models, point estimation of the parameters for the Bayesian solution is not necessary. An attempt at a Bayesian solution is found in Meyer and Christensen [2000], but Judd [2003] raises problems in the statistical formulation of the problem. Our formulation is equivalent to Judd [2003], but, instead of considering the posterior density of the initial state and the static parameter, we analyse forecasting performance by computing the sequential predictive densities and calculating the score.

Model 13. *Bayesian Model*

In this model, we assume that

$$\boldsymbol{\alpha} \sim \pi(\alpha), \tag{6.5}$$

$$\boldsymbol{\theta}_0 \sim \pi(\theta_0|\alpha), \tag{6.6}$$

$$\boldsymbol{\theta}_i \sim \pi(\theta_i|f(\theta_{i-1}, \alpha)), \tag{6.7}$$

$$\boldsymbol{x}_i \sim \pi(x_i|\theta_i), \tag{6.8}$$

where $f(\theta, \alpha)$ is the forward map defining the evolution of the state.

When f is a linear function, and both the density of the state evolution $\pi(\theta_i|\theta_{i-1})$, and the measurement likelihood $\pi(x_i|\theta_i)$, are Normal, the solution to the filtering problems $\pi(\theta_t|\mathring{x}_{1:t})$, and smoothing problems $\pi(\theta_i|\mathring{x}_{1:t})$, are known as the Kalman filter. The Kalman filter exploits properties of the Normal distribution, in that, if the prior density, state evolution density, and likelihood density, are Normal, the posterior density is also Normal. The likelihood function of the general case is

$$\pi(\mathring{x}_{1:t}|\alpha, \theta_0) = \prod_{i=1}^{t} \pi(\mathring{x}_i|f^i(\theta_0, \alpha)), \tag{6.9}$$

so that the posterior is

$$\pi(\alpha, \theta_0|\mathring{x}_{1:t}) \propto \prod_{i=1}^{t} \pi(\mathring{x}_i|f^i(\theta_0, \alpha))\pi(\theta_0|\alpha)\pi(\alpha). \tag{6.10}$$

When the density of the likelihood function is translation-invariant, $\pi(x_i|\theta) := \pi(x_i - \theta_i)$, and $\theta_i = f^i(\theta_0)$ is the i-fold composition, the formulation of Judd [2003] is recovered. When f is a nonlinear function, the posterior density is not necessarily Normal, even if the prior and likelihood densities are Normal.

Using the posterior density we write, for each i, the predictive density as

$$\pi(x_i|\mathring{x}_{1:i-1}) = \int\int \pi(x_i|\theta_i)\pi(\theta_i, \alpha|\mathring{x}_{1:i-1})d\theta_i d\alpha. \tag{6.11}$$

We can write, using the Markov property of the model,

$$\pi(\theta_i, \alpha|\mathring{x}_{1:i-1}) = \int \pi(\theta_i|\theta_{i-1}, \alpha)\pi(\theta_{i-1}, \alpha|\mathring{x}_{1:i-1})d\theta_{i-1}, \tag{6.12}$$

where $\pi(\theta_{i-1}, \alpha|\mathring{x}_{1:i-1})$ is the posterior density. In particular, for $i = 1$, $\pi(\theta_0, \alpha|\mathring{x}_{1:0}) = \pi(\theta_0|\alpha)\pi(\alpha)$ is the joint prior of the initial state and static parameter. If the evolution of the state space is modelled using a deterministic equation, we can use a Dirac delta function to write (6.12) as

$$\pi(\theta_i, \alpha|\mathring{x}_{1:i-1}) = \int \delta(\theta_i - f(\theta_{i-1}, \alpha))\pi(\theta_{i-1}, \alpha|\mathring{x}_{1:i-1})d\theta_{i-1}. \tag{6.13}$$

The properties of the Dirac delta function can be used to simplify the integral using the following lemma.

Lemma 2. *If $g(x)$ is a sufficiently differentiable function with a finite number of isolated simple roots then*

$$\delta(g(x)) = \sum_j \frac{\delta(x - r_j)}{|\nabla g(r_j)|} \tag{6.14}$$

where, for all j, $g(r_j) = 0$.

Proof. Since the roots of the function are assumed to be isolated, consider a partition of the underlying space into non-intersecting sets $\mathcal{X}_j$, such that $\cup_j \mathcal{X}_j = \mathcal{X}$ and $\mathcal{X}_j \cap \mathcal{X}_j = \emptyset$ where, for each j, $r_j \in \mathcal{X}_j$. Then, for any test function ϕ, we can write

$$\int_{\mathcal{X}} \phi(x)\delta(g(x)) = \sum_j \int_{\mathcal{X}_j} \phi(x)\delta(g(x))dx. \tag{6.15}$$

Let $z = g(x)$ so that $dx = |\nabla g(x)|^{-1} dz$. Then, for each j, write

$$\int_{\mathcal{X}_j} \phi(x)\delta(g(x))dx = \int_{\mathcal{X}_j} \phi(x)\delta(g(x))|\nabla g(x)|^{-1} dz. \tag{6.16}$$

For each j, the only non-zero contributions to the integral are the points at which the delta function is zero. It follows that $\delta(g(x)) = 0 \iff g(x) = 0$. For each j, $g(x) = 0 \implies x = r_j$. Then, the right-hand side of (6.16), summed over j, is

$$\sum_j \phi(r_j)|\nabla g(r_j)|^{-1} = \int \phi(x) \frac{\delta(x - r_j)}{\sum_j |\nabla g(r_j)|} dx. \tag{6.17}$$

Since the test function is arbitrary, the result follows. □

Employing Lemma 2, the predictive density is computed as

$$\pi(x_i|\mathring{x}_{1:i-1}) = \int\int\int \pi(x_i|\theta_i) \sum_j \frac{\delta(\theta_{i-1} - r_j))}{|\nabla_{\theta_{i-1}} f(r_j, \alpha)|} \pi(\theta_{i-1}|x_{1:i-1}) d\theta_{i-1} d\theta_i d\alpha \tag{6.18}$$

$$= \int\int \pi(x_i|\theta_i) \sum_j \frac{\pi(r_j|x_{1:i-1})}{|\nabla_{\theta_{i-1}} f(r_j, \alpha)|} d\theta_i d\alpha, \tag{6.19}$$

where r_j depends on θ_i and α, satisfying $f(r_j, \alpha) = \theta_i$ for each j.

6.2.2 Sampling-Noise Model

Secondly, we consider forecasts made by adding a random sample of the known measurement-noise to the previous measurement, and then applying the forward map to create a forecast ensemble. This model is known as the sampling-noise model, and was utilised to analyse the problem of parameter estimation in Du and Smith [2012].

The sampling-noise model takes advantage of the evolution of the state, as well as the measurement-noise, to produce an ensemble forecast of the next measurement. Ensemble members are generated by simulating $\epsilon^{(s)} \sim \boldsymbol{N}(0, \sigma^2)$, and applying the forward map so that

$$x_i^{(s)} = f(\mathring{x}_{i-1} + \epsilon^{(s)}, \alpha). \tag{6.20}$$

Each ensemble member constitutes an indistinguishable state, in that the measurement-noise allows for a number of values to be consistent with the previous observation [Judd and Smith, 2001]. Then, the predictive density can be computed as

$$\pi(x_i|\mathring{x}_{1:i-1}) = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{N}(x_i, x_i^{(s)}, \xi^2), \tag{6.21}$$

where ξ^2 is the variance of the Normal kernel used for the density estimation and, S is the number of ensemble members. In this model, we assume not to have knowledge of the static parameter. The parameter can be estimated by simulating a number of ensembles for a range of the parameter, and selecting the ensemble that minimises the prequential score. The ensemble with the lowest score is used to form the predictive density.

6.2.3 Random Analogue Prediction

The random analogue prediction model is constructed by finding trajectories in the observed time series that are similar to the most recent trajectory of measurements. These similar trajectories are referred to as 'analogues'. The use of analogues dates back to Lorenz [1969] and Lorenz [1955]. Paparella et al. [1997] suggest the use of random analogues for nonlinear prediction. The method of random analogue prediction has found success, more recently, in the problem of forecasting gross national products [Arora et al., 2013].

The number of analogues is defined by the size of neighbourhood, and the dimension of the space determined by the lags used in forming a delay vector. We estimate the neighbourhood size and dimension by minimising the forecast score, using a continuous ranked probability score, on the first one hundred time steps.

A prediction is made by randomly sampling from the set of similar trajectories, and issuing the subsequent measurement in the historical trajectory as the prediction for the next measurement. To make a forecast under the random analogue prediction model, we first find the

set

$$B(i, \delta, d) = \{\mathring{x}_k; ||(\mathring{x}_{k-1}, \ldots, \mathring{x}_{k-d}) - (\mathring{x}_{i-1}, \ldots, \mathring{x}_{i-d})||_2 \leq \delta\}, \tag{6.22}$$

where $||\cdot||_2$ is the Euclidean norm, d is the integer valued delay parameter, and $\delta > 0$ is the distance by which the trajectories are allowed to differ. The predictive density is

$$\pi(x_i|\mathring{x}_{1:i-1}) = \frac{1}{|B|} \sum_{k=1}^{|B|} \mathbb{1}\{x_i = \mathring{x}_k\}, \tag{6.23}$$

where $|B|$ is the cardinality of the set B, and $\mathbb{1}\{\cdot\}$ is the indicator function. Estimates of the model parameters d and δ are found by minimum score estimation.

6.2.4 Climatology Model

Lastly, as a 'baseline' forecast, we include, for comparison, the climatological forecast. The distribution function of the climatological model at time i is the empirical distribution

$$\Pi(x_i|\mathring{x}_{1:i-1}) = \frac{1}{i-1} \sum_{k=1}^{i-1} \mathbb{H}(x_i - \mathring{x}_k). \tag{6.24}$$

Alternatively, a kernel density estimator could be used. With a Normal kernel, the forecast density can be estimated as

$$\pi(x_i|\mathring{x}_{1:i-1}) = \frac{1}{i-1} \sum_{k=1}^{i-1} \boldsymbol{N}(x_i, \mathring{x}_k, \eta^2), \tag{6.25}$$

where the kernel variance, η^2, can be chosen using minimum score estimation.

6.3 Evaluating Models

An advantage to using a simulator is that a large number of simulations can be performed for fixed simulator settings. To evaluate the predictive performance of each statistical model,

we will analyse the deviations from the perfect model simulator by considering, for a fixed forward map $f(\theta, \alpha)$ and fixed corrupting map $q(\theta, \beta)$, simulators (6.4) of the form

$$\mathcal{F}(\theta, \alpha, \beta, p) := pf(\theta, \alpha) + (1-p)q(\theta, \beta), \tag{6.26}$$

varying $p \in [0, 1]$. When $p = 0$ we say we are in the perfect model scenario. Let $\mathring{\boldsymbol{x}}_{1:t}^{(p)}$ be a random variable whose realisations are simulated sequences of measurements

$$\mathring{x}_i^{(p)} = \mathcal{F}(\theta_{i-1}, \alpha, \beta, p) + \epsilon_i, \tag{6.27}$$

where ϵ_i is a realisation of the $\boldsymbol{N}(0, \sigma^2)$ random variable. For each model M the distribution of the random prequential score

$$\overline{\boldsymbol{\mathcal{S}}}_p(M) = \sum_{i=1}^{t} \mathcal{S}\left(\mathring{\boldsymbol{x}}_i^{(p)}, \pi(\cdot|\mathring{\boldsymbol{x}}_{1:i-1}^{(p)}, M)\right) \tag{6.28}$$

is estimated using a large number of simulations for fixed p. An analysis of model performance is given by inspecting the estimated distributions for various values of p.

As examples, two one-dimensional chaotic forward maps are analysed. For each map we compute the predictive density for the Bayesian model by using a novel numerical approximation of the Bayesian posterior. For fixed p, the distribution of $\overline{\boldsymbol{\mathcal{S}}}_p(M)$, the random prequential score is estimated by considering 200 simulations.

6.3.1 Simulator Settings

Each simulation is initialised by selecting a uniform random initial condition on the interval $[-1, 1]$. The simulator is run for a total of 500 time steps where the first 200 measurements are discarded leaving a sequence of simulated measurements of length 300. Of this sequence of simulated measurements the first 100 are used to fit the model. The models are evaluated by compute the proper score on the last 200 measurements. For these experiments we will utilise the logarithmic score to evaluate model performance.

As we vary p we can describe the predictive performance of the models as we deviate away from the perfect model scenario. For each map we shall fix the corrupting map as

$$q(\theta) = \frac{4}{5}\theta(1 - 2\theta^2 + \theta^3) \tag{6.29}$$

known as the quartic map. The chaotic maps under consideration are the *logistic map* and the *Gauss map*.

6.4 Two Chaotic Maps

6.4.1 The Logistic Map

The first example we consider is the logistic map defined by the forward map

$$f(\theta, \alpha) = 1 - \alpha\theta^2. \tag{6.30}$$

The simulator, for fixed p, is of the form

$$\mathcal{F}(\theta, p) := pf(\theta, 1.85) + (1 - p)q(\theta) \tag{6.31}$$

where f is given by (6.30) and q is given by (6.29). More recently, the logistic map has been used to elucidate problems in nonlinear parameter estimation. McSharry and Smith [1999] demonstrate the inadequacy of the maximum likelihood estimate. This result was extended more generally by Judd [2007]. Mukhin et al. [2006] employed a Bayesian approach to the problem, utilising the maximum posterior estimate. Judd demonstrates the inadequacy of the Bayesian formulation by evaluating the accuracy of point estimates of the initial system state and the static parameters using the posterior density. Creveling et al. [2008] estimated the parameters using a synchronisation technique, while Smith et al. [2010] introduce a geometric method. Du and Smith [2012] approach the problem of nonlinear parameter estimation by employing a minimum score estimate. In this chapter, in contrast to Du and Smith [2012], the focus is not parameter estimation, but instead the problem of forecasting.

6.4.2 Bayesian Solution

Again, using Lemma 2, an analytical solution to the sequential estimation of the Bayesian posterior is possible. A numerical approximation to the Bayesian posterior is described, and used to make probabilistic forecasts.

Model 14. *The Logistic Map*

Let $x_i \in \mathbb{R}, \theta_i \in [-1, 1], \alpha \in [0, 2]$

$$\boldsymbol{\alpha} \sim \boldsymbol{U}[0, 2] \tag{6.32}$$

$$\boldsymbol{\theta}_0 \sim \boldsymbol{N}(\theta_{_{prior}}, \sigma_0) \tag{6.33}$$

$$\boldsymbol{\theta}_i \sim \delta(\theta_i - f(\theta_{i-1}, \alpha)) \tag{6.34}$$

$$\boldsymbol{x}_i \sim \boldsymbol{N}(x_i|\theta_i, \sigma^2) \tag{6.35}$$

Notice that if $\theta_0 \in [-1, 1]$, then $\theta_i \in [-1, 1]$ for all i. To find (6.19) in the case of the logistic map, we have

$$\theta^{\pm} = \pm\sqrt{\frac{1-\theta_i}{\alpha}} \implies |\nabla f(\theta^{\pm})| = 2\alpha|\theta^{\pm}| = 2\sqrt{\alpha(1-\theta_i)}, \tag{6.36}$$

where $\theta^{\pm}$ are the two roots of the equation $\theta_i - f(\theta_{i-1}, \alpha) = 0$. The prior of (θ_i, α) at time i is

$$\pi(\theta_i, \alpha|\mathring{x}_{1:i-1}) = \frac{\pi(\theta^+|\mathring{x}_{1:i-1}) + \pi(\theta^-|\mathring{x}_{1:i-1})}{2\sqrt{\alpha(1-\theta_i)}}. \tag{6.37}$$

The predictive density (6.19) at time i is obtained by integrating the posterior against the likelihood for the i-th measurement. To implement an approximate Bayesian solution, we approximate the posterior density using a grid as described in the next section.

Numerical Implementation

A two dimensional grid is used to approximate the probability density. Let $\theta \in \{-1 : \Delta_\theta : 1\}$ and $\alpha \in \{0 : \Delta_\alpha : 2\}$, where Δ_θ and Δ_α are the grid step-sizes for θ and α respectively. For

each pair (θ, α), the value on the grid at this point, is computed as

$$\pi_{\text{post}}(\theta,\alpha) = \frac{1}{2\sqrt{\alpha(1-\theta)}}\Big(\pi_{\text{prior}}(\theta^+,\alpha) + \pi_{\text{prior}}(\theta^-,\alpha)\Big) \text{ where } \theta^{\pm} = \pm\sqrt{\frac{1-\theta}{\alpha}}. \tag{6.38}$$

There are three important aspects to computing the grid approximation to (6.38). Firstly, we require $\theta^{\pm}$ to be real numbers. To ensure that $\theta^{\pm}$ are real numbers, we must have the argument in the square root to be positive. Notice that $\pi_{\text{prior}}(\alpha,\theta) = 0$ for $\theta < (1-\alpha)$ after the initial time step.

Secondly, note that it will not be true in general that $\pi_{\text{prior}}(\alpha,\theta^+)$ and $\pi_{\text{prior}}(\alpha,\theta^-)$ are points on the grid, and so we use linear interpolation between the grid points to compute the approximating value. Lastly, note that (6.38) is undefined at $\theta = 1$. To avoid this problem, we approximate the density, near $\theta = 1$, with the linear function $g(\theta)$

$$\frac{1}{\sqrt{\alpha(1-\theta)}} \approx g(\theta) = a(1-\theta) + b(\theta - 1 + \epsilon), \tag{6.39}$$

where a and b are constants that need to be determined. At a distance $\epsilon > 0$ from $\theta = 1$ the approximation (6.39) should be exact, so that we have

$$\frac{1}{\sqrt{\alpha(1-\theta)}} = g(1-\epsilon) \implies \frac{1}{\sqrt{\alpha\epsilon}} = a\epsilon. \tag{6.40}$$

Here $a = \frac{1}{\epsilon\sqrt{\epsilon\alpha}}$ is determined from (6.39) by setting $\theta = 1-\epsilon$, which yields (6.40). To determine b, we require that our approximation $g(\theta)$ yields an equivalent probability density in the sense that the measures be equal, or, in other words, the total probability under the curve is equal. To ensure this, the following equation

$$\int_{1-\epsilon}^{1} \frac{1}{\sqrt{\alpha(1-\theta)}} d\theta = \int_{1-\epsilon}^{1} \left(a(1-\theta) + b(\theta - 1 + \epsilon)\right) d\theta \tag{6.41}$$

should be satisfied. Substituting $a = \frac{1}{\epsilon\sqrt{\alpha\epsilon}}$ into (6.41), so that

$$\int_{1-\epsilon}^{1} b(\theta - 1 + \epsilon) d\theta = \int_{1-\epsilon}^{1} \frac{1}{\sqrt{\alpha(1-\theta)}} d\theta - \int_{1-\epsilon}^{1} \frac{1}{\epsilon\sqrt{\alpha\epsilon}}(1-\theta) d\theta. \tag{6.42}$$

Integrating each term separately yields

$$b\frac{1}{2}(\theta - 1 + \epsilon)^2\Big|_{1-\epsilon}^{1} = \frac{-2\sqrt{1-\theta}}{\sqrt{\alpha}}\Big|_{1-\epsilon}^{1} - \frac{-(1-\theta)^2}{2\epsilon\sqrt{\alpha\epsilon}}\Big|_{1-\epsilon}^{1} \tag{6.43}$$

$$b\frac{1}{2}\epsilon^2 = \frac{2\sqrt{\epsilon}}{\sqrt{\alpha}} - \frac{\epsilon^2}{2\epsilon\sqrt{\alpha\epsilon}} = \frac{3\epsilon}{2\sqrt{\alpha\epsilon}} \implies b = \frac{3}{\epsilon\sqrt{\alpha\epsilon}}. \tag{6.44}$$

Now that we have found a and b, we can determine our approximation. At $\theta = 1$ we have

$$\left.\frac{1}{2\sqrt{\alpha(1-\theta)}}\right|_{\theta=1} \approx \left.\frac{1}{2}g(\theta)\right|_{\theta=1} = \frac{b\epsilon}{2} = \frac{3}{2\sqrt{\alpha\epsilon}}. \tag{6.45}$$

With the numerical implementation in hand, we plot, in Figure 6.1, the probabilistic predictions against the measurements to visualise the performance of the numerical Bayesian model. Only in the perfect model scenario can the notion of a true parameter make sense. In the perfect model scenario, the minimum score accurately estimates the true parameter value of the logistic map. The predictions generated using the sampling-noise ensembles are visualised in Figure 6.2, where the minimum continuous ranked probability score estimate for α is used to select the ensemble used for prediction.

For the random analogue prediction model the estimates of the delay and distance parameters are the minimum continuous ranked probability score estimates. The forecasts are visualised in Figure 6.3. Lastly the climatological model is applied and the forecasts are visualised in Figure 6.4.

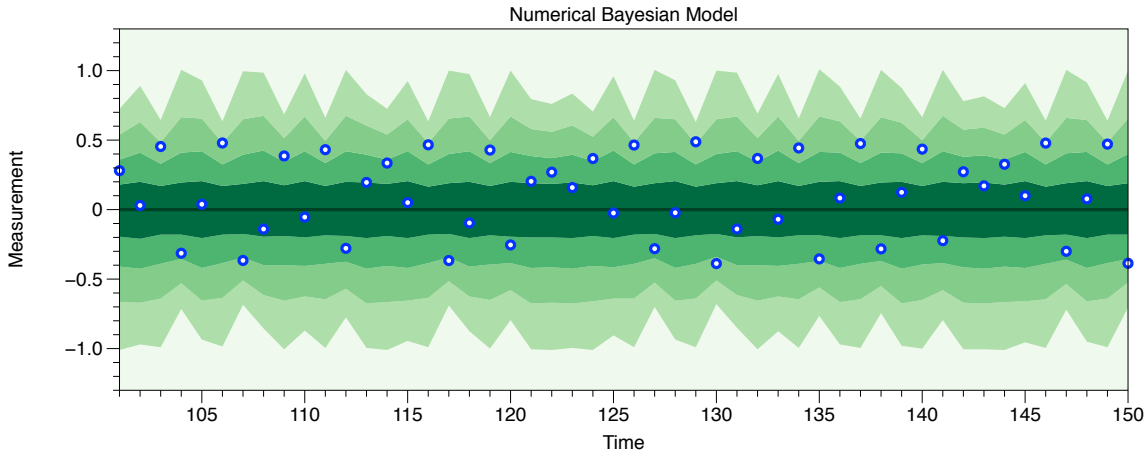


Figure 6.1: For fifty time steps the probabilistic forecasts produced by the numerical Bayesian solution are plotted along with the actualising values of the measurements. The graph is shifted about the median (black line) to help visualise which quantile the measurements fall into. Each shade of green above or below the median (black line) is a decile.

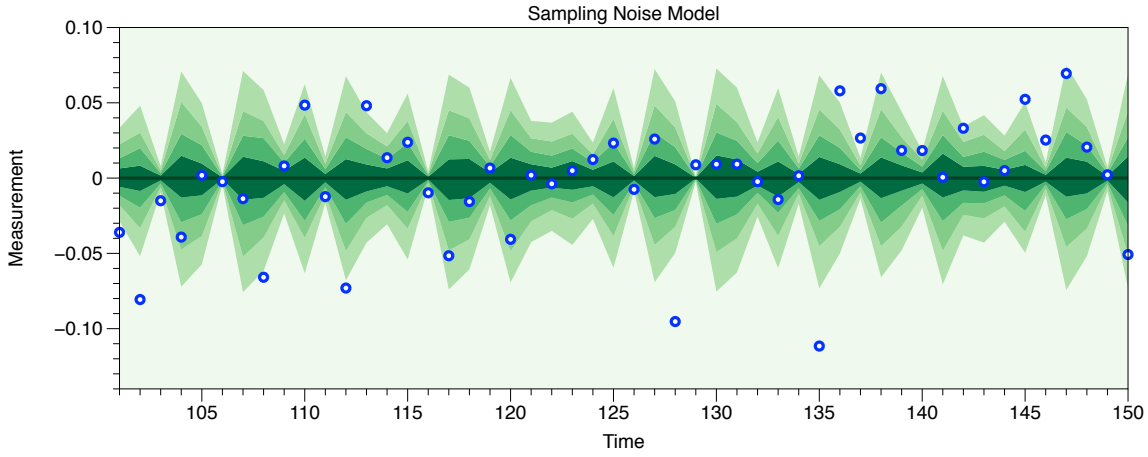


Figure 6.2: For one hundred time steps, the probabilistic forecasts produced by the sampling noise model are plotted along with the actualising values of the measurements. The graph is shifted about the median (black line) to help visualise which quantile the measurements fall into. Each shade of green above or below the median (black line) is a decile.

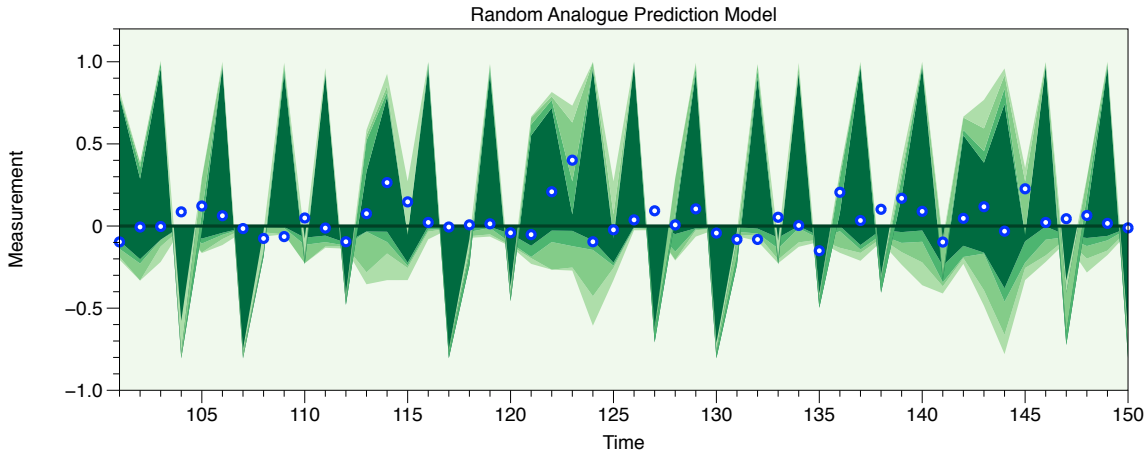


Figure 6.3: For fifty time steps, the probabilistic forecasts produced by the random analogue prediction model are plotted along with the actualising values of the measurements. The graph is shifted about the median (black line) to help visualise which quantile the measurements fall into. Each shade of green above or below the median (black line) is a decile.

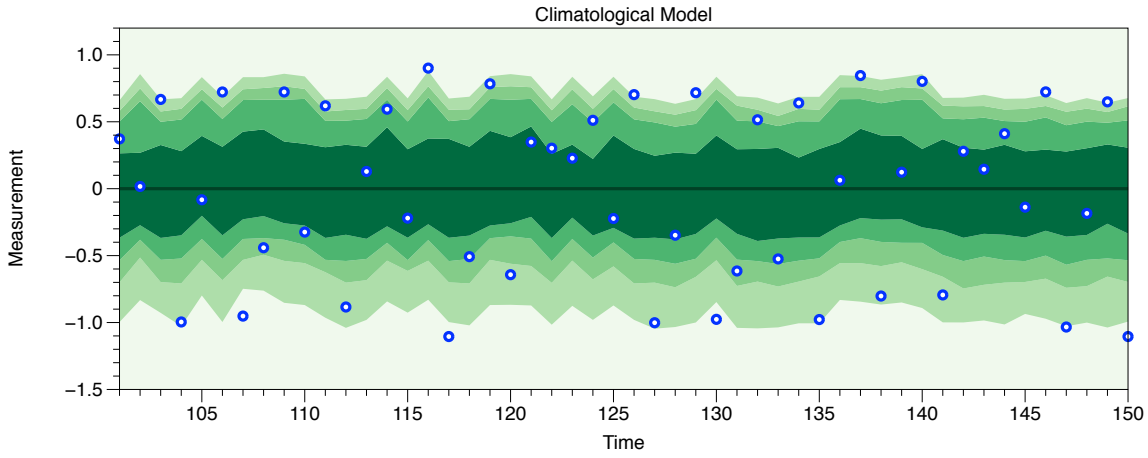


Figure 6.4: For fifty time steps, the probabilistic forecasts produced by the climatological model are plotted along with the actualising values of the measurements. The graph is shifted about the median (black line) to help visualise which quantile the measurements fall into. Each shade of green above or below the median (black line) is a decile.

6.4.3 The Gauss Map

As a second example, consider the Gauss map defined by the forward map

$$f(\theta, \alpha, \beta) = e^{-\alpha\theta^2} + \beta. \tag{6.46}$$

The simulator, for fixed p, is of the form

$$\mathcal{F}(\theta, p) := pf(\theta, (8, -0.58)) + (1-p)q(\theta). \tag{6.47}$$

where f is given by (6.46) and q is given by (6.29).

The Bayesian Model

Using Lemma 2, an analytical solution to the sequential estimation of the Bayesian posterior is possible. A numerical approximation to the Bayesian posterior is described, and used to make probabilistic forecasts.

Model 15. *The Gauss Map*

Let $x_i \in \mathbb{R}, \theta_i \in [-1, 1], \alpha \in [0, 2]$

$$\boldsymbol{\alpha} \sim \boldsymbol{U}[5, 10] \tag{6.48}$$

$$\boldsymbol{\beta} \sim \boldsymbol{U}[-1, 1] \tag{6.49}$$

$$\boldsymbol{\theta}_0 \sim \boldsymbol{N}(\theta_{prior}, \sigma_0) \tag{6.50}$$

$$\boldsymbol{\theta}_i \sim \delta(\theta_i - f(\theta_{i-1}, \alpha, \beta)) \tag{6.51}$$

$$\boldsymbol{x}_i \sim \boldsymbol{N}(x_i|\theta_i, \sigma^2) \tag{6.52}$$

where $f(\theta, \alpha, \beta) = e^{-\alpha\theta^2} + \beta$.

To apply Lemma 2 we solve, for fixed θ_i, the equation

$$\theta_i = e^{-\alpha\theta_{i-1}^2} + \beta. \tag{6.53}$$

The roots of (6.53) are

$$\theta^{\pm} = \pm\sqrt{\frac{-\log(\theta_i - \beta)}{\alpha}} \implies |\nabla_\theta f(\theta^{\pm}, \alpha, \beta)| = 2\alpha|\theta^{\pm}|e^{-\alpha(\theta^{\pm})^2}, \tag{6.54}$$

so that the posterior density is computed using Lemma 2 as

$$\pi(\theta, \alpha, \beta|\mathring{x}_{1:i}) = \frac{1}{\alpha|\theta^+|} e^{\alpha(\theta^+)^2} \pi(\theta^+|\mathring{x}_{1:i-1}) + \frac{1}{\alpha|\theta^-|} e^{\alpha(\theta^-)^2} \pi(\theta^-|\mathring{x}_{1:i-1}) \tag{6.55}$$

$$= \frac{\pi(\theta^+|\mathring{x}_{1:i-1}) + \pi(\theta^-|\mathring{x}_{1:i-1})}{(\theta_i - \beta)\sqrt{-\alpha \log(\theta_i - \beta)}}. \tag{6.56}$$

The predictive density (6.19) at time i is obtained by integrating the posterior against the likelihood for the i-th measurement. To implement an approximate Bayesian solution, we approximate the posterior density using a grid as described in the next section.

Numerical Implementation

A three dimensional grid is used to approximate the probability density. Let $\alpha \in [\alpha_l : \Delta_\alpha : \alpha_u]$ and $\beta \in [\beta_l : \Delta_\beta : \beta_u]$ and $\theta \in [\beta_l : \Delta_\theta : 1 + \beta_u]$. For each triplet (α, β, θ), the value on the grid is computed as

$$\pi_{\text{post}}(\alpha, \beta, \theta) = \frac{1}{(\theta - \beta)\sqrt{-\alpha \log(\theta - \beta)}} \Big(\pi_{\text{prior}}(\alpha, \beta, \theta^+) + \pi_{\text{prior}}(\alpha, \beta, \theta^-) \Big), \tag{6.57}$$

where

$$\theta^\pm = \pm \sqrt{\frac{-\log(\theta - \beta)}{\alpha}}. \tag{6.58}$$

There are three important aspects to computing the grid based approximation. Firstly, we require $\theta^\pm$ to be real numbers. To ensure that $\theta^\pm$ are (well defined) real numbers, we must have

$$-\log(\theta - \beta) > 0 \implies 0 < \theta - \beta \leq 1. \tag{6.59}$$

Secondly, note that it will not be true in general that $\theta^\pm$ are points on the grid, and so we use linear interpolation between the grid points to compute the approximating value. Lastly, notice that (6.57) is undefined at $\theta = 1 + \beta$. To avoid this problem, we approximate the density, near $\theta = 1 + \beta$, with the linear function as

$$\frac{1}{(\theta - \beta)\sqrt{-\alpha \log(\theta - \beta)}} \approx g(\theta) = a(1 + \beta - \theta) + b(\theta - (1 + \beta - \epsilon)). \tag{6.60}$$

At a distance $\epsilon > 0$ from $1+\beta$ the approximation should be exact, so that we have, at $\theta = 1+\beta-\epsilon$,

$$\frac{1}{(1-\epsilon)\sqrt{-\alpha\log(1-\epsilon)}} \approx g(1+\beta-\epsilon) = a\epsilon \tag{6.61}$$

$$\implies a = \frac{1}{\epsilon(1-\epsilon)\sqrt{-\alpha\log(1-\epsilon)}}. \tag{6.62}$$

To determine b, we require that our approximation $g(\theta)$ yields an equivalent probability density in the sense that the measures be equal, or, in other words, the total probability under the curve is equal. To ensure this, the following equation,

$$\int_{1+\beta-\epsilon}^{1+\beta} \frac{1}{(\theta-\beta)\sqrt{-\alpha\log(\theta-\beta)}} d\theta = \int_{1+\beta-\epsilon}^{1+\beta} a(1+\beta-\theta) + b(\theta-(1+\beta-\epsilon))d\theta \tag{6.63}$$

should be satisfied, where a is given by (6.62), and b remains to be determined. The left hand side of (6.63) is

$$\int_{1+\beta-\epsilon}^{1+\beta} \frac{1}{(\theta-\beta)\sqrt{-\alpha\log(\theta-\beta)}} d\theta = -2\sqrt{\frac{-\log(\theta-\beta)}{\alpha}}\Bigg|_{1+\beta-\epsilon}^{1+\beta} = 2\sqrt{\frac{-\log(1-\epsilon)}{\alpha}}. \tag{6.64}$$

The right-hand side of (6.63) can be evaluated in two parts. The first is

$$\int_{1+\beta-\epsilon}^{1+\beta} a(1+\beta-\theta)d\theta = -\frac{1}{2}a(1+\beta-\theta)^2\Big|_{1+\beta-\epsilon}^{1+\beta} = \frac{1}{2}a\epsilon^2. \tag{6.65}$$

The second is

$$\int_{1+\beta-\epsilon}^{1+\beta} b(\theta-(1+\beta-\epsilon))d\theta = \frac{1}{2}b(\theta-(1+\beta-\epsilon))^2\Big|_{1+\beta-\epsilon}^{1+\beta} = \frac{1}{2}b\epsilon^2. \tag{6.66}$$

Rewriting (6.63), using (6.64), (6.65), and (6.66) yields

$$\frac{1}{2}b\epsilon^2 = 2\sqrt{\frac{-\log(1-\epsilon)}{\alpha}} - \frac{1}{2}a\epsilon^2 = 2\sqrt{\frac{-\log(1-\epsilon)}{\alpha}} - \frac{\epsilon}{2(1-\epsilon)\sqrt{-\alpha\log(1-\epsilon)}} \tag{6.67}$$

$$\implies b = \frac{1}{\epsilon^2\sqrt{\alpha}}\left(4\sqrt{-\log(1-\epsilon)} - \frac{\epsilon}{(1-\epsilon)\sqrt{-\log(1-\epsilon)}}\right). \tag{6.68}$$

So that, for $\theta = 1+\beta$ the approximation is

$$g(1+\beta) = b\epsilon = \frac{1}{\epsilon\sqrt{\alpha}}\left(4\sqrt{-\log(1-\epsilon)} - \frac{\epsilon}{(1-\epsilon)\sqrt{-\log(1-\epsilon)}}\right). \tag{6.69}$$

Along with the numerical Bayesian model we apply the sampling-noise model, the random analogue prediction model and the climatological model.

6.5 Assessing Predictive Performance

Consider a sequence of measurements generated in the perfect model scenario. Since this model is almost completely known one would expect forecasts made using the numerical Bayesian solution, and the sampling-noise model, to be more accurate than climatology or random analogue prediction. In the perfect model scenario it is demonstrated that this is the case. However, when we move into the imperfect model scenario by increasing the corruption mixing parameter, we see that the numerical Bayesian solution is significantly non-optimal. In particular, for a large enough choice of the corruption mixing parameter, the climatology model is optimal relative to the numerical Bayesian solution. The proper scoring rule framework allows for all statistical models, Bayesian or not, to be compared. That is, by understanding prediction as a Bayesian decision problem we can make a Bayesian comparison of all competing statistical models to find the best one.

6.5.1 Perfect Model Scenario

To compare the models, we compute the logarithmic score for each model across 200 experiments in the perfect model scenario for three levels of measurement noise. The ensemble of scores for each model is used to approximate the density of the score. For each model, a box plot, indicating the quartiles as well as the 5% and 95% quantiles is plotted. For each noise level, the box plots are collected in Figure 6.5 in the example using the logistic map and in Figure 6.6 for the Gauss map. From the box plots displayed in Figures 6.5 and 6.6, it is clear that the numerical Bayesian solution outperforms all other models, including, notably, the sampling-noise model.

Notice that the sampling-noise model and the random analogue models perform better than the climatology model. There is a noticeable gain in score between the sampling-noise model and the random analogue prediction model. This is to be expected, since the sampling-noise model resembles, more closely, the simulation process. The numerical approximation to the

Bayesian solution yields a score that is clearly smaller than any of the alternatives. This demonstrates the optimality of the Bayesian model in the perfect model scenario.

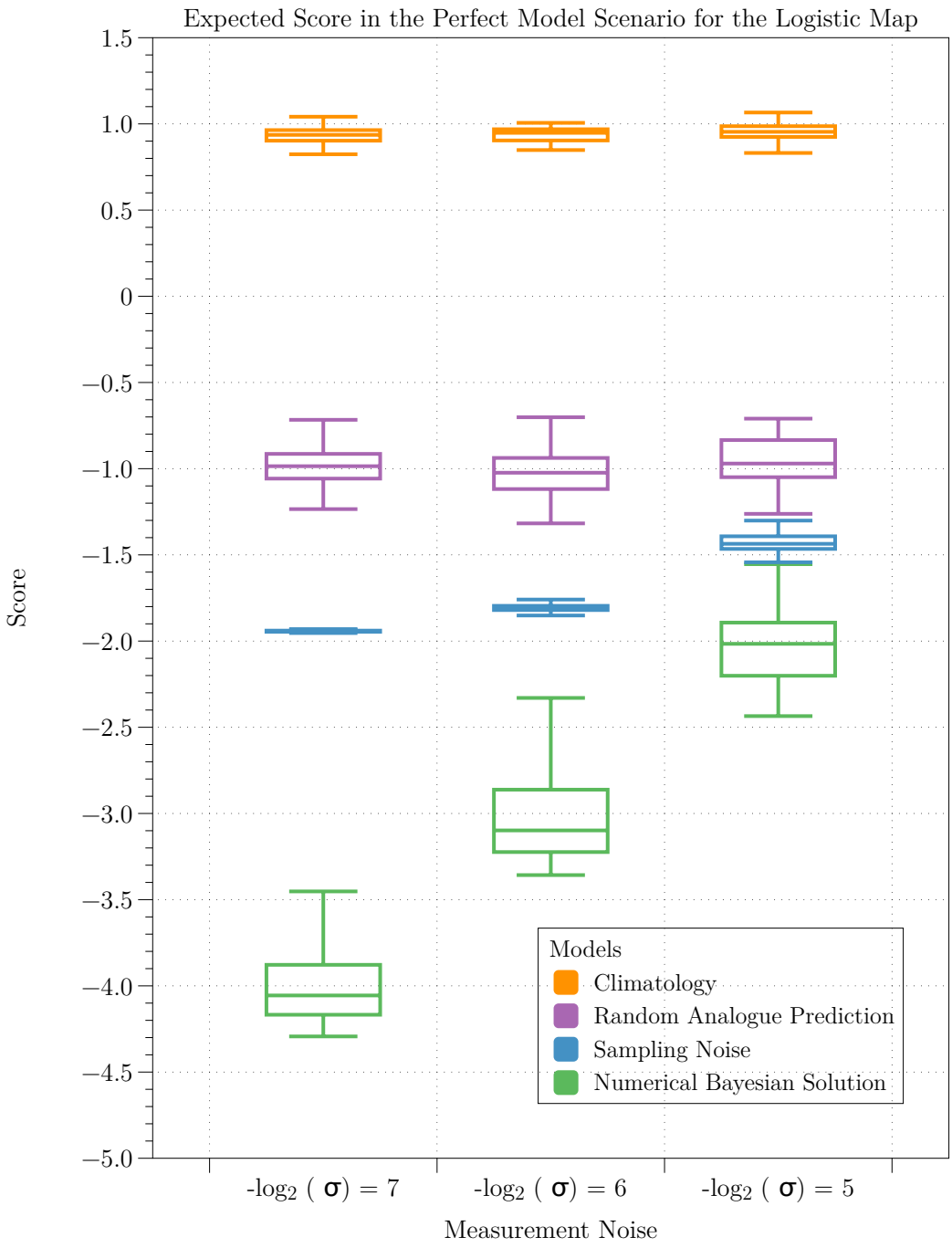


Figure 6.5: In the perfect model scenario, for varying noise levels, the box plots show that the model based numerical Bayesian solution outperforms the other non-perfect models applied to measurements generated using the logistic map.

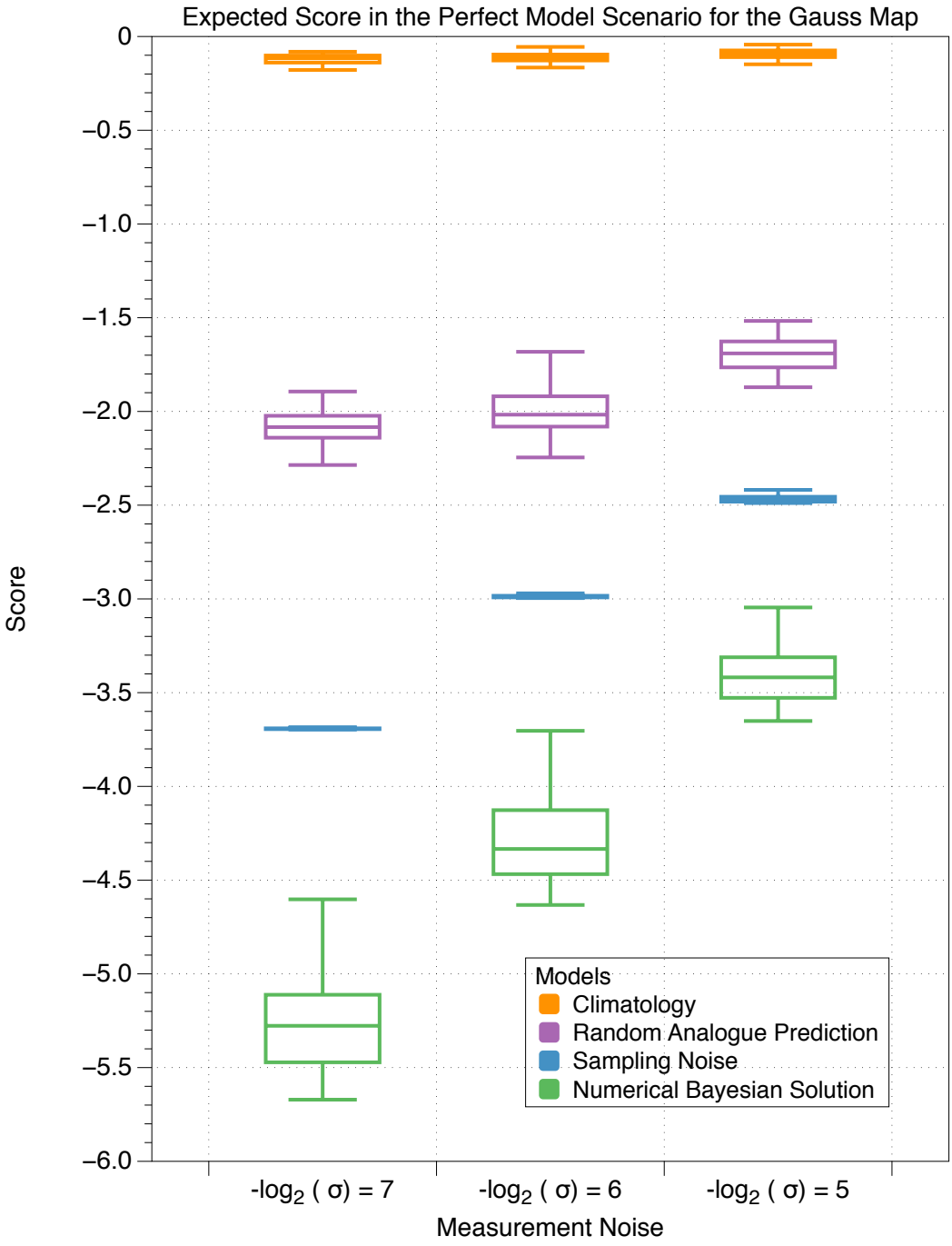


Figure 6.6: In the perfect model scenario, for varying noise levels, the box plots show that the model based numerical Bayesian solution outperforms the other non-perfect models applied to measurements generated using the Gauss map.

Benchmarking Simulators

Low-dimensional systems, such as the logistic map and the Gauss map, can be used to test the effects of errors in predictive models. Sequences of measurements are simulated using a known set of equations. The resulting simulations can be analysed using any predictive model. Climatology can be used to benchmark other models. If a model does not perform better than climatology, it should be discarded. Instead, one can just use climatology and make better predictions.

In the perfect model scenario, models that explicitly account for the true dynamics can be used to benchmark other forecasting systems. For example, if we have a fully Bayesian solution for the simulator available, like the one derived, this can act as an opposing benchmark. The score for an alternate model can be compared to that of the fully Bayesian solution for the simulator. The noise sampling model also takes advantage of the known dynamics of the simulator. The random analogue prediction model and the sampling-noise model yield similar scores, suggesting that the random analogue prediction model, which does not require any knowledge of the underlying true dynamics, performs almost as well as the sampling-noise model.

The proper scoring rule framework allows for a principled approach to model development. Models such as the climatological model and the random analogue prediction model can be used, along with the actualising measurements to determine benchmarks when applied to the actualising measurements made. Models like the sampling-noise model and the numerical Bayesian solution which explicitly take into account dynamics, can be applied to both the actualising measurements as well as simulated measurements made in the perfect model scenario.

One can then determine the differences between the score of the noise-sampling model and the numerical Bayesian solution and the benchmarks provided by the climatological model and the random analogue prediction model. A large positive difference indicates that the benchmarks should be used instead of the dynamical models. A large negative difference

indicates that the dynamical models are useful for prediction.

Additionally one can compare the score of the noise-sampling model and numerical Bayesian solution in the perfect model scenario, using simulated measurements, to the score obtained using the actualising measurements. This can give an indication as to how the models are performing outside the perfect model scenario. A small positive difference between the score obtained using the actualising measurements and the score obtained using simulated measurements in the perfect model scenario indicates that the model is a *good* description of the underlying phenomena.

6.5.2 Imperfect Model Scenario

We compare the approximate numerical Bayesian solution to the alternative models for each value of the corruption mixing parameter $p = 0, \frac{1}{2}\%, 1\%, 5\%, 10\%$. The results for the logistic map are illustrated in Figure 6.7 and in Figure 6.9 for the Gauss map. One can see that, when the corruption-mixing parameter is small, $p \leq 1\%$, the numerical Bayes' solution for both the logistic map and the Gauss map outperforms the alternative models. At 5% corruption, the Bayesian solution for both maps fails to beat the climatological forecast. In fact, the magnitude of the scores achieved by the Bayesian model greatly exceeds that of the other models.

One notices that the random analogue prediction model and the sampling-noise model produce similar magnitudes of scores across each corruption scenario. In Figure 6.7, the relative magnitude of the numerical Bayesian solution for the logistic map makes it difficult visually to compare the random analogue prediction model and the sampling noise model. The box plots in Figure 6.8 for the logistic map and in Figure 6.9 for the Gauss map illustrate that, if we increase the corruption-mixing parameter, the random analogue prediction begins to outperform the sampling-noise model. For the logistic map, when $p = 10\%$, the random analogue prediction model outperforms the sampling-noise model. In the case of the Gauss map the random analogue prediction model begins to outperform the sampling-noise model

at $p = 5\%$ corruption.

These results suggest that models that do not explicitly model the underlying dynamics can beat near perfect models when the signal is sufficiently corrupted. In the imperfect model scenario proper scoring rules can still be used to evaluated forecast performance even if a perfect model is unavailable. In practice, we are always in the imperfect model scenario, and this analysis suggests that caution should be exercised when using forecasting methods that model the dynamics explicitly, as, indeed, all forecasting methods should be criticised and scored in a principled way.

6.6 Conclusion

In this chapter, we designed experiments to compare the performance of various forecasting models for noisy measurements simulated in the perfect model scenario (Definition 13) and the imperfect model scenario (Definition 14). Rather than focusing on the classical problem of parameter estimation, we use chaotic dynamical systems to illustrate the problem of forecasting. Four different models were considered. Two of the models take advantage of knowledge of the measurement generating process, while the other two forecasting strategies make no use of knowledge of the underlying dynamics. In all instances where parameters need to be estimated, this is done selecting the minimum score estimation using the continuous ranked probability score.

A fully Bayesian solution, (6.19), for the sequential posterior density is derived using properties of the Dirac delta function. In both the logistic map and the Gauss map, finding the roots of the dynamical equation in closed form is possible, so an analytical Bayesian solution is possible. Implementing a numerical solution on a grid requires resolution of problems at the boundary. In particular the density at the boundary is linearly approximated subject to a conservation of probability requirement.

In the fully Bayesian solution, parameter estimation is performed implicitly in the marginal-

isation step. The sampling-noise model, (6.20), explicitly takes into account knowledge of the dynamics of the state evolution, as well as knowledge of the measurement-noise. The random analogue prediction model, (6.22), finds historical trajectories that are close to the most recent trajectory, and produces a probabilistic forecast by randomly sampling from the set of similar trajectories. Lastly, the climatological model, (6.25), models the measurements as independent and identically distributed, and the predictive distribution is just the empirical distribution.

In the perfect model scenario, the numerical Bayesian solution is shown to be optimal as it outperforms all other models. This is illustrated in Figure 6.5 for the logistic map and in Figure 6.6 for the Gauss map. The sampling-noise model outperforms the random analogue prediction model and the climatological model. This is expected, as both the Bayesian and sampling-noise models explicitly take into account the underlying dynamics of the observed time series.

To simulate deviations from the perfect model scenario and into the imperfect model scenario, a number of measurement sequences are generated using the logistic map, and are corrupted by mixing in influence from the quartic map. Once again, the four models are applied to the simulated sequences in the imperfect model scenario, illustrated in Figure 6.7. As the corruption parameter increases in magnitude, the Bayesian solution goes from being optimal to being significantly worse than even the baseline climatological forecast. Furthermore, as shown in Figure 6.8, as the corruption-mixing parameter increases, the random analogue prediction model outperforms the sampling-noise model.

These experiments illustrate that, while statistical models that explicitly model the dynamical evolution of the sequence of measurements perform well in the perfect model scenario, small deviations from perfection into the imperfect model scenario can result in poor performance relative to other models that do not model the dynamics explicitly, and instead rely on identifying patterns in the observed measurements. That is to say, when proposing complicated dynamical models to describe a sequence of measurements, one should be cautious.

Experiments, in both the perfect and the imperfect model scenario, should be performed to assess the predictive performance of the models. Analysis like that described in this chapter, illustrates the deterioration of the model performance in the imperfect model scenario and can be used to investigate how *robust* a model is to deviations away from the perfect model scenario.

In this analysis only one approach to defining deviations away from the perfect model scenario is considered. For a more thorough analysis of the model performance, a number of different types of deviations could be investigated. This type of analysis may allow the forecaster to illustrate model behaviour in a variety of scenarios.

In reality, when we are analysing measurements in the imperfect model scenario, solutions that attempt to model complicated dynamics, like the numerical Bayesian solution for a specific dynamical model, may not perform well under small deviations from the perfect model scenario. The proper scoring rule framework allows for such a hypothesis to be investigated. Using a variety of experiments in both the perfect and imperfect model scenario any number of models can be compared and contrasted in a principled way.

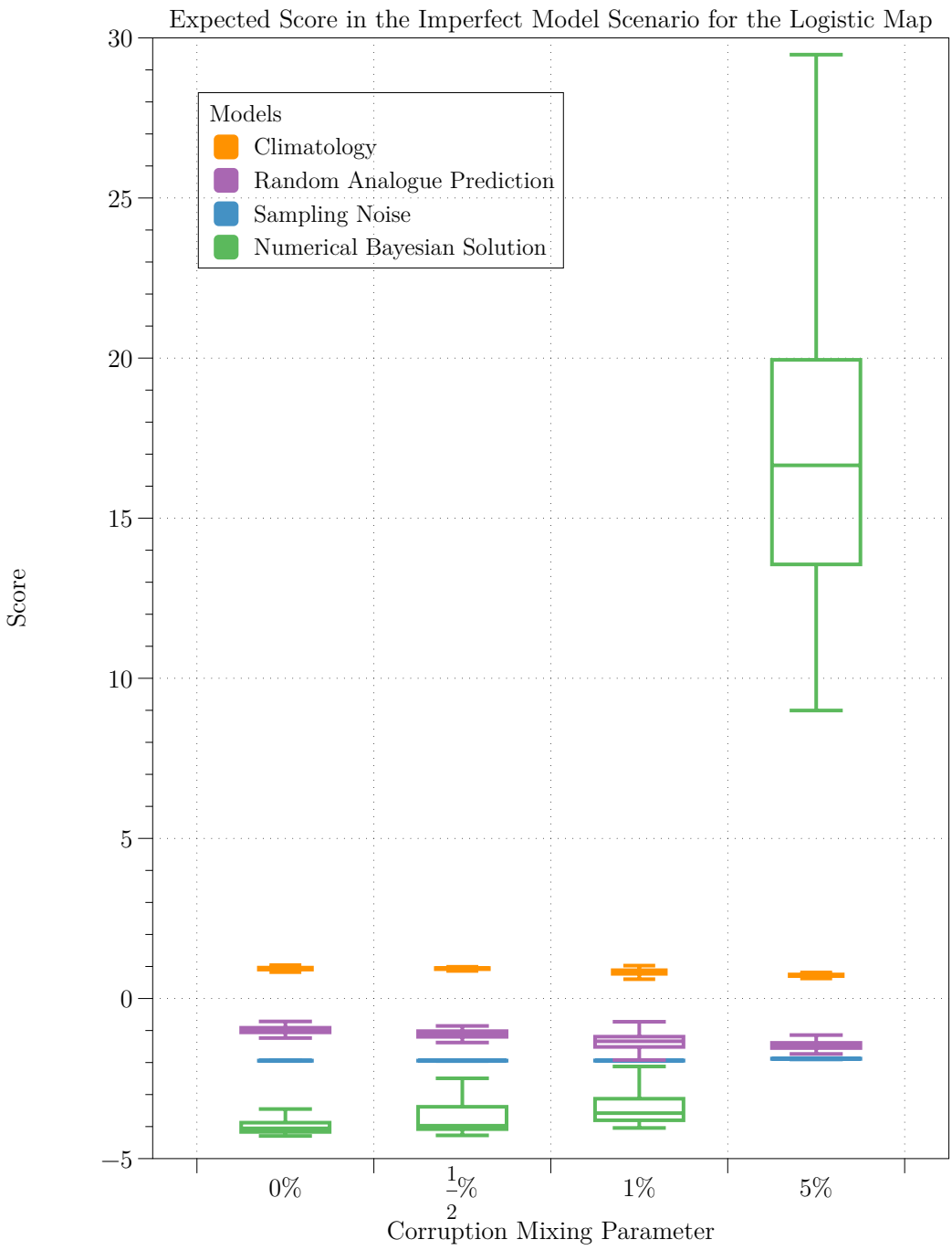


Figure 6.7: For various small values of the corruption-mixing parameter, each model is applied to simulations from the logistic map to generate a box plot estimating the density of the expected score. As the corruption increases, the numerical Bayesian solution goes from being the optimal model to being significantly suboptimal.

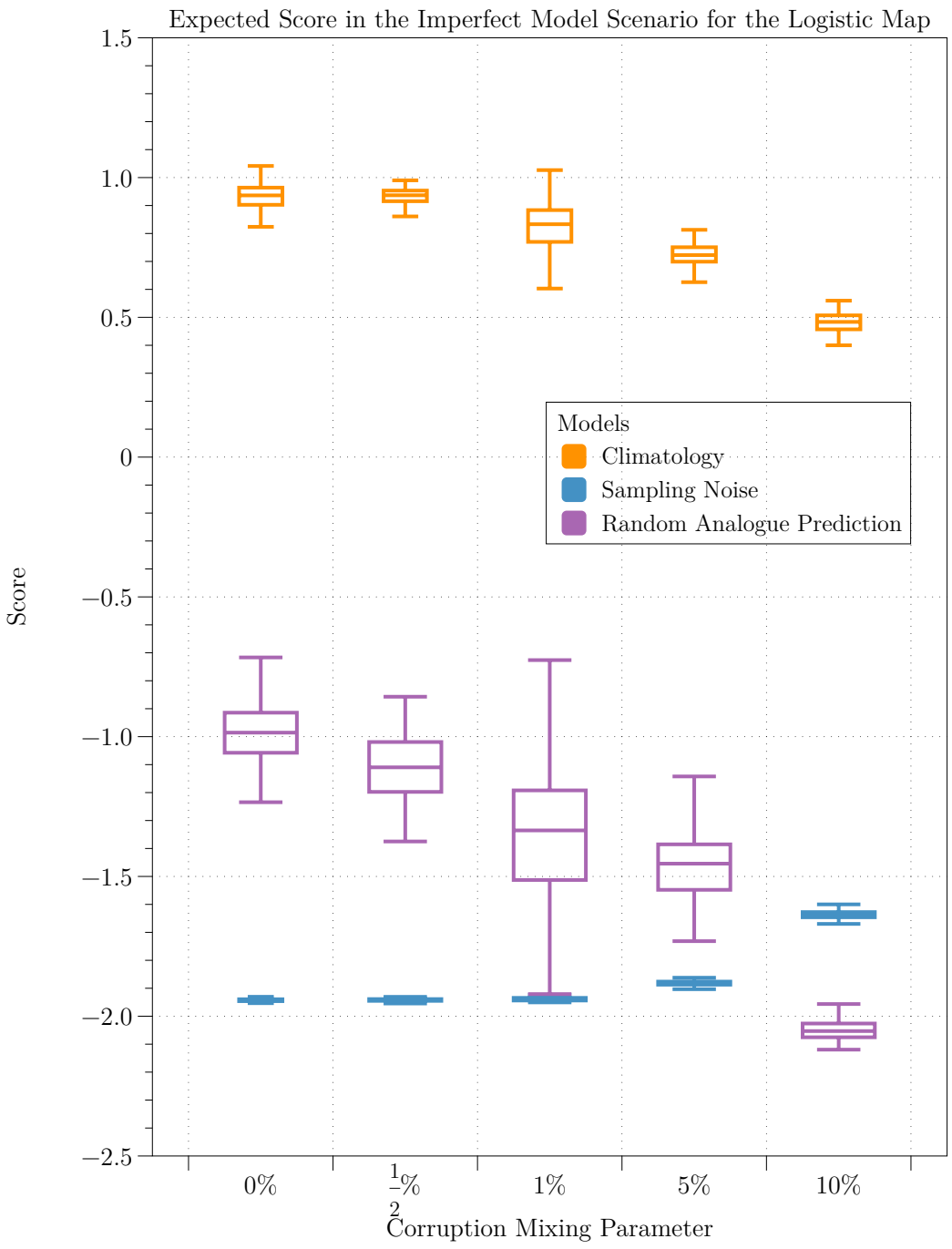


Figure 6.8: For various small values of the corruption mixing parameter, each model is applied to simulations from the logistic map, to generate a box plot estimating the density of the expected score. The numerical Bayesian solution is omitted. As the corruption increases, the random analogue prediction model outperforms the sampling-noise model.

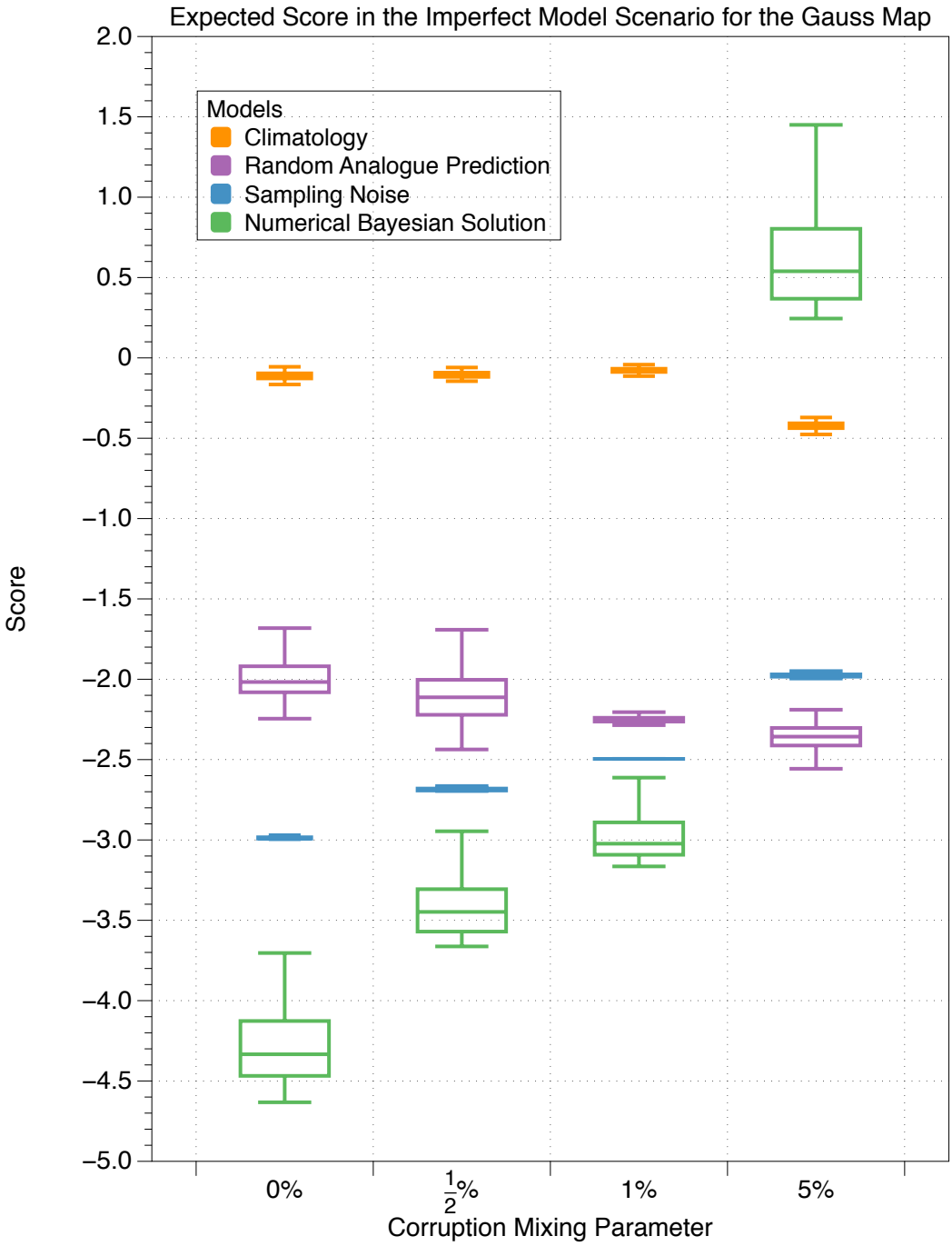


Figure 6.9: For various small values of the corruption mixing parameter, each model is applied to simulations from the Gauss map, to generate a box plot estimating the density of the expected score. As the corruption increases to 5%, the random analogue prediction model outperforms the sampling noise model while the Bayesian model underperforms relative to all other models.

Chapter 7

Conclusion

7.1 Making and Evaluating Predictions

This thesis considers the problem of making and evaluating predictions from models and a sequence of measurements. Once a sequence of measurements is collected, we first propose one or more statistical models to analyse the sequence of measurements. A statistical model is designed by specifying conditional probabilistic relationships between the measurements and model parameters. In the first chapter, we considered several different models and model structures, and reviewed the principles of statistical inference.

The Core Genome

In the next chapter, we introduced decision problems. As an example of a statistical decision, we decide which genes from an observed set of genes are to be considered core genes. This model is applied to a variety of genomic datasets, and the resulting list of core genes is used to motivate further biological analysis.

Prediction as a Decision

We frame the problem of prediction as a decision problem. Proper scoring rules are derived as appropriate loss functions, obtained by generalising the action from a point to a probability density. Next, a general construction of proper scoring rules, based on an arbitrary convex function, is presented. This indicates that there are any number of proper scoring rules, though it remains to be shown whether or not some are more useful than others.

To establish the link to parameter inference, we consider the limiting theory of the minimum score estimate. The derivative of a proper scoring rule yields an estimating function, which is used to produce the minimum score estimate. Under the logarithmic score for particular models, the minimum score estimate is equivalent to the maximum likelihood estimate. The limiting behaviour of the minimum score estimate procedure is utilised in deriving the classical model selection criteria.

Evaluating Predictions

We select the model that minimises the prequential score. A comparison between the prequential score and classical model selection criteria is given. Model selection using the Bayes factor is shown to be equivalent to comparing the prequential logarithmic score, suggesting the generalisation to using a prequential score with arbitrary proper scoring rules. Model selection using the classical Frequentist criterion proceeds by selecting the model that has the smallest estimated score of a future measurement. A generalisation to arbitrary proper scoring rules is given, though it is only valid for models with conditionally independent and identically distributed measurements. Using the prequential score as the model selection criterion allows for any collection of statistical models to be scored and compared.

Chaotic Dynamical Systems

Lastly, we consider the chaotic logistic and Gauss maps as examples of dynamical systems that exhibit complicated behaviour. A numerical approximation to the Bayesian solution is derived and compared to a few other models. We show, using proper scoring rule analysis, that the numerical Bayesian solution outperforms the other models in the perfect model scenario. In the imperfect model scenario, however, the numerical Bayesian solution does not perform nearly as well as other models, suggesting that complicated dynamical models may fail when the measurements are not exactly generated using the model assumptions.

7.2 Predictive Modelling

Statistical models are employed to describe natural systems. Often, these models involve a large number of variables interacting in various and complicated ways. It may be infeasible to establish a set of equations from fundamental physics as a forward model, let alone solve the inverse problem. To analyse the behaviour of a complicated natural system, simplifications in the representations of their structure are necessary, if we are hoping to capture at least some aspects of the underlying phenomenon. A statistical model is just a set of equations postulating conditional probabilistic relationships between the variables. The types of equations that could be used range from time-dependent, partial differential equations aiming to describe, explicitly, the underlying dynamics, to forecasting strategies that exploit patterns in the measurements. These are all models that can be evaluated within the underlying Bayesian framework.

Models can be altered by adding and removing assumptions through modification of the variables and their relationships. It is not necessary to describe the fundamental behaviour of the underlying phenomenon, though understanding aspects may lead to better predictions. If some alteration improves the predictive power of a model, it is considered reasonable and legitimate.

This thesis suggests a relationship between proper scoring rules and robust statistics. Further investigation into the connections between these areas may allow for a formal description of model robustness.

This thesis describes a formal probabilistic approach to making and evaluating predictions. In everyday life, personal analogues of statistical models are used informally by everyone. We make a prediction of the outcome of some uncertain event, conditional on our observations and prior knowledge, and then make a decision on an action, based on our own personal analysis.

Bibliography

Luis A Aguirre and Christophe Letellier. Modeling nonlinear dynamics and chaos: a review. *Mathematical Problems in Engineering*, 2009, 2009.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

Dario Albarello and Vera D'Amico. Testing probabilistic seismic hazard estimates by comparison with observations: an example in italy. *Geophysical Journal International*, 175(3): 1088–1094, 2008.

Siddharth Arora and James W Taylor. Short-term forecasting of anomalous load using rule-based triple seasonal methods. 2013.

Siddharth Arora, Max A Little, and Patrick E McSharry. Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the us gross national product. *Studies in Nonlinear Dynamics and Econometrics*, 17(4):395–420, 2013.

J.M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3): 686–690, 1979.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 2001.

D. Bernoulli. The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. *Acta Academiae Petropolitanae*, pages 3–33, 1777.

E. Borel. *Probabilities and Life*. Dover Publications, 1962.

G.E.P Box. *Robustness in the strategy of scientific model building*. University of Wisconsin-Madison, Mathematics Research Center, 1979.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Christopher L Bremer and Daniel T Kaplan. Markov chain monte carlo estimation of nonlinear dynamics from time series. *Physica D: Nonlinear Phenomena*, 160(1):116–126, 2001.

G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Jochen Bröcker and Leonard A Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C Wilcox. Extreme-scale uq for bayesian inverse problems governed by pdes. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, page 3. IEEE Computer Society Press, 2012.

Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2002.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

Simon L Cotter, Masoumeh Dashti, and Andrew M Stuart. Approximation of bayesian inverse problems for pdes. *SIAM Journal on Numerical Analysis*, 48(1):322–345, 2010.

David Roxbee Cox. *Principles of statistical inference*. Cambridge University Press, 2006.

Daniel R Creveling, James M Jeanne, and Henry DI Abarbanel. Parameter estimation using balanced synchronization. *Physics Letters A*, 372(12):2043–2047, 2008.

Nicholas J Croucher, Jonathan A Finkelstein, Stephen I Pelton, Patrick K Mitchell, Grace M Lee, Julian Parkhill, Stephen D Bentley, William P Hanage, and Marc Lipsitch. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6): 656–663, 2013.

A Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292, 1984.

A Philip Dawid. Prequential analysis. *Encyclopedia of Statistical Sciences*, 1997.

A Philip Dawid and Vladimir G Vovk. Prequential probability: principles and properties. *Bernoulli*, pages 125–162, 1999.

AP Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

Philip Dawid, Monica Musio, and Laura Ventura. Minimum scoring rule inference. *arXiv preprint arXiv:1403.3920*, 2014.

Bruno De Finetti. *Theory of probability*. CUP Archive, 1970.

M.H. DeGroot. *Optimal statistical decisions*. McGraw-Hill Series in Probability and Statistics, 1970.

Paul Adrien Maurice Dirac. The principles of quantum mechanics. *The International Series of Monographs on Physics, Oxford: Clarendon Press, 1947*, 1, 1947.

Hailiang Du and Leonard A Smith. Parameter estimation through ignorance. *Physical Review E*, 86(1):016213, 2012.

Richard M Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

C.F. Gauss. *Theoria combinationis observation erroribus minimis obnoxiae*. Dietrich, 1821.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

VP Godambe. Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63(2):277–284, 1976.

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

Patrick J Heagerty and Thomas Lumley. Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, 95(449): 197–211, 2000.

Nils Lid Hjort, CC Holmes, Peter Müller, and Stephen G Walker. Bayesian nonparametrics. *AMC*, 10:12, 2010.

Paul Hoyningen-Huene. *Systematicity: The nature of science*. Oxford University Press, 2013.

Peter J Huber. *Robust statistics*. Springer, 2011.

E.T. Jaynes and G.L. Bretthorst. *Probability theory: the logic of science*. Cambridge Univ Pr, 2003.

Jooyoung Jeon and James W Taylor. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79, 2012.

Ian T Jolliffe and David B Stephenson. *Forecast verification: a practitioner's guide in atmospheric science*. Wiley, 2012.

Biju Joseph, Roland F Schwarz, Burkhard Linke, Jochen Blom, Anke Becker, Heike Claus, Alexander Goesmann, Matthias Frosch, Tobias Müller, Ulrich Vogel, et al. Virulence evolution of the human pathogen neisseria meningitidis by recombination in the core and accessory genome. *PLoS One*, 6(4):e18441, 2011.

Kevin Judd. Chaotic-time-series reconstruction by the bayesian paradigm: Right results by wrong methods. *Physical Review E*, 67(2):026212, 2003.

Kevin Judd. Failure of maximum likelihood methods for chaotic dynamical systems. *Physical Review E*, 75(3):036210, 2007.

Kevin Judd and Leonard Smith. Indistinguishable states: I. perfect model scenario. *Physica D: Nonlinear Phenomena*, 151(2):125–141, 2001.

T.H. Jukes and C.R. Cantor. Evolution of protein molecules. *Manmmalian Protein Metabolism*, pages 21–132, 1969.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

J.L. Klein. *Statistical visions in time: a history of time series analysis, 1662-1938*. Cambridge Univ Pr, 1997.

Ying Kong, Jennifer H Ma, Keisha Warren, Raymond SW Tsang, Donald E Low, Frances B Jamieson, David C Alexander, and Weilong Hao. Homologous recombination drives both sequence diversity and gene content variation in neisseria meningitidis. *Genome biology and evolution*, 5(9):1611–1627, 2013.

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

Ada Lau, Patrick McSharry, et al. Approaches for multi-step density forecasts with application to aggregated wind power. *The Annals of Applied Statistics*, 4(3):1311–1341, 2010.

A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des cometes*. F. Didot, 1805.

Dennis V Lindley. *Understanding uncertainty*. John Wiley & Sons, 2013.

Dennis Victor Lindley. *Introduction to Probability and Statistics. Part 2: Inference*. CUP Archive, 1965.

Max A Little, Patrick E McSharry, and James W Taylor. Generalized linear models for site-specific density forecasting of uk daily rainfall. *Monthly weather review*, 137(3), 2009.

Edward N Lorenz. Available potential energy and the maintenance of the general circulation. *Tellus*, 7(2):157–167, 1955.

Edward N Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4):636–646, 1969.

Jun Lu, Alexios Vlamis-Gardikas, Karuppasamy Kandasamy, Rong Zhao, Tomas N Gustafsson, Lars Engstrand, Sven Hoffner, Lars Engman, and Arne Holmgren. Inhibition of bacterial thioredoxin reductase: an antibiotic mechanism targeting bacteria lacking glutathione. *The FASEB Journal*, 27(4):1394–1403, 2013.

Patrick E McSharry and Leonard A Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Physical review letters*, 83(21):4285–4288, 1999.

Renate Meyer and Nelson Christensen. Bayesian reconstruction of chaotic dynamical systems. *Physical Review E*, 62(3):3535, 2000.

Yves Meyer, Bob B Buchanan, Florence Vignols, and Jean-Philippe Reichheld. Thioredoxins

and glutaredoxins: unifying elements in redox biology. *Annual review of genetics*, 43:335–367, 2009.

DN Mukhin, AM Feigin, EM Loskutov, and Ya I Molkov. Modified bayesian approach for the reconstruction of dynamical systems from time series. *Physical Review E*, 73(3):036211, 2006.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Allan H Murphy and Robert L Winkler. Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79(387):489–500, 1984.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Anthony O'Hagan and Jonathan J Forster. *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*, volume 2. Arnold, 2004.

F Paparella, A Provenzale, Leonard A Smith, C Taricco, and R Vio. Local random analogue prediction of nonlinear processes. *Physics Letters A*, 235(3):233–240, 1997.

Pierre Pinson. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):555–576, 2012.

Pierre Pinson and George Kariniotakis. Conditional prediction intervals of wind power generation. *Power Systems, IEEE Transactions on*, 25(4):1845–1856, 2010.

K.R. Popper. *The Logic of Scientific Discovery*. Basic Books (New York), 1959.

Peter Poulsen and Kaj Frank Jensen. Three genes preceding¡ i¿ pyre¡/i¿ on the¡ i¿ escherichia coli¡/i¿ chromosome are essential for survival and normal cell morphology in stationary culture and at high temperature. *Research in microbiology*, 142(2):283–288, 1991.

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Wiley, 1951.

Jingshan Ren, Masayo Kotaka, Michael Lockyer, Heather K Lamb, Alastair R Hawkins, and David K Stammers. Gtp cyclohydrolase ii structure and mechanism. *Journal of Biological Chemistry*, 280(44):36912–36919, 2005.

Vijay K Rohatgi and AK Md Ehsanes Saleh. *An introduction to probability and statistics*, volume 910. John Wiley & Sons, 2011.

V.K. Rohatgi and A.K.M.E. Saleh. *An introduction to probability and statistics*. Wiley, 2009.

Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-Hill New York, 1964.

Simo Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

Claudia Schillings and Christoph Schwab. Sparse, adaptive smolyak quadratures for bayesian inverse problems. *Inverse Problems*, 29(6):065011, 2013.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

Claude Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(1):379–423, 1948.

J McLean Sloughter, Tilmann Gneiting, and Adrian E Raftery. Probabilistic wind speed forecasting using ensembles and bayesian model averaging. *Journal of the American Statistical Association*, 105(489):25–35, 2010.

Dmitry A Smirnov, Vladislav S Vlaskin, and Vladimir I Ponomarenko. Estimation of parameters in one-dimensional maps from noisy chaotic time series. *Physics Letters A*, 336 (6):448–458, 2005.

L.A. Smith, M.C. Cuéllar, H. Du, and K. Judd. Exploiting dynamical coherence: A geometric approach to parameter estimation in nonlinear models. *Physics letters A*, 374(26):2618–2623, 2010.

A.M. Stuart. Inverse problems: a bayesian perspective. *Acta Numerica*, 19(1):451–559, 2010.

Kei Takeuchi. Distribution of information statistics and criteria for adequacy of models. *Math. Sci*, 153:12–18, 1976.

A. Tarantola. Popper, bayes and the inverse problem. *Nature Physics*, 2(8):492, 2006.

James W Taylor. Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3):534–549, 2012.

Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, 2005.

Constantino Tsallis et al. *Introduction to nonextensive statistical mechanics*. Springer, 2009.

Andries J van Tonder, Shilan Mistry, James E Bray, Dorothea MC Hill, Alison J Cody, Chris L Farmer, Keith P Klugman, Anne von Gottberg, Stephen D Bentley, Julian Parkhill, et al. Defining the estimated core genome of bacterial populations using a bayesian decision model. *PLoS computational biology*, 10(8):e1003788, 2014.

Cristiano Varin, Nancy Margaret Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

Robert L Winkler. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 64(327):1073–1078, 1969.

Ziheng Yang. *Computational molecular evolution*, volume 284. Oxford University Press Oxford, 2006.