New focused approaches to topics within model selection and approximate Bayesian inversion

Martin Jullum

Dissertation presented for the degree of
Philosophiae Doctor (PhD)

Department of Mathematics
University of Oslo
December 2015

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1732

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Preface

This thesis comprises the work conducted during three years as a PhD student at the Department of Mathematics at the University of Oslo (UiO). I will be looking back at these years as an enjoyable phase of my life, in which I learned a lot. The period has also been tough, though – possibly even tougher than I expected it to be. With the finished thesis in my hand, there is however no doubt it was worth it. Still, I have to agree with the author Joseph Epstein who noted that it is a lot better to *have written*, than to actually *be writing*. The final product constituting my PhD thesis is inarguably positively correlated with the original project description, although the ρ is far from one. Through these years I have been working with methodology within a broad range of fields across the science of statistics. Among them are approximate Bayesian inference, asymptotic theory, Bayesian statistics, copulae, density estimation, frequentist statistics, functional differentiation, Gaussian distribution theory, geostatistics, inverse problems, Markov chain Monte Carlo (MCMC), model averaging, model selection, spatial statistics, stochastic process theory, survival analysis, time series modelling – and I even had to learn a little bit of geophysics, petrophysics and rock physics. I do by no means claim to master all these subjects, but I have learned a fair amount about all of them, and for that I feel incredibly lucky.

Upon completing this thesis, I am deeply indebted to my two supervisors Nils Lid Hjort and Odd Kolbjørnsen. I am truly grateful for how you inspired me, and the eagerness you showed while working with the various projects. I will sincerely like to thank you both for that. You also supported me and made it possible for me to spend the autumn of 2014 at Stanford University, visiting Paul Switzer at the Department of Statistics. Paul was an outstanding host during some incredible months over there – our delightful academic and non-academic discussions will not be easily forgotten. Being founded by Statistics for Innovation (SFI^2), a centre for research-based innovation, I was lucky enough to be awarded with two offices in Oslo; one at the campus at Blindern and one at the Norwegian Computing Center (NR) at Forskningsparken. Without even being employed at NR, I was very well taken care of and included in the SAND group. I am thankful for that additional dimension and opportunity to learn, and for being exposed to the weekly dose of Thursday-buns – that will be missed! I would also like to thank all my colleagues, both at the statistics group at UiO and the SAND group at NR. Special thanks go to my 'roommates' Marie at NR and Reinaldo at UiO for all our inviting discussions and fascinating conversations, and to Gudmund Horn Hermansen for co-authoring one of the papers in the thesis. Finally, I would like to thank my friends, family, and 'family-in-law' for filling my life with joy – especially my wonderful Elin for putting up with me, supporting and understanding me, even though I know you really wished I was rather spending those late evenings and weekends with you.

Oslo, December 2015
Martin Jullum

List of papers

Paper I

JULLUM, M. & HJORT, N. L. (2016). Parametric or nonparametric: The FIC approach. *Submitted for publication in Statistica Sinica*

Paper II

JULLUM, M. & HJORT, N. L. (2015a). What price semiparametric Cox regression? *Submitted for publication in Scandinavian Journal of Statistics*

Paper III

HERMANSEN, G. H., HJORT, N. L. & JULLUM, M. (2015). Parametric or nonparametric: The FIC approach for stationary time series. *Technical report, Department of Mathematics, University of Oslo*

Paper IV

JULLUM, M. & KOLBJØRNSEN, O. (2016). A Gaussian-based framework for local Bayesian inversion of geophysical data to rock properties. *Accepted for publication in Geophysics*

Contents

Preface i

List of papers ii

1 Introduction 1

2 Parametric and nonparametric modelling and asymptotics 3
2.1 Fully observed independent observations 3
2.2 Survival analysis 7
2.3 Time series analysis 10

3 Model selection 12
3.1 Classical information criteria for parametric models 13
3.2 Parametric or nonparametric? 14
3.3 Focus parameter and the focused information criterion (FIC) 17

4 Geophysical/rock physical inversion and approximate Bayes 21
4.1 Geophysical data and rock physics 21
4.2 The inverse problem and the Bayesian approach 22
4.3 General approximate Bayesian inference 25
4.4 Approximate Bayesian inference within the geosciences 29

5 Summary of papers 31
5.1 Paper I 31
5.2 Paper II 32
5.3 Paper III 32
5.4 Paper IV 33

6 Discussion 34
6.1 Extensions and further work 34
6.2 New vs. original FIC 35
6.3 Localising the inverse problem 37
6.4 The leitmotifs and frequentist vs. Bayesian statistics 38

References 39

Papers I-IV with supplementary material 49

1 Introduction

Due to the pervasive technological developments the last couple of decades, the modern society is overwhelmed with information or data of various types and forms. Parallel to this, analysis of data keeps spreading to new industries and sciences, which are eager to learn and gain scientific data-based insight. As a consequence, the complexities of models and analytical methods have increased rapidly in attempts to extract new knowledge from the data. The scientific challenges one currently faces are therefore somewhat different from those typically encountered only a few decades ago. Back then, the data were smaller and the incentives for pushing the limit of model complexity where not as strong as today. The new challenges concern both collection and storing of data, but perhaps more importantly, how to extract the interesting features from the data. Universal statistical approaches will in many settings be too general to capture and exploit the important features of the data. Such specific statistical challenges rather call for specialised, focused methodology which is specifically tuned towards certain inference tasks, for which scrutiny is considered valuable. Further, despite recent computational advances, computational handling of large amounts of data and complex models remains a bottleneck. Thus, some kind of scientific simplification, approximation, or intelligent elimination of redundancy, is sometimes required to facilitate computationally feasible data analysis.

This thesis explores two topics related to specifically targeted statistical methodology, where asymptotic theory and other statistical rationale are utilised for approximation purposes. More precisely, new focused methodology for topics within statistical model selection (Papers I-III) and approximate Bayesian inversion (Paper VI), have been developed. Despite both being pieces of the puzzle of the new statistical frontier, statistical model selection and approximate Bayesian inversion may come across as two fairly different topics. Firstly, the former is mainly a frequentist problem, while the latter is fully Bayesian. Secondly, model selection is essential in the final stage of inference – once alternative statistical models are described and fitted, and one needs to decide which of the alternative conclusions that should be trusted. In approximate Bayesian inversion, on the other hand, all efforts are used to get an idea of the features that caused the observed data. Thirdly, a partly unspecified semi-/nonparametric candidate model plays a key role in Papers I-III, while the Bayesian framework in Paper IV is fully specified. The four papers also deal with quite different types of data: (I) The simplest type of data, where the observations are independent, fully observed and have the same distribution (i.i.d. data); (II) partly observed data where the distribution of the observations may depend on covariate values (censored survival or event history data with covariates); (III) data consisting of repeated observations of a process evolving over time (time series data); and (IV) high dimensional, highly correlated, and noisy application specific geophysical measurement data, typically being processed by specialists before they are presented to the statistician.

In view of the specific scientific problems being addressed, the four papers comprising this thesis appear to span quite broadly. They are however more closely connected in a more general context, as methodological contributions to the statistical science fundamentally governed by being (A) *focused*, and (B) based on careful statistical *approximations*. In terms of (A), all projects are in some sense focused towards a particular pre-specified task defined as the ultimate goal of analysis. In terms of (B), both the model selection projects and the approximate inversion project concern finding the right amount of structure – i.e. to what degree should one approximate? These running themes will be elaborated more upon in Section 6, once all concepts and contributions of the papers have been thoroughly presented. Note however, already at this point, the following general remark related to (B): When modelling a data generating process by a (parametric) statistical model, one is in practice simplifying or approximating a true unknown data generating process. Apart possibly from a few cases, like quantities within physics having an exact Gaussian distribution, statistical models are not exact. Still, statistical models can give insight and precise knowledge from data. Some assign this characteristic to the perhaps overly simplified quote of George Box: “All models are wrong, but some are useful.” Having an unusual strong interest for maps myself, I personally prefer an analogue between models and maps. Maps are two-dimensional simplifications of the real world, but a map of the world is not wrong just because one’s backyard are not to be found there – or as John Michael Steele states regarding the same analogue: “If I say that a map is wrong, it means that a building is misnamed, or the direction of a one-way street is mislabelled. I never expected my map to recreate all of physical reality, and I only feel ripped off if my map does not correctly answer the questions that it claims to answer. My maps of Philadelphia are useful. Moreover, except for a few that are out-of-date, they are not wrong.”

The two main aims of the thesis concern development of fundamentally new statistical methodology. The aims are:

- To bridge the gap between the use of semi-/nonparametric statistical modelling procedures and fully parametric alternatives by developing model selection procedures for comparing these fundamentally different models – a challenge which has not been properly addressed in the literature before.

- To develop computationally efficient, yet sufficiently precise, methodology for handling Bayesian inversion problems within the geosciences – whose earlier proposed solutions are either too rough and inaccurate, or too slow for real large-scale applications.

The remainder of the thesis is structured as follows: Section 2 introduces parametric and nonparametric modelling approaches for fully observed independent observations, and within survival- and time series analysis. Section 3 gives an introduction to model selection for parametric models, discussing also the difficulties of including semi- and nonparametric models. Section 4 introduces the data and the problems commonly dealt with in the geosciences, in addition to the inverse problem. A brief survey of existing methodology for conducting approximate Bayesian inference is also presented. Section 5 consists of summaries of each of the four papers constituting this thesis. Finally, Section 6 gives pointers to further work, summarises the higher level contribution of the thesis, and discusses a few selected topics in greater depth.

2 Parametric and nonparametric modelling and asymptotics

There are roughly speaking two main approaches to performing statistical modelling of an unknown distribution G: Parametrics and nonparametrics. A parametric modelling approach is characterised by having a model space restricted to a family of (cumulative) distributions F_θ indexed and described completely by a finite dimensional parameter θ, which is estimated based on the data. The precise family is typically chosen for reasons of simplicity, tradition, prior knowledge, computational efficiency or eased interpretation. The approach involves two steps: (a) deciding upon a parametric family of distributions F_θ and (b) estimating the parameter θ. While part (a) naturally depends on the application and data available, part (b) is methodologically standardised. That is, the same type of estimation procedures are typically used independently of the type of application and available data. This is one of the main advantages of parametric modelling approaches, in addition to the natural interpretations which they often carry with them.

A nonparametric modelling approach may, despite its name, be thought of as a parametric modelling approach where the θ-parameter is infinite dimensional. In this regard, the nonparametric model has an infinite number of parameters rather than none. The somewhat counterintuitive term 'nonparametric' stems merely from the fact that in practice, no parameters are being estimated. The approach is in some sense distribution free and does not restrict the space of available models to the extent that parametric approaches do, even though certain restrictions may be imposed in practice. In contrast to the parametric modelling approach, the unstructured nature of the nonparametric modelling approach imposes procedures restricted to certain data types and settings. The underlying ideas of the nonparametric methods are however often conceptually alike, even for fundamentally different data types. The main idea is to let the data speak for themselves. Freedom and insurance against misspecification are the main advantages of the nonparametric modelling approach.

2.1 Fully observed independent observations

Independent identically distributed data are possibly the simplest data type for which statistical inference is performed. A collection of observations $Y_1, \ldots, Y_n$ are then assumed to stem from a common unknown (cumulative) distribution G, having density or probability mass function g. This distribution may either be modelled parametrically by $F_{\widetilde{\theta}_n} = F(\cdot; \widetilde{\theta}_n)$ for a data dependent fitted value $\widetilde{\theta}_n$ of the p-dimensional parameter θ; or nonparametrically by some $\widetilde{G}_n$.

2.1.1 Parametric

The by far most frequently applied procedure for specifying $\widetilde{\theta}_n$ in the parametric modelling setting, is that of maximum likelihood. The maximum likelihood estimator $\widetilde{\theta}_n = \widehat{\theta}$, developed and largely popularised by R.A. Fisher (Aldrich, 1997) in the 1910s, is defined as

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L_n(\theta) = \underset{\theta}{\operatorname{argmax}}\ \ell_n(\theta), \tag{2.1}$$

where $L_n(\theta) = \prod_{i=1}^n f(Y_i;\theta)$ is the likelihood and $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^n \log f(Y_i;\theta)$ is the log-likelihood. If the unknown distribution G lies within the parametric family F_θ, then there exists a true value θ_{true} of θ, and the parametric model is said to be correct. The classical textbook result is then that

$$\widehat{\theta} \to_p \theta_{\mathrm{true}} \text{ and } \sqrt{n}(\widehat{\theta} - \theta_{\mathrm{true}}) \to_d \mathrm{N}(0, J(\theta_{\mathrm{true}})^{-1}) \text{ as } n \to \infty, \tag{2.2}$$

where $J(\theta) = -\mathrm{E}[\partial^2 f(Y_i;\theta)/\{\partial\theta\partial\theta^{\mathrm{t}}\}]$ is the well-known Fisher information matrix. This results is rather restrictive, however, as G is seldom within the parametric family F_θ. Nonetheless, a similar asymptotic result due to White (1982) holds also when the parametric class is misspecified. Of course, then no θ_{true} exists; instead a so-called least false parameter value θ_0 takes its place. This least false parameter value is defined as the value of θ that minimises the Kullback–Leibler divergence (Kullback & Leibler, 1951) from g to the family f_θ. For continuous densities this is defined as

$$\mathrm{KL}(g, f_\theta) = \int g(y) \log \frac{g(y)}{f(y;\theta)}\,\mathrm{d}y = \int g(y)\log g(y)\,\mathrm{d}y - \int g(y) \log f(y;\theta)\,\mathrm{d}y. \tag{2.3}$$

With this definition of θ_0 and additional rather mild regularity conditions (see e.g. White (1982)) the result in (2.2) generalises to

$$\widehat{\theta} \to_p \theta_0 \text{ and } \sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}(0, J(\theta_0)^{-1}K(\theta_0)J(\theta_0)^{-1}), \tag{2.4}$$

where $K(\theta) = \mathrm{Var}\{\partial f(Y_i;\theta)/\partial\theta\}$. This result allows for model robust inference with parametric models. One may for instance establish confidence intervals for functions of θ with the correct asymptotic coverage even when the assumed parametric distribution is incorrect.

Although the maximum likelihood approach is the most popular method for estimating θ due to its intuitive interpretation and favourable properties, there exist several other procedures. The more general concept of M-estimators generalises maximum likelihood estimation, see e.g. van der Vaart (2000, Ch. 5). The method of moments, dating all the way back to Pearson (1895), is different in spirit and typically easier to handle analytically. With the computational power of modern times, the incentives for using this simpler estimator are however weakened compared to the theoretically superior maximum likelihood and M-estimation procedures. The framework of generalised method of moments (Hansen, 1982) may be viewed as a generalisation of the method of moments, and (in some sense) also maximum likelihood and M-estimation (Imbens, 2002). The method does not require a full likelihood, but only a set of moment conditions related to θ. This makes it suitable for estimation in certain semiparametric models, particularly within economy and finance where models are often specified solely in terms of such conditions.

2.1.2 Nonparametrics

The most natural nonparametric estimator for G is the empirical distribution (function) $\widetilde{G}_n(y) = \widehat{G}(y) = n^{-1}\sum_{i=1}^{n}\mathbf{1}_{\{Y_i \le y\}}$, which essentially assigns equal weight $1/n$ to each of the observations. Profound studies have through the history revealed a number of powerful and useful theoretical properties of this estimator. The law of large numbers (Bernoulli, 1713; de Moivre, 1733; Khinchin, 1929) shows pointwise consistency of this estimator, i.e. $\widehat{G}(y) \to_p G(y)$ as $n \to \infty$ (holding also with almost sure convergence $\to_{a.s.}$). This result is made uniform in terms of the uniform norm $\sup_y \|\cdot\|$ by what is often referred to as the 'fundamental theorem of statistics', namely the Glivenko–Cantelli theorem (Glivenko, 1933; Cantelli, 1933) stating that $\sup_y \|\widehat{G}(y) - G(y)\| \to_{a.s.} 0$ as $n \to \infty$. Further, the pointwise asymptotic distribution of $\sqrt{n}(\widehat{G}(y) - G(y))$ is zero-mean Gaussian by the central limit theorem. This result is also made uniform by the so-called Donsker theorem which states that $\sqrt{n}(\widehat{G}(\cdot) - G(\cdot))$ process converges (weak convergence) to a zero-mean Gaussian process $Z(\cdot)$ with covariance function $\Sigma(x,y) = G(\min(x,y)) - G(x)G(y)$, also commonly referred to as a G-Brownian bridge. Furthermore, a result known as the functional delta method (van der Vaart, 2000, Theorem 20.8) (a generalisation of the familiar delta method for vectors (van der Vaart, 2000, Theorem 3.1)), may be used to extend the Donsker theorem to smooth functionals of G, i.e. functions taking the process G as input. For a sufficiently smooth functional T, the result says that $\sqrt{n}\{T(\widehat{G}) - T(G)\} \to_d T'(Z)$, where T' is another functional known as the functional derivative at G. In particular, $T'(Z)$ is a Gaussian process in the relevant space. The precise smoothness condition required, is that of Hadamard or compact differentiability; see e.g. van der Vaart (2000, p. 297) for a precise definition. In particular, this very general result provides the limit distribution for any continuously differentiable function of one or more quantiles $G^{-1}(p)$ and means $\int h(y)\,\mathrm{d}G(y)$.

The functional delta method does not work for $g(y)$, however. For discrete distribution, the probability mass function $g(y) = \Pr(Y_i = y)$ is also easy to estimate nonparametrically by the size of the discrete jumps in $\widehat{G}$. For continuous distributions, however, the density $g(y) = \partial G(y)/\partial y$ is more troublesome. The histogram, apparently first introduced by Pearson (1895), is often used as a visual tool for inspecting the density. The histogram divides the data range into bins (starting at some y_0): $B_j = [y_0 + (j-1)h, y_0 + jh), j = 1, \ldots$ for some bin-width h, hereby called the bandwidth, and uses

$$\widetilde{g}(y) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j}\mathbf{1}_{\{(Y_i \in B_j,\, y \in B_j\}}, \tag{2.5}$$

resulting in a bar diagram showing the frequencies of the data falling into each bin once plotted. If the bandwidth $h = h_n$ decreases to zero slower than n^{-1} as the sample size increases, i.e. if $h_n \to 0$ while $h_n n \to \infty$ as $n \to \infty$, then $\widetilde{g}(y)$ is a consistent estimator of $g(y)$ (Härdle, 1991, p. 15). In addition to being a discontinuous step-function, which is somewhat unappealing when estimating a continuous distribution, the optimal convergence rate of $O(n^{-2/3})$ for this estimator is not impressive (Wasserman, 2006, Theorem 6.11). The convergence rate is faster for the related naïve density estimator

$$\widehat{g}(y) = \#\{Y_i \in [y-h, y+h); i = 1, \dots, n\}/(2nh),$$

for some small bandwidth h. Even though this estimator has a faster convergence rate and keeps the same consistency properties, it is not very robust and efficient, and is still discontinuous; see e.g. Silverman (1986) for further details. Writing the estimator as

$$\widehat{g}(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{y - Y_i}{h}\right), \tag{2.6}$$

with $K(x) = \frac{1}{2}\mathbf{1}_{\{|x| \leq 1\}}$, motivates using a smooth kernel function K, rather than the 'boxcar' kernel which gave the naïve density estimator. Under weak conditions on the kernel function, the general class of estimators of the form (2.6) has an optimal convergence rate of $O(n^{-4/5})$ and estimates $g(y)$ consistently, as long as $h = h_n \to 0$ and $h_n n \to \infty$ as $n \to \infty$ (Lehmann, 1998, Theorem 6.4.1 and Corollary 6.4.1). Typical choices of the kernel function are the standard normal distribution $K(x) \propto \exp(-x^2/2)$, the Epanechinkov kernel $K(x) \propto (1 - x^2)\mathbf{1}_{\{|x| \leq 1\}}$, and the tricube kernel $K(x) \propto (1 - |x|^3)^3 \mathbf{1}_{\{|x| \leq 1\}}$. The bandwidth is often set using cross-validation (see e.g. Jones et al. (1996)), or using plug-in selectors like Silverman's 'rule of thumb' (Silverman, 1986), corresponding to $h_n = 1.059\widehat{\sigma} n^{-1/5}$ for $\widehat{\sigma}$ the empirical standard deviation.

Motivated by the above smoothed density estimation procedures, authors have also suggested using

$$\widetilde{G}(y) = \int_{-\infty}^{y} \widehat{g}(x)\,\mathrm{d}x, \tag{2.7}$$

with $\widehat{g}$ as in (2.6), as a smooth estimator for the distribution function G. This estimator is asymptotically equivalent to $\widehat{G}(y)$ in the sense that $\sqrt{n}(\widehat{G}(\cdot) - G)$ and $\sqrt{n}(\widetilde{G}(\cdot) - G)$ converges in distribution to the same limit process (Watson & Leadbetter, 1964).

2.1.3 Parametric and nonparametric regression

Consider now the more general regression setting where, in addition to Y_i, covariate vectors $X_i, i = 1, \dots, n$ are also available and assumed to have an influence on the distribution of the data. That is, the conditional distribution $G(\cdot|x)$ depends on the covariates having value x. For parametric modelling it is then more natural to describe the parametric family by $F(\cdot; \theta|x)$, resulting in a maximum likelihood estimator on the same form as (2.1), but with $L_n(\theta) = \prod_{i=1}^{n} f(Y_i; \theta|X_i)$ and an analogous definition for $\ell_n(\theta)$. The results in (2.2) and (2.4) still hold with the new definitions of $\ell_n(\theta)$, modulo slightly stronger regularity conditions and that θ_0 rather minimises the more general weighted Kullback–Leibler divergence

$$\mathrm{KL}(f_\theta, g) = \int\int g(y|x) \log \frac{g(y|x)}{f(y; \theta|x)}\,\mathrm{d}y\,\mathrm{d}C(x),$$

where C is the distribution of the covariates. For further details and discussion, see e.g. Claeskens & Hjort (2008a, Ch. 2).

There exist several different procedures for nonparametric regression. Most of them are however in principle analogous to those mentioned for density estimation above. In fact, the density estimation problem may be translated to a nonparametric regression problem such that all techniques applicable to the latter become available also for the former (Nussbaum, 1996; Brown et al., 2010). As these types of problems shall not concern us particularly, we will not go further into details on this topic. See e.g. Wasserman (2006, Ch. 5) for further concepts and details.

2.2 Survival analysis

In the survival analysis community, there is a strong tradition for using non- or semiparametric modelling approaches, as opposed to fully parametric ones. The reason for this is possibly the formers' intuitive setups and simple interpretations. In the covariate free case, one has data of the form $(T_i, D_i), i = 1, \ldots, n$ observed over a time window $[0, \tau]$ where T_i is the possibly censored survival time (or more generally time to some event), while D_i is the indicator of T_i being equal to the uncensored, but unobserved, event time $T_i^{(0)}$. For analysis of such data, the counting process $N_i(t) = \mathbf{1}_{\{T_i \leq t, D_i = 1\}}$ and the individual at-risk processes $Y_i(t) = \mathbf{1}_{\{T_i \geq t\}}$ for $i = 1, \ldots, n$, play key roles along with the martingales based on these processes $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha(s)\,\mathrm{d}s$, where $\alpha(s)$ is the hazard rate defined below. Denote by G the common distribution of the fully observed survival times $T_i^{(0)}$, which we shall here assume have a continuous event density g. For notational convenience we also assume there are no tied events.

For such data one is often interested in modelling the survival function $S(t) = 1 - G(t) = \Pr(T_i^{(0)} > t)$ and the hazard rate $\alpha(t) = g(t)/S(t)$, or the latter's cumulative $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. The Kaplan–Meier estimator is a nonparametric estimator of the survival function, first proposed by Böhmer (1912) and later re-introduced and popularised by Kaplan & Meier (1958). It takes the form

$$\widehat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{1}{\sum_{i=1}^n Y_i(T_j)}\right).$$

The Nelson–Aalen estimator (Nelson, 1969, 1972; Aalen, 1975, 1978) is a related nonparametric estimator of the cumulative hazard rate. It is given by

$$\widehat{A}(t) = \sum_{T_j \leq t} \frac{1}{\sum_{i=1}^n Y_i(T_j)}.$$

Under appropriate conditions, these are consistent and both $\sqrt{n}\{\widehat{S}(\cdot) - S(\cdot)\}$ and $\sqrt{n}\{\widehat{A}(\cdot) - A(\cdot)\}$ process converges to certain explicit zero-mean Gaussian processes, see e.g. Andersen et al. (1993, Ch. IV).

A perfectly valid alternative to the above type of nonparametric modelling is to rely on a parametric class of distributions, event densities or hazard rates. As the hazard rate plays such an important role in survival analysis, the parametric model families are often described in terms of families of hazard rates. The exponential distribution is the simplest family, having a constant hazard rate $\alpha_{\text{exp}}(t;\theta=\lambda)=\lambda$. The Weibull and Gompertz distributions are also popular extensions, having respecitve hazard rate functions $\alpha_{\text{wei}}(t;\theta=(\lambda,\gamma))=\gamma\lambda(\lambda t)^{\gamma-1}$ and $\alpha_{\text{gom}}(t;\theta=(\lambda,\gamma))=\lambda\exp(\gamma t)$. Inference for these parametric models typically proceeds via maximum likelihood estimation, similarly to the case for the fully observed data in Section 2.1. Since the observations are only partially observed, the likelihood takes a somewhat different form, however. The log-likelihood is in particular given by

$$\ell_n(\theta)=\sum_{i=1}^{n}\int_0^{\tau}\left[\log\alpha_{\text{pm}}(s;\theta)\,\mathrm{d}N_i(s)-Y_i(s)\alpha_{\text{pm}}(s;\theta)\right]\,\mathrm{d}s. \tag{2.8}$$

Note that unless the censoring mechanism is uninfluenced by θ, then (2.8) is not a true likelihood. This has no consequence for inference, however.

The maximum likelihood estimator $\widehat{\theta}$ specifies the precise form of the parametric hazard rate function and its cumulative. The precise formula for the survival function is further found via the link $S(t)=\exp\{-A(t)\}$. In principle any other parameter or function related to the distribution of the survival times may be found via similar transformations. Under appropriate conditions, the properties of the maximum likelihood estimator derived in Section 2.1, both under model conditions (Borgan, 1984) and under general misspecification (Hjort, 1992), hold also for this case.

Let us turn to the more general regression case, where also covariate vectors X_i are available for each individual. In the same manner as nonparametrics are popular for the covariate free case, semiparametrics are widely applied in the regression case, in particular due to Cox's partial likelihood formulation (Cox, 1972, 1975). In his proportional hazard regression setup, the hazard rate function is assumed to take the form

$$\alpha(t|x)=\alpha_0(t)\exp(x^{\mathrm{t}}\beta), \tag{2.9}$$

with β some vector of regression coefficients and $\alpha_0(t)$ some unspecified baseline hazard function. The β represents the parametric part of the model, while leaving the baseline hazard unspecified indeed makes (2.9) semiparametric. Since α_0 is completely unstructured, a full likelihood is not attainable. Due to the proportional form of (2.9), relative risks (or more precisely hazard ratios) $\alpha(t|x_2)/\alpha(t|x_1)$ are independent of the baseline hazard and thus fully parametric. Cox utilised this property to construct a relative risk type of partial likelihood which is independent of the baseline hazard α_0. The likelihood yields

$$L_{\text{partial},n}(\beta)=\prod_{i=1}^{n}\left\{\frac{\exp(X_i^{\mathrm{t}}\beta)}{R_n(T_i;\beta)}\right\}^{D_i}, \tag{2.10}$$

where $R_n(s;\beta)=\sum_{i=1}^{n}Y_i(s)\exp(X_i^{\mathrm{t}}\beta)$ is the 'cumulative risk' of all the individuals at time s. Since the only unknown quantity in (2.10) is the regression coefficients β, one may estimate β by maximising the partial likelihood. This estimator $\widehat{\beta}_{\text{cox}}=\text{argmax}_{\beta}L_{\text{partial},n}(\beta)$ is known as the maximum partial likelihood estimator. The exponential form of the proportional hazard setup also provides an intuitive and simple interpretation of the individual regression coefficients, as

$\exp(\beta_j)$ is indeed the hazard ratio between two individuals whose covariate information differs only by a unit of 1 in the j-th covariate.

The Breslow estimator (Breslow, 1972) is typically called upon when one wishes to estimate features of the underlying distribution which does not depend solely on the regression coefficients, but also on the baseline hazard α_0. Examples of such are cumulative hazards, survival functions, life-time quantiles and similar, possibly conditioned on certain covariate values. The Breslow estimator estimates the cumulative hazard function by

$$\widehat{A}_{\text{cox}}(t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{d}N_i(s)}{R_n(s; \widehat{\beta}_{\text{cox}})}.$$

Semiparametric estimates of the cumulative hazard and survival functions conditioned on some covariates given by x, are then given by respectively $\widehat{A}(t|x) = \widehat{A}_{\text{cox}}(t)\exp(x^{\mathrm{t}}\widehat{\beta}_{\text{cox}})$ and $\widehat{S}(t|x) = \exp\{-\widehat{A}(t|x)\}$. Under various conditions (including that the Cox model in (2.9) indeed holds), these estimators are consistent, while $\sqrt{n}\{\widehat{A}(\cdot\,|x) - A(\cdot\,|x)\}$ and $\sqrt{n}\{\widehat{S}(\cdot\,|x) - S(\cdot\,|x)\}$ have explicit zero-mean Gaussian process limits; see e.g. Andersen et al. (1993, Ch. VII.2.2).

An alternative to leaving the baseline hazard $\alpha_0(t)$ completely unspecified is to assume it takes a certain parametric form, say $\alpha_{0,\text{pm}}(t;\theta)$, typically corresponding to those mentioned for $\alpha_{\text{pm}}(t;\theta)$ in the covariate free case above. A fully specified likelihood with parameters (θ,β) may then be established. The resulting log-likelihood extends that of (2.8) and is given by[1]

$$\ell_n(\theta,\beta) = \sum_{i=1}^n \int_0^\tau \left[\{\log\alpha_{0,\text{pm}}(s;\theta) + X_i^{\mathrm{t}}\beta\}\,\mathrm{d}N_i(s) - Y_i(s)\alpha_{0,\text{pm}}(s;\theta)\exp(X_i^{\mathrm{t}}\beta)\,\mathrm{d}s\right].$$

This gives rise to fully parametric alternatives $\widehat{A}_{\text{pm}}(t|x) = A(t;\widehat{\theta})\exp(x^{\mathrm{t}}\widehat{\beta})$ and $\widehat{S}_{\text{pm}}(t|x) = \exp\{-\widehat{A}_{\text{pm}}(t|x)\}$ to the semiparametric estimators $\widehat{A}(t|x)$ and $\widehat{S}(t|x)$. As for the covariate free case, Borgan (1984); Hjort (1992) establish comforting asymptotic properties for these estimators under suitable conditions.

Similar relative risk type of partial and fully parametric likelihoods may also be formulated for other proportional hazard models, i.e. when the proportionality function r in $\alpha(t|x) = \alpha_0(t)r(x^{\mathrm{t}}\beta)$ is not log-linear in the β as with $r(a) = \exp(a)$. Examples are the regular linear form $r(a) = 1 + a$, or the logistic form $r(a) = \exp(a)/\{1+\exp(a)\}$. Although such formulations gives rise to new semiparametric and parametric estimators, the regression coefficients typically have less appealing interpretations and their asymptotic properties are more troublesome to handle (Prentice & Self, 1983). Such formulations are therefore only sporadically used in practice, see also Aalen et al. (2008, Ch. 4.1). Yet other semiparametric and fully parametric regression procedures appear to be used occasionally. Among them are Aalen's additive hazard regression (Aalen, 1989) which assumes that the conditional hazard $\alpha(t|x)$ is a linear combination of regression coefficients β, and Efron's parametric logistic regression approach for discretised survival times with grouped covariates (Efron, 1988).

[1] Once again, with no consequences for inferences, this is a true likelihood only if the censoring mechanism and covariate distribution are independent of the parameters θ and β.

2.3 Time series analysis

Time series analysis concerns the study and estimation of underlying features and mechanisms of processes observed over time. That is, analysis of an observed sequence of real valued random variables $Y_1, \dots, Y_n$, which we shall take as stationary and observed at discrete equidistant time points $t = 1, \dots, n$. However, in practice, time series processes often evolve over time, corresponding to a mean or trend function. Thus, we shall here assume that the time series $Y_1, \dots, Y_n$ has been initially detrended, and rather concentrate on modelling and analysing the dependencies of this detrended time series. When analysing such data, it is quite common to assume that the data generating process is Gaussian, in which case the covariance function $C(h) = \mathrm{Cov}(Y_t, Y_s), h = |s - t|$ determines the distribution of the time series completely.

There are essentially two parallel approaches or domains for analysing time series data. In the possibly most natural *time domain*, one works directly with the covariance function $C(h)$, and tries to estimate that 'directly' based on the observations. By Wold's theorem (Priestley, 1981, p. 222) the covariance function may also be represented as $C(h) = \int_{-\pi}^{\pi} \cos(\omega h)\, \mathrm{d}G(\omega)$ where G is the so-called spectral distribution, having the usual properties of a distribution function on $(-\pi, \pi)$. When G is everywhere differentiable, the spectral density $g(\omega) = \partial G(\omega)/\partial\omega$ exists for all $\omega \in (-\pi, \pi)$ and may be represented by

$$g(\omega) = \frac{C(0)}{2\pi} + \frac{1}{\pi}\sum_{h=1}^{\infty} C(h)\cos(\omega h), \quad \text{for } \omega \in (-\pi, \pi). \tag{2.11}$$

Thus, time series data may analogously be studied and described in terms of the spectral density $g(\omega)$, which by (2.11) is seen to be symmetric around $\omega = 0$. In this *spectral* or *frequency domain*, one works with the spectral density $g(\omega)$ and the spectral distribution $G(\omega) = \int_{-\pi}^{\omega} g(\omega)\, \mathrm{d}\omega$, trying to estimate and learn about these.

The most common parametric time series model is the autoregressive (AR) model. Defining $Y_t = 0$ for $t \le 0$, this models assumes that

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \varepsilon_t,$$

for p scalar parameters $\alpha_1, \dots, \alpha_p$, and an i.i.d. zero-mean error term ε_t. The more general autoregressive moving average (ARMA) model includes q additional scalar parameters $\beta_1, \dots, \beta_q$, and takes the form:

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t. \tag{2.12}$$

The α- and β-parameters may be represented by a parameter vector θ specifying a parametric form of the covariance function $C_{\mathrm{pm}}(h;\theta)$, which further describes the parametric forms of a spectral density $f(\omega;\theta)$ and distribution $F(\omega;\theta)$. The parameters in the AR-model may be properly estimated based on the so-called Yule-Walker equations, without making further distributional assumptions about the underlying time series process; see e.g. Brockwell & Davis

(1991, Ch. 8). For the more general ARMA-model, conditional and unconditional least squares type of estimation procedures may be used under such circumstances. Alternatively, if one is willing to make fully descriptive distributional assumptions on the underlying time series process, one may put up a full likelihood formula and estimate the parameters via e.g. maximum likelihood. Under the traditional Gaussian assumption, the log-likelihood takes the form

$$\ell_n(\theta) = -\tfrac{1}{2}\{n\log(2\pi) + \log|\Sigma_n(\theta)| + (Y_1,\ldots,Y_n)\Sigma_n(\theta)^{-1}(Y_1,\ldots,Y_n)^{\mathrm{t}}\}, \tag{2.13}$$

where $\Sigma_n(\theta)$ is the covariance matrix of the observations, having elements $C_{\mathrm{pm}}(|s-t|;\theta)$ for $s,t = 1,\ldots,n$. Due to analytical and computational complexities, Whittle (1953) suggested to rather maximise the following approximation to the Gaussian log-likelihood in (2.13):

$$\widetilde{\ell}_n(\theta) = -\frac{n}{2}\left[\log 2\pi + \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\{2\pi f(\omega;\theta)\}\,\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_n(\omega)}{f(\omega;\theta)}\,\mathrm{d}\omega\right], \tag{2.14}$$

where for once writing i for the imaginary unit,

$$I_n(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} Y_t\exp(-i\omega t)\right|^2 = \frac{1}{2\pi n}\left[\left\{\sum_{t=1}^{n} Y_t\cos(\omega t)\right\}^2 + \left\{\sum_{t=1}^{n} Y_t\sin(\omega t)\right\}^2\right], \tag{2.15}$$

is the periodogram. For large samples sizes n, it is computationally less intensive to evaluate this likelihood, which is $O(n\log n)$, compared to the full likelihood, which is $O(n^3)$. It is also easier to handle analytically. Under suitable regularity conditions, the maximum likelihood, the Whittle approximated maximum likelihood, the estimators based on the Yule-Walker equations, and certain conditional and unconditional least squares estimators, are all asymptotically equivalent (Shumway & Stoffer, 2011, Property P.3.10). These have behaviour inside and outside model conditions corresponding to those in (2.2) and (2.4), see e.g. Dahlhaus & Wefelmeyer (1996, Theorem 3.3). Other parametric forms may be handled similarly by maximum likelihood or the Whittle approximation, either via parametric covariance functions like the Matérn or 'rational quadratic' forms, or by describing the parametric forms of the spectral density directly.

The periodogram introduced above has the property that it estimates the spectral density nonparametrically; in fact, under appropriate conditions, one may show that $\mathrm{E}\{I_n(\omega)\} = g(\omega) + O(n^{-1})$ (Brillinger, 1975, Theorem 5.5.2.). Thus, a natural nonparametric estimator of the spectral distribution is

$$\widehat{G}(\omega) = \int_{-\pi}^{\omega} I_n(u)\,\mathrm{d}u.$$

Since $C(h) = \int_{-\pi}^{\pi}\cos(\omega h)\,\mathrm{d}G(\omega)$, the above estimators may be transformed to a nonparametric estimator of any quantity determined by the covariance function $C(h)$. This involves, in particular, the variance and any lag covariances or correlations. In fact, under certain conditions $\sqrt{n}\{\widehat{G}(\cdot) - G(\cdot)\}$ has a zero-mean Gaussian process limit, see e.g. Priestley (1981, Ch. 6.2.5). This implies, in particular, asymptotic normality of covariances and correlations, and smooth functions of these.

3 Model selection

When being presented with data, the statistician or data analyst usually comes up with more than a single suggestion for how the data might be modelled. The consequence of this is often that a set of models are being fitted, each with their own consequences and conclusions if being trusted upon for further inference. This situation appears in the classical regression setup where there are 2^q different models defined by including or excluding each of the q covariates; when data with an unimodal distribution might deviate from the Gaussian in terms of skewness (skewed Gaussian distributions) or heavier tails (Cauchy, Laplace or t-distribution); and when a counting process are to be modelled, and both the Poisson and more general renewal processes appear appropriate. Model selection is the task of selecting the 'best' model from such a fixed set of candidate models. Figure 3.1 illustrates two basic model selection problems. In the left panel, the question is whether to trust the exponential, the Weibull or the gamma model being fitted to the positive data shown by the histogram. In the right panel, the question is whether a linear, quadratic or cubic regression model should be used to describe the relationship between two random variables X and Y. In addition to these parametric models one may of course consider nonparametric candidates, like those surveyed in Section 2.

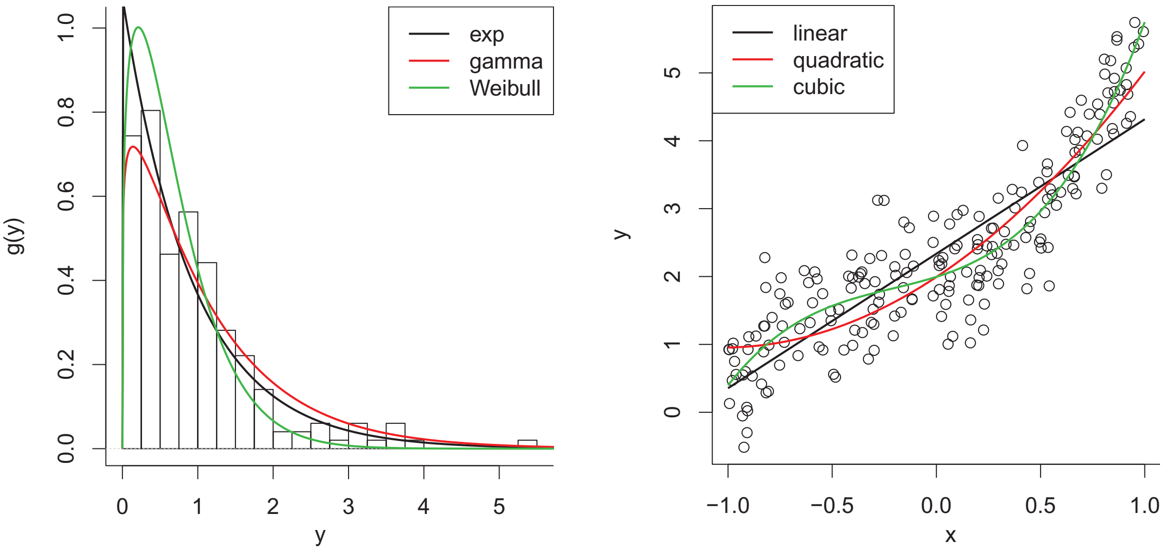

Figure 3.1: Two illustrations of typical model selection problems. The left panel shows i.i.d. data on the positive axis along with the density of three fitted parametric models. The right panel shows a scatterplot of data corresponding to X and Y, along with three fitted parametric regression models.

Various definitions of 'best' have led to a broad range of procedures for model selection, ranging from visual inspection and checks of p-values, to goodness of fit procedures and various information criteria. Below we first review some classical information criteria for parametric models and discuss model selection between parametric and nonparametric models, before we introduce and discuss more thoroughly the focused approach to model selection.

3.1 Classical information criteria for parametric models

When selecting among parametric models, some of the most popular methods are defined in terms of information criteria. The information criteria are characterised by formulae which assign a data dependent score to each candidate model. These scores are then used to rank the models in terms of their performance. The model with the best score is selected and trusted for further inference, while the others are discarded.

Several of the most famous and frequently applied information criteria take the following simple penalised log-likelihood form:

$$\text{IC} = 2\ell_n(\widehat{\theta}) - \text{penalty},$$

where the model reaching the highest scores is deemed the best. The likelihood term measures how well the model fits the data, while the penalty term penalises for model complexity – since more parameters typically means increased variance for estimators based on the fitted model. Different penalty terms corresponding to different information criteria, having different properties and possibly different motivations.

The simplest, most famous, and frequently applied information criterion is surely Akaike's information criterion (AIC) (Akaike, 1974). For p the dimension of θ, the AIC uses penalty $=$ $2p$, yielding

$$\text{AIC} = 2\ell_n(\widehat{\theta}) - 2p. \tag{3.1}$$

Although the AIC formula seems quite natural, it stems form quite involved theoretical reasoning. Consider for simplicity the i.i.d. situation. Modulo asymptotically negligible terms, the AIC score is then proportional to a bias adjusted estimator of $\int g(y) \log f(y;\theta)\,\mathrm{d}y$, the decisive ingredient in the Kullback–Leibler divergence $\mathrm{KL}(g, f_\theta)$ given in (2.3). The penalty is in this sense a large sample motivated bias adjustment, rather than a direct measure of model complexity.

The way that the AIC adjusts for this bias has certain limitations, however, and there exists numerous papers attempting to correct for various drawbacks of this bias estimator. For instance, Sugiura (1978); Hurvich & Tsai (1989) use penalty $= 2pn/(n-p-1)$ to correct improper small sample properties of the AIC, while the TIC of Takeuchi (1976) uses penalty $= 2\widehat{p}^*$, where $\widehat{p}^*$ is an estimate of the generalised dimension of the parameter space $p^* = \text{trace}(J^{-1}K)$. The TIC may be viewed as a model robust version of AIC. In the derivation of AIC's bias adjustment it is implicitly assumed that the fitted model is correct – which is of course unrealistic in a model selection setting. This AIC-assumption has the consequence that $J = K$ which gives $p^* = p$, and thereby simplistic AIC-formula in (3.1). The TIC takes this part more seriously, and estimate both J and K to get TIC $= 2\ell_n(\widehat{\theta}) - 2\text{trace}(\widehat{J}^{-1}\widehat{K})$. The generalised information criterion (GIC) of Konishi & Kitagawa (1996) generalises the idea behind TIC to allow for more general parameter estimation procedures, like that of M-estimation.

Besides the AIC and its cousins, the Bayesian information criterion (BIC) (Schwarz, 1978), sometimes also referred to as Schwarz's information criterion (SIC), is in frequent use. The BIC uses penalty $= p \log n$, that is

$$\text{BIC} = 2\ell_n(\widehat{\theta}) - p \log n.$$

It thus penalises model complexity harder than the above criteria for all but tiny sample sizes. Although the BIC is structurally very similar to AIC and its cousins, its underlying motivation is very different. As the name reveals, the BIC has Bayesian roots. Write now BIC_M for the BIC score of candidate model M. From a Bayesian perspective, $\exp(\text{BIC}_M/2)$ approximates a quantity which is proportional to the posterior probability that model M is correct, assuming all candidate models are a priori equally likely. Despite the Bayesian motivation, the criterion is mainly used, in its original form, for frequentist model selection problems. For various other information criterion, like the DIC, minimum description length, Mallows CP, and so on, see e.g. Claeskens & Hjort (2008a, Ch. 2-3).

3.2 Parametric or nonparametric?

The nonparametric (and semiparametric) modelling approaches are powerful and robust tools for inference, largely due to their consistency and Glivenko–Cantelli type properties. However, when the truth is 'close' to some parametric model with reasonably few parameters, it is typically preferable to proceed with such a parametric model. The reason for this is that parametric modelling approaches typically impose a smaller variance, making these more efficient. In the response to the discussants of his pioneering Cox regression paper (Cox, 1972), Sir David Cox himself emphasises the importance of considering parametric options: "[The semiparametric Cox model] is only one way of proceeding and the possibility of a parametric representation of [$\alpha_0(t)$] will often be worth consideration." The importance of this somewhat undercommunicated principle is also stressed by Bradley Efron in Efron (1988).

It is evident that model selection involving a set of appropriate fully parametric models and a lightly structured nonparametric (or semiparametric) modelling approach, constitutes an important model selection problem. Despite its importance, there is surprisingly little literature on comparison and selection among parametric and nonparametric modelling approaches. This is possibly caused by the difficulty of naturally raising or handling such a problem. The range of information criteria reviewed in Section 3.1 cannot be directly extended to include such models, since they all rely on likelihoods – a feature the nonparametrics do not possess. To my knowledge, there exists no specifically constructed criterion or procedure for selection among parametrics and nonparametrics. Below we shall discuss the problems associated with the most intuitive approaches to such selection.

3.2.1 Goodness-of-fit tests

Goodness-of-fit tests are characterised by their ultimate goal of testing whether a set of observations stems from a specified statistical model or family. The goodness-of-fit is typically measured in terms of a statistic which is small for good fits. The null hypothesis that the data stem from the described model is then rejected if the observed value of the test statistic is too large compared to what is to be expected under the null hypothesis.

Pearson's chi squared test (Pearson, 1900) for categorical or categorised data, is possibly the oldest and most famous goodness-of-fit test. Assume there are k different categories defined in only *one* direction (i.e. a single variable defines the categorical affiliation of an observation). Consider first the case of checking a fully specified null hypothesis, i.e. whether the probabilities associated with each of these k categories $(p_1, \ldots, p_k)$ fulfil $p_j = p_{j,0}, j = 1, \ldots, k$ for fully specified probabilities $p_{1,0}, \ldots, p_{k,0}$. The test statistic then takes the form

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},$$

where O_j is the number of observations in category j, while E_j is the expected number of observations under the null hypothesis. That is $E_j = np_{j,0}$. The test statistic may consequently be expressed in terms of frequencies: $X^2 = n\sum_{j=1}^{k}(O_j/n - p_{j,0})^2/p_{j,0}$. Under the null hypothesis, the limiting distribution of X^2 (as $n \to \infty$) is χ^2 with $k-1$ degrees of freedom, which is then used to test the null hypothesis. In applications, it is usually more interesting to test null hypotheses constituting a family of probabilities: $p_j = p_{j,0}(\theta), j = 1, \ldots, k$ with flexibility imposed by a q-dimensional parameter θ. With this formulation, each value of θ constitutes a separate value of the X^2 statistic above. Indeed it follows that the smallest value of this tests statistic, i.e. $\min_\theta X^2(\theta)$ (with $X^2(\theta) = n\sum_{j=1}^{k}(O_j/n - p_{j,0}(\theta))^2/p_{j,0}(\theta)$) follows a χ^2 distribution with $k-1-q$ degrees of freedom. In some situations the θ has been fitted initially via e.g. maximum likelihood, perhaps before categorising the data, and one actually wants to test whether this specific fitted parametric model suits the data well. The distribution of the test statistic is no longer χ^2 distributed, but has a precise asymptotic distribution which is stochastically somewhat larger than the χ^2 distribution (Chernoff & Lehmann, 1954). In the case where the categorisation depends on θ, the complexity increases even further. Such cases are discussed in e.g. Moore & Spruill (1975).

Instead of working with categorical or categorised data, the Cramér–von-Mises (Cramér, 1928; von Mises, 1928) and Kolmogorov–Smirnov tests (Kolmogorov, 1933; Smirnov, 1948) work directly with the empirical distribution function. Let us return to the case where one ought to test a fully specified null hypothesis. Consider testing of the null hypothesis $G = F_0$ for some fully specified distribution function F_0. The two test statistics are then given by respectively

$$\text{CvM} = \int (\widehat{G}(y) - F_0(y))^2 \, \mathrm{d}F_0(y) \quad \text{and} \quad \text{KS} = \sup_y |\widehat{G}(y) - F_0(y)|.$$

Under the null hypothesis, the quantities nCvM and $\sqrt{n}$KS have precise asymptotic distributions being transformations of Gaussian processes independent of F_0; see e.g. Anderson & Darling (1952). These tests reject the null hypothesis if the observed test statistics deviate significantly from the centrality of their asymptotic null distributions. In model comparison settings it is usually more appropriate to compare a family of distributions, say $F(\cdot;\theta)$, than a specific data independent configuration of such a family, like above. Estimating the θ parameter of such a family, and applying the above procedures as if $F(\cdot;\widehat{\theta})$ was given a priori, is however not generally valid. 'Luckily', Durbin (1973a,b) provide asymptotic distributions for the two test statistics which hold when the parameters are estimated based on the same data. This may be utilised to test these more general null hypothesis. In addition, tables of critical values have been developed to correct for small sample inaccuracies in these distributions.

Although the goodness-of-fit tests above are not constructed for the purpose of performing model selection, they may in principle be used for selection among parametric candidate models and the standard nonparametric competitor which puts equal weight on each observation. The strategy pans out as follows: Rank the parametric models according to some goodness-of-fit measure, accounting also for model complexity. If all measures exceed the threshold value set according to a certain (asymptotic) significance level then the corresponding natural nonparametric model is chosen; otherwise the parametric model associated with the smallest goodness-of-fit measure is selected. Although such a procedure appears natural and consistent in theory, and there is a rich literature on multiple testing, the outlined model selection procedure does not seem to have been applied or written out explicitly in the literature before. This is perhaps due to the difficulty of setting threshold values with several, possibly partially nested, models fitted from data – which should correct also for model complexity. Another possible drawback with this approach is that the performance or uncertainty of the nonparametric model is not explicitly measured or accounted for.

3.2.2 The nonparametric likelihood

Above we claimed that the nonparametric model has no likelihood. This is in some sense not entirely correct for e.g. the nonparametric kernel density estimator $\widehat{g}$ of the form (2.6). The 'likelihood' of this model is simply $\prod_{i=1}^{n} \widehat{g}(Y_i)$. This should in principle allow comparison with other parametric models by consulting a likelihood based information criterion. Since the bandwidth is typically set using the data, the number of parameters would be 1. The problem with this approach is however that the bandwidth is not selected based on the likelihood; instead, the smaller the bandwidth – the larger the likelihood. Thus, the whole concept collapses as the bandwidth $h \to 0$ and the 'likelihood' reaches infinity. For other thoughts in this direction, particularly related to the concept of 'empirical likelihood', see e.g. Owen (2001).

Another way to view the nonparametric model is, as mentioned in Section 2, as a parametric model with an infinite dimensional θ parameter. Provided with a finite sample from some distribution, it is of course not possible to find such a parameter. For certain classes of parametric models, one may however fit a parametric model $F(\cdot;\theta)$ with a very large number of parameters. For i.i.d. data one may for instance 'log-expand' any density (or probability mass function) f_0 by a sequence of exponential families like in Barron & Sheu (1991):

$$f(y;\theta) \propto f_0(y) \exp\left\{ \sum_{j=1}^{k} \theta_j \Psi_j(F_0(y)) \right\},$$

for a fairly large number of orthonormal basis functions $\Psi_1, \ldots, \Psi_k$. For time series data one may 'similarly' use e.g. an AR(p) model of very large order p. The problem with such procedures is however that the data are not used as efficiently as in standard nonparametric approaches. Parametric models with lots of parameter flexibility typically estimate features of the distribution implying very detailed structures and properties, i.e. they overfit. This may result in parametric models whose estimators have larger variance than that of the nonparamatric candidate. Although the parametric models with large number of parameters may appear smoother and visually better-looking, their statistical properties typically make them unfavourable compared to both the parametric models with few parameters and nonparametric models.

3.3 Focus parameter and the focused information criterion (FIC)

The ultimate goal of statistical modelling is often to estimate and perform inference for one or a few specific population quantities. Which quantities that are of interest typically depends on the setting, the data type, and indeed why the analysis is carried out in the first place. Examples ranges from a measure of centrality like the mean, median or mode, to a measure of spread like the standard deviation, the interquartile or interdecile range, a quantile, and the probability that a certain data dependent event occurs – all of these possibly conditioned on specific covariate values when such are available. We hereby refer to such a pre-specified population quantity as a focus parameter, and denote it by μ.

When the purpose of the analysis is mainly to perform inference for one or a few such focus parameters, it seems natural to take this information into account also when selecting model. This breaks with the principle of the classical parametric model selection criteria mentioned in Section 3.1 (which rather seek the model with best *overall* fit and properties) and with the goodness of fit approaches mentioned in Section 3.2.1. The focused information criterion (FIC) introduced by Claeskens & Hjort (2003), however, is constructed exactly to accommodate the principle of utilising the purpose of the analysis. Rather than seeking the model with the best overall properties, the FIC seeks to find the model which best estimates a single pre-defined focus parameter μ. This is achieved by measuring the accuracy of the model based estimator $\widehat{\mu}$ of μ in terms of its mean squared error. The propriety of such approaches are emphasised also by Hand & Vinciotti (2003) and Longford (2005).

As a proof of concept for such a focused model selection strategy, recall the setup and notation in Section 2.2 and consider the following simulated survival analysis case: A set of $n = 50$ survival times are sampled from the Weibull distribution with cumulative hazard $A(t) = t^{1.3}$ (i.e. having scale = 1 and shape = 1.3). The observed survival times are right censored, and the censoring distribution is exponential with constant rate 1/4, resulting in about 20% censoring. We repeat this simulation procedure several times, and for each such simulated data set we fit the parametric Weibull model and the simpler exponential model, and compute the fitted survival functions $S(t) = \Pr\{T_i^{(0)} > t\}$, in addition to the nonparametric Kaplan–Meier estimator. The Monte Carlo mean squared errors of the model based estimators are then computed across a range of t-values. Figure 3.2 shows the root of these Monte Carlo mean squared errors (rmse) in addition to the mean of the estimated survival probabilities. As seen from the figures, the risk is uniformly smaller for the estimator based on the fitted Weibull model compared to the Kaplan–Meier estimator, with a clear efficiency gain. This is perhaps not very surprising, as the Weibull model is indeed the true model here. It is then more constructive to compare the simpler,

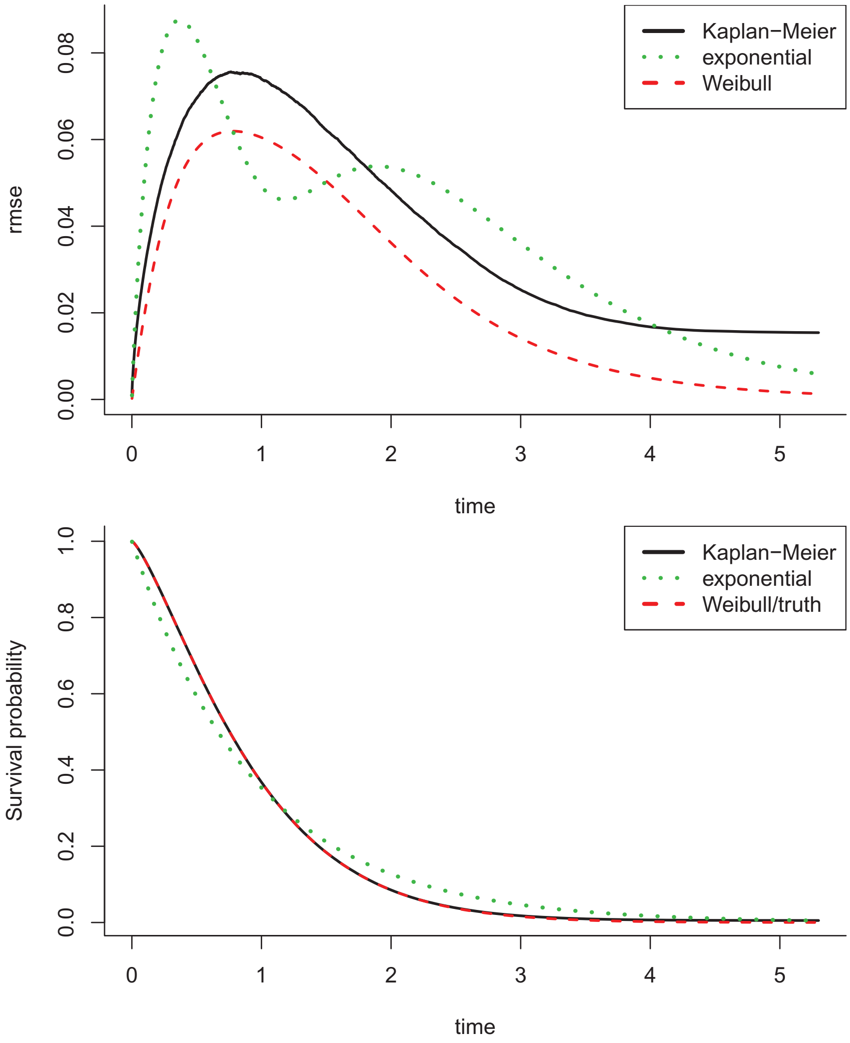

Figure 3.2: Root mean squared error based on simulated Weibull survival times, and average estimates of survival times.

biased estimator based on the exponential model, say $\exp(-\widehat{\theta}t)$, to the Kaplan–Meier estimator. The former has larger risk for small and intermediate valued t, but smaller risk for the largest t and also within an interval around $t = 1$. This pinpoints that sometimes, but certainly not always, it is wise to rely on even a biased parametric estimator, rather than the asymptotically unbiased nonparametric estimator. Note that in an interval around 1, the estimator based on the exponential model also outperforms the Weibull model. This tells us that a simpler, misspecified parametric model *may* be useful even when a bigger parametric model is indeed correct. Hence, which model or estimator that should be used depends on which questions are deemed more important, i.e. on the (focus) parameter selected for scrutiny.

3.3.1 More on the focused information criterion

Below we dig further into the details underlying the original focused information criterion of Claeskens & Hjort (2003), where we for presentational simplicity concentrate on the i.i.d. case. The authors work with nested parametric models which all lie between a narrow model described by a p-dimensional parameter vector θ and the wide model which uses an additional

q-dimensional parameter vector γ. The family of densities or probability mass functions of the narrow and wide model is thereby described by respectively $f(y;\theta) = f(y;\theta,\gamma_{\text{narr}})$ and $f(y;\theta,\gamma)$, where γ_{narr} is the value of the γ parameter making the wide model reduce to the narrow model. In deriving large sample results, the authors work under a so-called local misspecification framework (Hjort & Claeskens, 2003a). This framework assumes that the true data generating distribution has a density or probability mass function

$$g_n(y) = f(y;\theta_{\text{true}},\gamma_{\text{narr}} + \delta/\sqrt{n}). \tag{3.2}$$

Here, the unknown parameter value θ_{true} determines the limiting true distribution $f(y;\theta_{\text{true}}) = f(y;\theta_{\text{true}},\gamma_{\text{narr}})$. The unknown q-dimensional δ parameter describes the $O(1/\sqrt{n})$ distance between the true and the narrow model – in the direction of the γ parameter. Thus, the true model shrinks with increasing sample size, and reaches the narrow model in the limit. This gives squared biases of the same asymptotic 'size' as the variances, namely $O(n^{-1})$.

Each candidate model M, between (and including) the narrow and wide models, constitutes an estimator $\widehat{\mu}_M$ for the focus parameter μ, having true value $\mu_{\text{true},n}$ under the local misspecification framework in (3.2). The authors show that under weak regularity conditions related to the focus parameter and the candidate models themselves, $\sqrt{n}(\widehat{\mu}_M - \mu_{\text{true},n}) \rightarrow_d \Lambda_M \sim \mathrm{N}(\text{bias}_M, \text{var}_M)$, with precise expressions for bias_M and var_M of each candidate model M. The bias_M and var_M depend on the focus parameter μ, the local deviation parameter δ and how the models are nested. As a consequence, the n-scaled squared loss function has a well-defined limit $L_{n,M} = \{\sqrt{n}(\widehat{\mu}_M - \mu_{\text{true},n})\}^2 \rightarrow_d L_M = \Lambda_M^2$. The expectation of this limiting loss L_M is thus

$$\text{risk}_M^* = \mathrm{E}\{L_M\} = \text{bias}_M^2 + \text{var}_M.$$

This risk_M^* is a natural approximation to n times the actual (finite sample) mean squared error $\text{mse}_M = \mathrm{E}[(\widehat{\mu}_M - \mu_{\text{true},n})^2]$. In view of this mean squared error approximation, the FIC scores are estimates of the risk_M^*, and may hence be expressed as

$$\text{FIC}_M = \widehat{(\text{bias}_M^2)} + \widehat{\text{var}}_M.$$

These are computed for each candidate model M using the observed data, and the model with the smallest FIC-score is selected and trusted as the best model for estimating the focus parameter μ. Thus, instead of aiming at a model with good overall fit and properties, the motivation of the FIC is that the intended use of the model and focus of the investigation should be the central part of the selection procedure. This has the consequence that even for the exact same data, different models typically are deemed the best for estimating different focus parameters.

The FIC apparatus is specially geared towards model selection where estimation of a single prespecified focus parameter μ is the main goal. The framework suggests that in situations where several focus parameters are of interest, the FIC procedure should be repeated for each such μ, possibly resulting in different models being selected for different focus parameters. In some cases, however, there may be good reasons to endorse only a single final model for estimating μ. An extension of the FIC, termed the weighted or averaged focused information criterion (AFIC) and developed in Claeskens & Hjort (2008b), deals with this type of problem. The criterion aims at selecting the model which best estimates the whole set of focus parameters $\mu(u)$ for u in some index set. Thus, consider the loss function

$$L'_{n,M;W} = \int [\sqrt{n}\{\widehat{\mu}_M(u) - \mu_{\mathrm{true},n}(u)\}]^2 \,\mathrm{d}W(u),$$

for W some cumulative weight function chosen from the context to reflect the relative importance of the different $\mu(u)$. In the local misspecification framework we then typically have $L'_{n,M;W} \to_d L'_{M;W} = \int \Lambda_M(u)^2 \,\mathrm{d}W(u)$. The expectation of the limiting loss function is thus

$$\mathrm{risk}^*_{M;W} = \mathrm{E}\{L'_{M;W}\} = \int \mathrm{E}\{\Lambda_M(u)^2\} \,\mathrm{d}W(u) = \int \{\mathrm{bias}_M(u)^2 + \mathrm{var}_M(u)\} \,\mathrm{d}W(u).$$

Similarly to the reasoning underlying the FIC, the AFIC strategy is to estimate $\mathrm{risk}^*_{M;W}$ for all candidate models, and select the model with the smallest estimated risk. Such a strategy may in particular be fruitful when a single precise focus parameter cannot be defined. The data analyst might for instance be asked to find a model which produces good estimates of the ‘upper tail’ of the distribution, without further specifics of what is meant by the ‘upper tail’. AFIC may then be applied with a focus parameter set including all quantiles from say 0.8 to 0.999, along with a suitable weight function perhaps peaking at the 0.9- or 0.95-quantile. Although not 100% focused like the FIC, such a strategy ought to return a model better tuned towards the intended use of the model than those returned by overall model selection procedures like the AIC.

In addition to the i.i.d. setting, Claeskens & Hjort (2003) presents FIC for the more general regression setting, using a generalisation of the misspecification framework in (3.2). With similar local misspecification frameworks, FIC-procedures have also been developed for generalised additive partial linear models (Zhang & Liang, 2011), Cox’s proportional hazards semiparametric regression model (Hjort & Claeskens, 2006), certain versions of the Aalen model (Gandy & Hjort, 2013), autoregressive and autoregressive moving average time series models (Claeskens et al., 2007; Rohan & Ramanathan, 2011), for missing data (Sun et al., 2014), certain classes of semiparametric models (Claeskens & Carroll, 2007), quantile regression (Behl et al., 2014), in addition to a Bayesian version (Nguefack-Tsague & Bulla, 2014). Such have also been utilised in certain applications within economics (Behl et al., 2012), finance (Brownlees & Gallo, 2008), fisheries science (Hermansen et al., 2016), personalised medicine (Rolling & Yang, 2014) and for population size estimation (Bartolucci & Lupparelli, 2008). Yet others have worked with other loss functions than the squared loss. Claeskens et al. (2006) work with the more general L_p loss (having squared and absolute loss as special cases), while Brownlees & Gallo (2011); Zhang et al. (2012) work also with the nonsymmetric linear exponential (LinEx) loss (Varian, 1975; Zellner, 1986).

4 Geophysical/rock physical inversion and approximate Bayes

In contrast to the convention in the previous sections, we shall in this section use solely lowercase roman letters for (vector valued) variables. As commonly employed in Bayesian contexts, we also use $p(\cdot)$ as a generic notation for probability distributions (both densities and probability mass functions).

4.1 Geophysical data and rock physics

Geophysics is the study of the physical processes and properties of the earth. The field comprises several different types of applications enabling us to learn about plate tectonics, volcanoes, earthquakes and so on. In the petroleum industry, geophysical knowledge is essential for locating hydrocarbons in reservoirs thousands of meters below the surface. Extraction of hydrocarbons, which are the main ingredient in fossil fuels, is essential for the modern society to function.

The ultimate goal for petro- and geophysicists is often to predict/estimate unobservable properties of the rock in the subsurface, such as lithology (i.e. the rock type), porosity, permeability or saturation. We shall denote such latent quantities by r. The field of rock physics or petrophysics studies the relationship between these rock properties r, and what we denote by m and refer to as the geophysical properties or material characteristics. These geophysical properties typically consist of the density of the rock, in addition to elastic parameters which describe the speed at which compressional and shear sound waves move through the rock. Such geophysical properties may neither be measured directly (i.e. they are also latent), but well established geophysical models describe their relationship to observable geophysical data d. Thus, the pathway from rock properties r to geophysical data d has the simple structure illustrated in Figure 4.1. For more on seismic data, geophysics and rock physics, see e.g. Avseth et al. (2010).

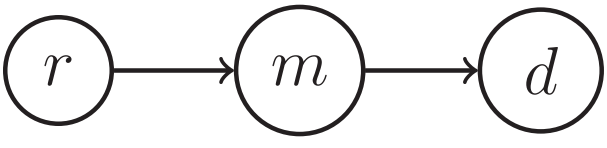

Figure 4.1: Geophysical and rock physical structural illustration

The geophysical data d typically stems from a seismic survey. Although such surveys are also carried out ashore, we will here focus on offshore marine seismic surveys, as illustrated in Figure 4.2. In a marine seismic survey, a seismic shock wave is ‘shot’ down in the water from an air gun towed behind a boat. As the wave reaches boundaries between two layers of rock

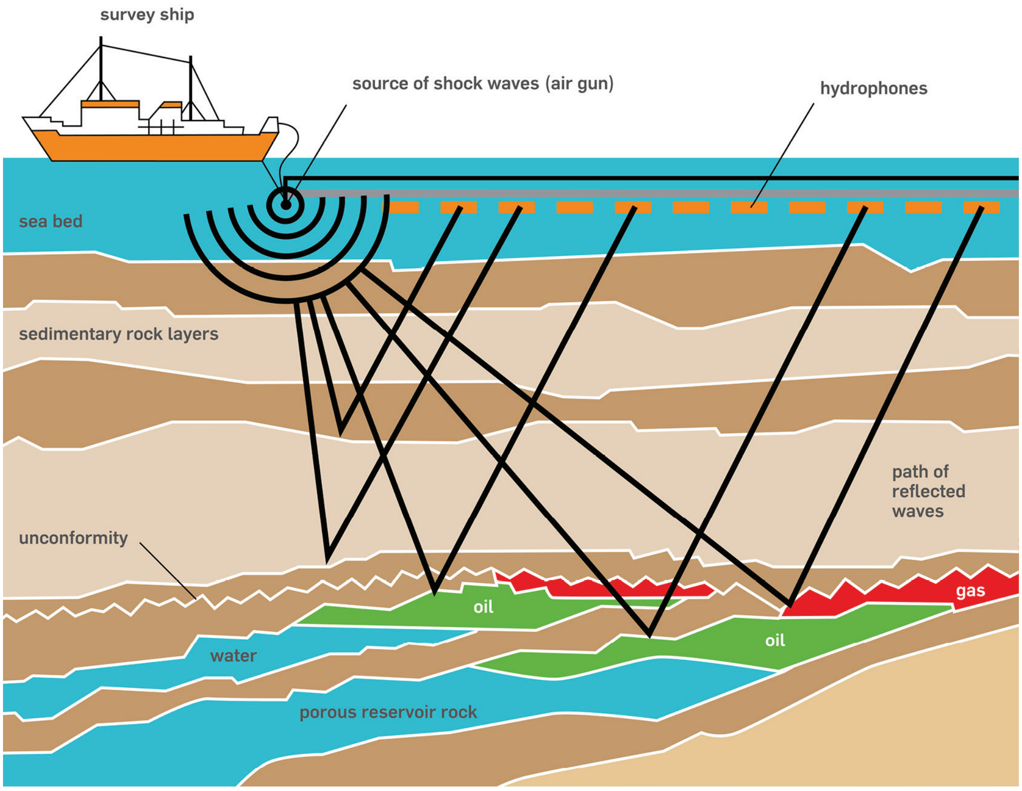

Figure 4.2: Sketch of marine seismic survey. Figure reprinted from krisenergy.com (2015).

with different properties, part of the wave energy is reflected while the rest continues through the layer to reach the next 'boundary'. As the reflected seismic waves approach the water surface, the amplitudes are recorded by hydrophones attached to long cables being towed behind the boat.

After the survey, a rather comprehensive processing procedure is usually carried out by geophysicists. The processing typically consists of deconvolving, stacking and migrating the data, see e.g. Buland & Omre (2003, Table 2.3) for an example. This process is required to clean, align, and adjust the data in order to make them conformable as input in established geological models. The most common form of processed data is seismic amplitude versus offset (AVO) data. Figure 4.3 shows a small illustrational 2D cross section of processed seismic AVO data from a larger 3D seismic survey offshore Norway.

4.2 The inverse problem and the Bayesian approach

In some sciences, including the geosciences, 'the inverse problem' is defined as the problem of finding an unobservable cause or source that has generated an observed set of data. In the geophysics/rock physics setting above, this corresponds to finding either the geophysical properties m, or ultimately the rock properties r, based on observed geophysical data d. It is referred to as the inverse problem simply because it is the opposite of the forward or direct problem. The forward/direct problem consists of finding the possible implications from a cause – i.e., according to the laws of physics and geological knowledge, what geophysical data would one expect

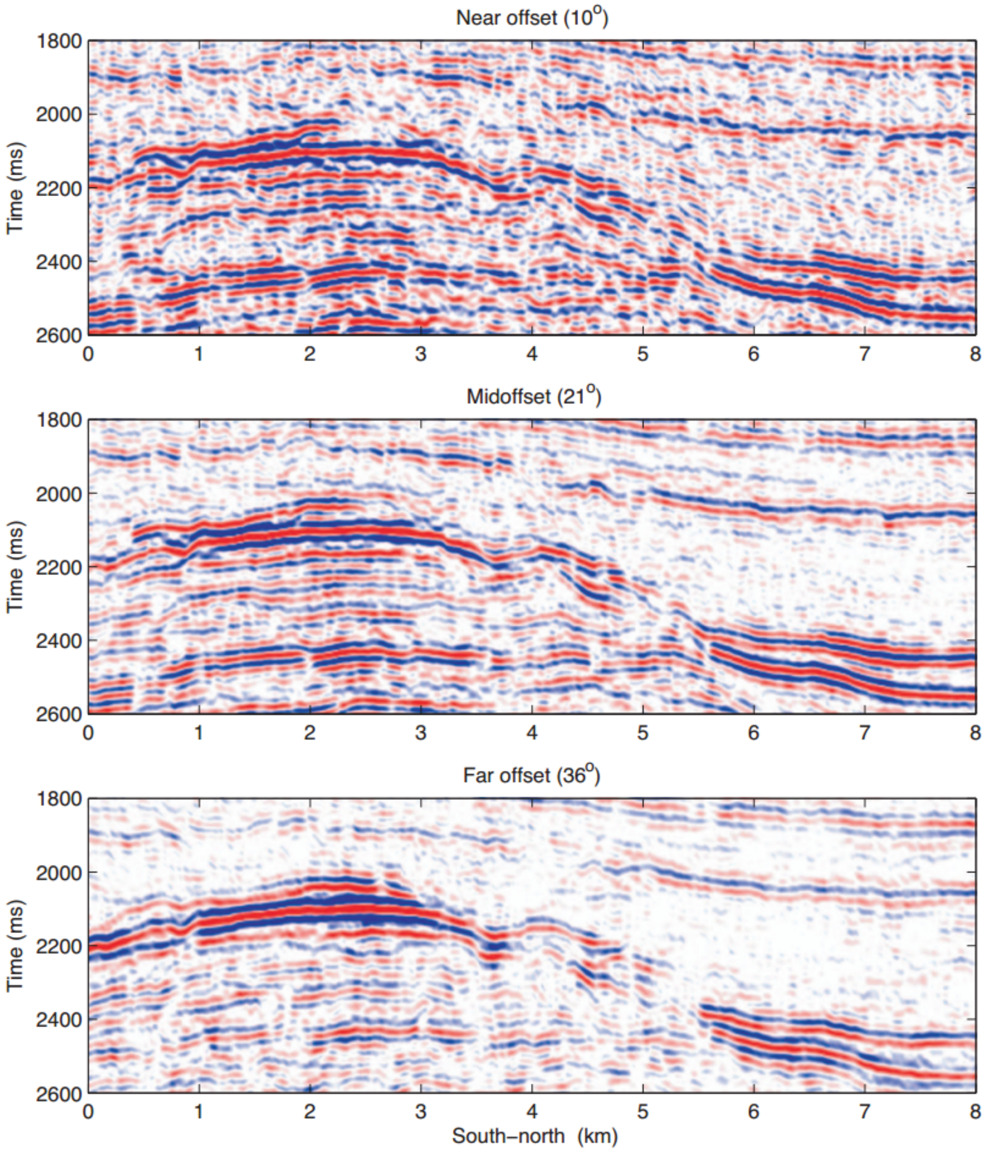

Figure 4.3: Example of processed 2D seismic AVO data for three different offsets extracted from a larger 3D seismic survey offshore Norway. Negative amplitudes are red, positive amplitudes are blue. Figure reprinted from Buland et al. (2008).

to observe if the subsurface had a certain composition of rock or geophysical properties. From a completely general mathematical perspective, these problems relate to the formulation

$$y = H(x), \tag{4.1}$$

where y denotes the observations, x is the unobservable (latent) cause or source, and H is some operator that describes the (causal) mechanism that produces y given x, possibly based on laws of physics etc. When uncertainty is involved, either related to the observation or recording of y, or inaccuracies in the operator H, one may write $H(x) = H_0(x) + \varepsilon$, such that (4.1) can be rewritten as

$$y = H_0(x) + \varepsilon. \tag{4.2}$$

In this representation, H_0 is a fixed, non-stochastic operator and ε is some stochastic error term. In the setting of Section 4.1, y represents the geophysical data d, while x is either the geophysical properties m, or ultimately the rock properties r.

The simplest case of (4.1) and (4.2) imaginable occurs when H_0 is a linear operator, i.e. a matrix. Even if this is seldom the case, that theory is by far the most developed. Thus, it is not uncommon to 'linearise' a nonlinear inverse problem and then employ methodology to the linearised problem. Although such an approach may work in some cases (which are close to linear), the approach is not generally valid, and may lead to severely wrong conclusions. Before we turn to approaches for solving inverse problems, note that inverse problems are typically 'ill-posed'. In the sense of Hadamard (1923) this means that the problem does not fulfil all criteria for a well-posed problem: existence, uniqueness and stability of a solution. For an inverse problem to be well-posed, the operator H_0 needs to be bijective, i.e. invertible. In the linear case this corresponds to H_0 being a square matrix with full rank, which is rarely the case.

Mathematicians often study the existence and uniqueness of solutions in noise-free frameworks with an infinite amount of data. In the presence of a noise term, this term is typically either neglected, or treated as deterministic in the studies. In the simplest linear problem, where the H_0 matrix has full rank, the unique solution is found by solving the resulting linear system. In the less restricted case where H_0 has more rows than columns and $H_0^{\mathrm{t}}H_0$ has full rank, the ordinary least squares solution $(H_0^{\mathrm{t}}H_0)^{-1}H_0^{\mathrm{t}}y$ is optimal in terms of the L_2-norm. 'Mathematical' optimal solutions to more involved problems are typically related to certain types of regularisation and generalised matrix inversion. For a more complete review of inverse problems casted in a mathematical theoretic setting, see e.g. Krisch (2011).

From a statistician's viewpoint, the treatment of the error term in the 'mathematical' approach is not very accommodating. It then seems more appropriate and realistic to treat the errors as random noise, which they in many cases indeed are. In a statistical framework, the inverse problem corresponds simply to carrying out statistical inference for the latent variable x, or certain aspects thereof, given the observed data y and the probability distribution for the noise term ε.

In fact, the linear inverse problem can be re-formulated as a linear regression problem. When $H_0^{\mathrm{t}}H_0$ has full rank, and the covariance matrix of ε is known, the standard frequentist solution is to use generalised least squares (Aitken, 1935), or more generally maximum likelihood for non-Gaussian error terms. When $H_0^{\mathrm{t}}H_0$ does not have full rank, a possible frequentist solution is to use a shrinkage method such as lasso (Tibshirani, 1996) or ridge regression (Hoerl & Kennard, 1970). However, for more general inverse problems, the most common frequentist approaches appear to be based on minimax estimation theory (Stark, 2000). The main benefit of treating the error term as random noise in the statistical approach, is the possibility to quantify the uncertainty of the solution. In a frequentist framework this is typically done in terms of one or more confidence intervals (or regions) with a confidence level α typically equal to 0.9 or 0.95. The interpretation of this interval (or region) is that if the 'experiment' was repeated several times, then a proportion α of these experiments would have intervals (or regions) containing the true solution to the inverse problem. From a frequentist viewpoint, the confidence level is however not a true probability. Also, typically the coverage proportion α holds only asymptotically as the effective sample size tends to infinity.

In a Bayesian statistical framework, linear and nonlinear inverse problems are handled analogously, modulo computational complexity. As in the frequentist approach one specifies H_0 and the probability distribution for the noise term ε, which describes the likelihood $p(y|x)$. However, in the Bayesian approach, the latent variable x is *also* treated as random. Thus, one needs to specify its prior probability distribution $p(x)$, reflecting the knowledge and belief of the statistician regarding this latent variable, before looking at the data. The Bayesian solution to the inverse problem is then simply to consult the posterior distribution of the latent variable $p(x|y)$. By Bayes' theorem (Bayes & Price, 1763), this distribution is given by

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \propto p(y|x)p(x). \tag{4.3}$$

The Bayes estimator is the point estimator, in which the Bayesians trust the most. This estimator depends not only on the resulting posterior distribution, but also on a loss function associated with the decision to be made. The most common Bayes estimators are the posterior mean, median or mode based on respectively the L_2-, L_1- and 0-1 loss functions. In terms of the point estimator, the Bayesian approach is no different from the frequentist – the distinction between the approaches lies in how the uncertainty is quantified. In the Bayesian framework, the uncertainty or belief in the solution is quantified by one or more credibility intervals (or regions). As opposed to the frequentist confidence intervals (or regions), the Bayesian credibility intervals (or regions) are to be interpreted as true probability distributions. Thus, it makes sense to say that the true value of the latent variable x lies in a certain interval or region with some probability derived directly from the posterior distribution. This interpretation is possibly the key tenet of the Bayesian approach. In addition, the ease of which prior information and additional knowledge about the problem may be incorporated directly in the solution, has given the Bayesian approach a large faithful fanbase.

4.3 General approximate Bayesian inference

In the Bayesian statistical approach one typically pose a problem like the above in a hierarchical setting by introducing also a set of parameters θ, typically corresponding to unknown model components of the distribution of either x and y. The full model description then requires also a prior distribution for these parameters $p(\theta)$, in addition to a prior for the latent field $p(x|\theta)$ and the likelihood $p(y|x,\theta)$. When the mechanisms involving θ are fairly well-known and stable, and the actual interest is solely in the latent variable (or cause) x, one may treat these mechanisms as part of the model formulation. This makes it possible to exclude the θ parameter from the setup and thereby reduce the complexity of the model formulation to that of (4.3). Although (4.3) is a very simple formula, deriving an analytical, explicit formula for $p(x|y)$ may be terribly difficult, especially in higher dimensions. The reason for this is that one typically needs to solve the integral in the denominator of the formula. When the prior $p(x)$ is so-called conjugate for the likelihood $p(y|x)$, however, an explicit formula for the posterior is available. This class of priors, first introduced by Schlaifer & Raiffa (1961), has the property that the resulting posterior distribution takes the same parametric form as the prior. When the inference parameter (x) is low dimensional, most of the classical likelihood models have a decent range of conjugate priors to choose from. The range is more limited in higher dimensions, and there is no guarantee that the conjugate priors available are able to represent the statistician's previous knowledge.

When the prior takes a non-conjugate form, one typically needs to employ some kind of numerical or analytical approximation. When the dimension of the latent variable is low, this typically amounts to solving the integral in the denominator of (4.3) by a numerical integration procedure like the quadrature, trapezoidal or Simpson's rule; see e.g. Ausín (2014) for a recent review. Alternatively, sampling based approaches like direct Monte Carlo or more sophisticated conditional/weighted Monte Carlo, importance sampling and acceptance rejection procedures, may be applied; see e.g. Thisted (1988, Ch. 5); Rubinstein & Kroese (2008, Ch. 2 and 5). However, in order to maintain the accuracy of the solution when the (effective) dimension of the integrand increases, the number of required evaluation points typically grows exponentially. This phenomenon is known as the curse of dimensionality (Thisted, 1988, Ch. 5.7.2). Thus, in the high dimensional settings where the inverse problems typically live (at least those of the type described in Section 4.1), one needs to consult other procedures for approximating the posterior $p(x|y)$. Below we briefly review some of the most well-known classes of such procedures.

4.3.1 Markov chain Monte Carlo (MCMC)

The simulation based Markov chain Monte Carlo (MCMC) procedures (Robert & Casella, 2005) are without doubt the most popular technique for approximating non-trivial posterior distributions. The procedure is based on constructing a Markov chain that has the posterior distribution as its stationary or equilibrium distribution. After a number of initial burn-in samples, the (dependent) samples of a random walk with transitions corresponding to such a Markov chain will be approximately distributed according to the posterior distribution.

Although there exists a wide variety of strategies for constructing Markov chains with the desired properties, most procedures (possibly except for the so-called slice sampler) are variations of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) or the Gibbs sampler (Geman & Geman, 1984). The Metropolis-Hastings algorithm requires specification of a jumping or transition distribution, say $q(x_0|x)$, from x to x_0. Then, after initiating the algorithm by a starting value $x^{(0)}$, the algorithm repeats the following two steps for $t = 1, \ldots$: A proposal $x_0^{(t)}$ is sampled according to $q(x_0^{(t)}|x^{(t-1)})$, where $x^{(t-1)}$ denotes the previous sample. The proposal is then accepted as a new sample $x^{(t)} = x_0^{(t)}$ with probability $\min(1, \alpha)$ where

$$\alpha = \frac{p(x_0^{(t)}|y)q(x^{(t-1)}|x_0^{(t)})}{p(x^{(t-1)}|y)q(x_0^{(t)}|x^{(t-1)})}.$$

Otherwise, $x^{(t)}$ is set equal to $x^{(t-1)}$.

The alternative Gibbs sampler divides the sampling vector x into k subvariables, say $x_1, \ldots, x_k$, and requires that sampling from the distribution $p(x_j|y, x_{-j})$ is accessible for $j = 1, \ldots, k$, where x_{-j} denotes the vector containing all elements of x except x_j. Then, after initiating a starting value for x, each iteration of the algorithm consists of sampling posterior values of each of the $x_j, j = 1, \ldots, k$, conditional on the current value of the other variables, i.e. sampling $x_j^{(t)}$-variables according to $p(x_j|y, x_{-j}^{(t-1)})$ where $x_{-j}^{(t-1)} = (x_1^{(t)}, \ldots, x_{j-1}^{(t)}, x_{j+1}^{(t-1)}, \ldots, x_k^{(t-1)})$.

The popularity of the MCMC approach is mainly due to its incredibly general form. No matter how analytically messy your likelihood and prior looks, MCMC can provide samples from an arbitrarily precise approximation to the posterior distribution after a sufficiently long 'burn-in' period. One drawback of the approach is that some setups require extremely long 'burn-in' sequences, and it might be difficult to determine if the algorithm has 'converged'. Another is that the samples can be highly dependent, requiring a very large number of samples. This is a consequence of the difficulty of finding the right balance between mixing of the proposals and a decent acceptance rate. Thus, for very high dimensional problems, it may be difficult to obtain samples whose empirical distribution approximates the true posterior distribution with sufficient precision within a decent amount of time. For a more thorough discussion of strengths, weaknesses, and variations of the procedures, see e.g. Robert & Casella (2005).

4.3.2 Integrated nested Laplace approximation (INLA)

The recent integrated nested Laplace approximation (INLA) methodology by Rue et al. (2009) approximates posterior distributions for so-called latent Gaussian (Markov random field) models. Let us for this subsection once again assume there is an additional unknown θ parameter involved in the model formulation. A latent Gaussian (Markov random field) model is characterised as follows: $p(\theta)$ is low dimensional; $p(x|\theta) \sim \mathrm{N}(0, Q(\theta)^{-1})$, where $Q(\theta)$ is a sparse precision matrix, i.e. the inverse of the covariance matrix; and $p(y|x, \theta) = \prod_{i=1}^n p(y_i|\eta_i, \theta)$ where the $\eta_i = \sum_{j=1}^n c_{ij} x_j$ are known linear combinations of the latent field. Thus, the data are conditionally independent, the latent field is Gaussian, and the Markov random field denotation imposes the precision matrix $Q(\theta)$ to be sparse.

The basic principle of the INLA methodology is, as the name reveals, to utilise several nested Laplace approximations. The posterior distribution of θ is approximated using the Laplace approximation

$$p(\theta|y) \propto \frac{p(y, x, \theta)}{p(x|y, \theta)}\Big|_{x=x'(\theta)} \approx p^*(\theta|y) = \frac{p(y, x, \theta)}{p^*(x|y, \theta)}\Big|_{x=x'(\theta)},$$

where the mode $x'(\theta) = \mathrm{argmax}_x p(x|y, \theta)$ is typically found via numerical Newton–Raphson type optimisation. The marginal posterior distribution for the elements of θ are then found by numerical integration over θ. To approximate the marginal posteriors of the latent field $p(x_j|y)$ for $j = 1, \ldots, n$ (i.e. without conditioning on θ), the procedure uses

$$p^*(x_j|y) \approx \int p^*(x_j|y, \theta) p^*(\theta|y) \, \mathrm{d}\theta,$$

where $p^*(x_j|y, \theta), j = 1, \ldots, n$ are established through separate Laplace approximations.

The procedure delivers very impressive accuracy when applied to the models of this particular class. Also, when these are high dimensional, the speed-up of the procedure compared to the MCMC approach is typically several orders of magnitudes. The success of the INLA procedure lies merely in the details, rather than the basic construction above. This involves clever utilisation of the sparseness of the precision matrix $Q(\theta)$ introduced by the Markov random field, the optimisation routines for finding the modes above, and the numerical integration. The full INLA procedure is implemented in the R-package `R-INLA`, see e.g. Martins et al. (2013); Blangiardo & Cameletti (2015) for further details.

4.3.3 Approximate Bayesian Computation (ABC)

The approximate Bayesian computation (ABC) procedure (Beaumont et al., 2002) is a class of sampling based procedures for approximating the posterior distribution specifically targeted to deal with situations where evaluation of the likelihood function $p(y|x)$ is the bottleneck. This is particularly the case when the dimension of y is extremely large (compared to x) or for some reason there is no explicit analytic closed form formula for $p(y|x)$. Such circumstances occur frequently, for instance in certain applications in genetics, ecology and epidemiology. The method produces independent samples from the posterior distribution without the need to evaluate the troublesome likelihood $p(y|x)$, which is an inevitable part of all other procedures.

The basic form of the procedure goes as follows: First, a large number of variables are sampled from the prior distribution $p(x)$. Then, given each of these samples, a new set of data y' is sampled according to the likelihood model. (Obtaining such a sample is often feasible without evaluating $p(y|x)$.) Then the new data y' is compared to the original data y, and the sample is accepted as a sample from the posterior if it is deemed sufficiently similar to the original data. Closeness is typically measured by the 'distance' between some informative (preferably sufficient) summary statistics S of the two data sets. Thus, for a distance measure ρ, the sample is accepted if $\rho(S(y), S(y')) \leq \varepsilon$, where ε typically needs to be strictly positive for the procedure to be computationally feasible. Although there exist various techniques for selecting the summary statistic S, the distance measure ρ, and the cut-off value ε, this remains an active field of research. See e.g. Blum et al. (2013) for a review of such techniques.

Thus, when the inverse problem at hand has a computationally infeasible likelihood $p(y|x)$, the ABC procedure is in principle the only applicable procedure. When evaluation of the likelihood is not the bottleneck, other approaches are more suitable. For more on the theory and practical application underlying the ABC, see e.g. the recent resource Sisson et al. (2015).

4.3.4 Variational Bayes (VB)

The variational Bayes (VB) technique is a procedure which, in contrast to the previous sampling based approaches, produces an analytical approximation to the posterior distribution $p(x|y)$. The approximated posterior $p^*(x|y)$ is the minimiser of the Kullback–Leibler divergence $\mathrm{KL}(q(x), p(x|y))$ given in (2.3), where $q(x)$ is a restricted class of probability distributions. The most common VB procedure, known as mean field VB, restricts the approximation class to distributions $q(x)$ which factorises into $k \leq \dim(x)$ different distributions, i.e. $q(x) = \prod_{j=1}^{k} q_j(x_j)$, but lays no other distributional assumptions on the approximation class. In such a situation, a general expression for the optimal solution is $p^*(x|y) = \prod_{j=1}^{k} q_j^*(x_j)$, where

$$q_j^*(x_j) \propto \exp(\mathrm{E}_{q_{-j}}[\log\{p(y|x)p(x)\}]), \tag{4.4}$$

where $\mathrm{E}_{q_{-j}}[\log\{p(y|x)p(x)\}] = \int \log\{p(y|x)p(x)\} \prod_{i \neq j}\{q_i(x_i)\,\mathrm{d}x_i\}$ denotes the expectation with respect to all factors of q except the j-th. Thus, the optimal solution will depend on the data y, and the parameters in the likelihood $p(y|x)$ and the prior $p(x)$. Except for a few special cases, the q_j^* in (4.4) does not take explicit forms. A recursive algorithm maximising and then updating one factor at a time may be employed. This algorithm is closely connected to the expectation maximisation (EM) algorithm (Dempster et al., 1977), and is guaranteed to converge to a (local) optimum due to the convexity of the optimisation problem.

The number of factors and precise subdivision of them are key in the mean field VB approach. As a consequence of the setup, any dependence between factorised variables is ignored. This generally results in underestimation of the variability in the joint posterior distribution. Thus, variables whose dependence is crucial should not be factorised. Fewer factors leads to a theoretically better approximation (in terms of the Kullback–Leibler divergence), but contrarily also a more difficult and computationally more time consuming optimisation problem. Hence, the factorisation of the variables should be guided by knowledge of the likelihood and prior model. Consider the case where the latent variables x correspond to spatial locations. When the locations are clustered, the factorisation is natural and the VB procedure should be very much suitable. When the locations are (approximately) given on a dense grid, it is typically more difficult to find a proper grouping of the variables. The approach is therefore perhaps less suited for such cases. This is often, but not always, the case for applications like in Section 4.1. For a thorough review of the VB approach from a pragmatic machine learning perspective, see e.g. Beal (2003). For a more careful probabilistic introduction, see e.g. Bishop (2006, Ch. 10).

Expectation propagation (EP) is another closely connected analytic approximation procedure. The EP procedure approximates the posterior distribution $p^*(x|y)$ by minimising the Kullback–Leibler divergence $\mathrm{KL}(p(x|y), q(x))$, i.e. the divergence from the true posterior $p(x|y)$ to a class of probability distributions $q(x)$ – opposite of the direction in mean field VB. In practice, the optimisation problem turns out rather different, however, and there is particularly no convergence guarantee for the standard iterative optimisation algorithm. See e.g. Minka (2009) for a thorough review of the EP procedure.

4.4 Approximate Bayesian inference within the geosciences

There appears to be few examples of the aforementioned Bayesian approximation procedures being directly applied to inverse problems within the geosciences. On the other hand, there is a large number of papers within the geoscience community which develop machinery to handle the application specific inverse problems. We review some of these methodological developments below.

Recalling the notation and setup in Section 4.1, the inversion problem is often viewed as a composition of two inversion problems: Geophysical inversion (inversion from geophysical data d to geophysical properties m) and rock physics inversion (inversion from geophysical properties

m to rock properties r). There are therefore essentially two approaches for dealing with the full inversion problem: The sequential two-step approach which handles the two problems separately, and the joint or simultaneous approach which performs full inversion in a single step. The joint (Bayesian) statistical approach is typically considered the most appropriate when the ultimate interest is indeed in the (latent) rock properties r (Bosch et al., 2010). Although it shall not concern us as such, note that there is a rich literature also on fully deterministic inversion procedures typically based on local or global optimisation, see e.g. Sen & Stoffa (2013) for an overview. Such approaches appear to be in less frequent use nowadays, possibly since such uncertainty-ignorant methodology is considered less appropriate (Francis, 2006a,b).

Buland & Omre (2003) describes a widely used geophysical inversion approach, often referred to as (linearised) Gaussian inversion. The approach utilises a direct Gaussian approximation for inverting seismic AVO data $y = d$ to elastic parameters and density $x = m$. More specifically, under the common assumption that $p(y|x)$ is Gaussian with a mean function which is linear in x and a covariance matrix which is independent of x, Buland & Omre (2003) approximates $p(x)$ by a Gaussian. This is a convenient, but typically very rough approximation. By this approximate construction, $p(x|y)$ is Gaussian with linear dependence on the data $y = d$ and a fixed covariance matrix – and thereby computationally extremely efficient. Rimstad & Omre (2014a,b) improves the approximation accuracy by extending the approximate prior $p(x)$ to the more general class of selection Gaussian distributions (Arellano-Valle et al., 2006). This class allows for skewness and multimodality while inheriting beneficial properties of the Gaussian distribution. Although the resulting approximation to the posterior $p(x|y)$ is computationally harder to evaluate, it is still analytically tractable. Grana & Della Rossa (2010) use a conceptually similar approach using Gaussian mixtures for inversion from geophysical data $y = d$, all the way to the rock properties $x = r$.

Some of the methodologies within the field are developed to deal with situations where the latent variable x is lithology or facies, i.e. a categorical variable. Several of these approaches are based on Larsen et al. (2006) which introduces discrete Markov dependencies for the spatial distribution of x, and includes an initial step inverting the geophysical data d to geophysical properties m using Gaussian inversion (Buland & Omre, 2003), even if the distribution of m is typically multimodal. A correction is subsequently made for this misspecification. There also exist methodologies which take the seismic processing uncertainty into account, see e.g. Houck (2002). For a more thorough review of (approximate) inversion procedures within the geosciences, see Bosch et al. (2010); Doyen (2007).

The main reason for the separate development of the geoscientific methodology lies in the computational feasibility. There appears to be a gap between the speed attained when applying the general methodology surveyed in previous subsections to inverse problems within the geosciences, and what is acceptable for the industry (Mosegaard & Tarantola, 2002). The MCMC approach seems to be the only general procedure which is utilised to some extent, see e.g. Eidsvik et al. (2004); Hammer et al. (2012); Malinverno (2002); Mosegaard & Tarantola (1995). These procedures are however mainly applied to subsets of the global inverse problem, and even then, they are quite time consuming. As indicated above, the geoscience specific inversion procedures are often somewhat rougher in nature than the more general procedures. Although one would always wish for as accurate a method as possible, this is a price the community has found it necessary to pay for computational feasibility. Furthermore, as the solutions to the inverse problems within the geosciences typically are used indirectly to make decisions, minor inaccuracies and biases may not be crucial. That being said, too rough and inaccurate methodology may lead to severely wrong conclusions. Hence, there is still a need for computationally feasible methodology with improved accuracy.

5 Summary of papers

5.1 Paper I

Jullum, M. & Hjort, N. L. (2016). Parametric or nonparametric: The FIC approach. ***Submitted for publication in Statistica Sinica***

Paper I founds a principally new branch of focused information criteria (FIC). This involves developing a strategy for constructing model selection criteria for selecting among a set of parametric models and a nonparametric alternative, specifically tuned towards optimal estimation of a pre-specified focus parameter μ. In contrast to earlier developments of FIC, this strategy does not utilize any kind of local misspecification framework. The criteria are rather derived completely without assumptions relating the parametric models to the true data generating distribution, allowing in particular the parametric candidate models to be non-nested. Generally speaking, a focus parameter μ, with true unknown value $\mu_{\rm true}$, can be estimated nonparametrically by $\widehat{\mu}_{\rm np}$ and by parametric estimators on the generic form $\widehat{\mu}_{\rm pm}$. Under suitable regularity conditions, we find that $\Lambda_{{\rm np},n} = \sqrt{n}(\widehat{\mu}_{\rm np} - \mu_{\rm true}) \to_d {\rm N}(0, v_{\rm np})$, and $\Lambda_{{\rm pm},n} = \sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_0) \to_d {\rm N}(0, v_{\rm pm})$ for each parametric model having its own least false focus parameter value μ_0. These limit distributions motivate the following types of approximate mean squared error formulae: ${\rm mse}_{\rm np} = n^{-1} v_{\rm np}$ and ${\rm mse}_{\rm pm} = b^2 + n^{-1} v_{\rm pm}$, where $b = \mu_0 - \mu_{\rm true}$ is the bias associated with using the parametric model. The joint limit of $\Lambda_{{\rm np},n}$ and $\Lambda_{{\rm pm},n}$ is then utilized to establish FIC scores being estimates of these approximate mean squared errors, which then rank the models and corresponding estimators accordingly. Provided the above limit distributions hold for a full set of focus parameters, the strategy extends to an average weighted criterion (AFIC) by essentially integrating over the FIC scores of the (possibly weighted) set of focus parameters.

The above strategy is used to derive FIC and AFIC in the i.i.d. setting for focus parameters which are Hadamard differentiable functionals of the distribution function. The asymptotic properties and behaviour of the schemes are studied, and extensions to other data types are sketched. In particular, we observe that when all parametric models are misspecified with respect to the focus parameter, the nonparametric model wins with a probability tending to one as the sample size increases. In addition, we show that application of the criteria may be viewed as implicit focused hypothesis tests having asymptotic significance level chosen by the theory itself. For the FIC, this level is $P(\chi^2_1 > 2) \approx 15.7\%$ independently of the focus parameter. For a particular application of AFIC to categorical or categorised data, such theoretical behavioural results shed new light on the classical Pearson chi-squared test. We also propose a model averaging routine which uses a weighted average across all candidate models as a final estimator of μ – with weights based on the obtained FIC scores. Finally, we discuss extensions and generalisations to a range of other data frameworks and situations, including density estimation and regression. The supplementary material following the paper (Jullum & Hjort, 2015b) includes, among other things, a simulation study showing promising results when comparing the FIC and AFIC to other information criteria. An asymptotic comparison with the original version of the FIC is also provided.

5.2 Paper II

JULLUM, M. & HJORT, N. L. (2015a). What price semiparametric Cox regression? ***Submitted for publication in Scandinavian Journal of Statistics***

The objective of Paper II is twofold. The first part concerns studying the ‘price paid’ by relying on the semiparametric Cox regression model as opposed to a model with a parametrically specified baseline hazard function, when the latter indeed is correct. This is accomplished by deriving, and then investigating, the limiting distributions of various estimators based on the two model types. These are then used to compute the asymptotic relative efficiency (ARE) for the exponential and Weibull models with different censoring proportions and covariate effects. We find that the efficiency gain of using more restrictive, fully parametric modelling approaches depends heavily on the specific quantity being estimated, in addition to the model complexity and the amount of censoring. When the baseline hazard is constant, we find that trusting parametrics to estimate large and small (conditional) cumulative hazards, survival probabilities, and quantiles gives a substantial gain. On the other hand, there is little to gain when estimating small regression coefficients. A certain outside-model-conditions extension of the ARE is also briefly examined.

The second part concerns development of methodology which can handle such selection problems in practice. Based on the strategy of Paper I, this is carried out by constructing FIC and AFIC schemes for the Cox regression setting, which allows for simultaneous focused model selection among a set of fully parametric alternatives and the semiparametric Cox regression model. The criteria also cover the case without covariates, where the nonparametric candidate corresponds to a transformation of the Nelson–Aalen or Kaplan–Meier estimator. The criteria are illustrated through applications to survival analysis data for patients with oropharynx carcinoma. Asymptotic properties and behaviour results similar to those in Paper I are also derived.

5.3 Paper III

HERMANSEN, G. H., HJORT, N. L. & JULLUM, M. (2015). Parametric or nonparametric: The FIC approach for stationary time series. ***Technical report, Department of Mathematics, University of Oslo***

Paper III uses the strategy of Paper I (and II) to construct FIC and AFIC schemes for stationary Gaussian time series, which selects among a set of fully parametric models and a nonparametric candidate model. Concentrating on focus parameters connected to the dependence structure of the time series, we restrict ourselves to detrended time series. In particular, we use some efforts to show that under weak conditions, the detrended time series may be handled theoretically as if it was the original series. The nonparametric estimator is based on an estimator of the periodogram, while the parametric models are treated generally and fitted either via maximum likelihood, or using the Whittle approximation. Although the methodology applies to general parametric models, we typically work with the most familiar autoregressive AR(p) and moving average MA(q) models. The FIC and AFIC criteria are developed for classes of focus parameters containing, in particular, covariance and correlations lags and intervals of the integrated spectrum. We provide a 'proof of concept' illustration, in addition to a brief simulation study to illustrate the practical performance of the derived FIC scheme. Asymptotic behavioural investigations show that also this FIC scheme behaves much like those in Papers I and II. Various pointers to further work are provided, including treatment of focus parameters depending on covariates and trends.

5.4 Paper IV

JULLUM, M. & KOLBJØRNSEN, O. (2016). A Gaussian-based framework for local Bayesian inversion of geophysical data to rock properties. ***Accepted for publication in Geophysics***

Paper IV develops a new procedure for approximate Bayesian inversion, specifically constructed to handle inversion from geophysical data d to the latent rock properties r. The model structure in question corresponds to that of Figure 4.1, involving also geophysical properties m. Rather than approximating a full, extremely high dimensional posterior distribution $p(r|d)$, we concentrate on approximating marginal posterior distributions for each single cell of the latent field r. Thus, we sequentially focus on each location of the rock property of interest and approximate its posterior distribution. The curse of dimensionality challenge is handled by only using the parts of the variables (r, m, d) which are spatially close and important for the cell being inverted. Where to draw the border for variable inclusion is a question of balancing methodological accuracy and computational speed. We typically assume that $p(d|m)$ is Gaussian with a linear mean and fixed covariance matrix, and also that we are able to sample locally from $p(m, r)$. Utilising Gaussian distribution theory, we construct a Gaussian approximation to a low dimensional local version of $p(d|r)$. This approximate Gaussian likelihood has a flexible dependence structure being modelled by a general nonlinear regression scheme which is fitted to samples from $p(m, r)$. A weighted Monte Carlo procedure is used to numerically solve a resulting lower dimensional integral, which finally approximates all marginal posterior distributions.

The procedure is illustrated on both synthetic and real CO_2 monitoring cases related to the Sleipner CO_2 injection project. Based on time lapse seismic AVO data, the project aims at monitoring CO_2 which has been injected in the Utsira formation at the Sleipner field offshore Norway. We use our procedure to approximate the spatial dispersion of the injected CO_2 through the CO_2 saturation. The accuracy of the approximations is evaluated both by comparison to the synthetic truth, and to the posterior distribution obtained by an MCMC procedure. The approximation accuracy of the central tendency and the coverage of credibility intervals are both considered adequate. We also illustrate that our procedure significantly improves the ability to resolve features compared to the widely used direct Gaussian inversion approach of Buland & Omre (2003). The inversion results also match well with previously published qualitative interpretations for the real case.

6 Discussion

There are several different directions for which the methodology in the four papers can be extended, some of which are discussed in the papers themselves. Several of these directions are listed below, after which I point to a few more. Then, rather than touching the surface of all of the possible discussion topics relevant for this thesis, I concentrate on two topics related to the papers, in addition to the recurring and non-recurring themes. First, however, I summarise the main contributions of each of the papers.

The main contribution of Paper I is the introduction and construction of the new FIC-paradigm which enables focused model selection among statistical models with and without likelihoods. While Paper I pans out the strategy and focuses on the theoretic properties and development for the simplest i.i.d. data situation, the main contributions of Papers II and III are to extend the scope of that apparatus to proportional hazard regression with censoring, and to stationary time series. Combined, these three papers allow focused model selection to be carried out, via natural and intuitive criteria, for a wide range of data and model settings where appropriate statistical comparison and selection were not previously possible.

The main contribution of Paper IV is the development of a new procedure for approximate Bayesian inversion within the geosciences. By localising the problem and utilising the parallelisation properties of the methodology, the procedure gives approximate local solutions. The obtained (adjustable) combination of accuracy and speed is out of reach for existing methodology.

6.1 Extensions and further work

The below list shows directions for further work mentioned within the papers themselves:

I Allow for other parametric estimation techniques like M-estimation; derive methodology for handling candidate models with different convergence rates; derive methodology for more general loss functions.

II In addition to the extensions mentioned in I, allow for implicit covariate selection and time-dependent covariates.

III In addition to the extensions mentioned in I, lift the framework to non-Gaussian time series; derive methodology for handling focus parameters related to the trend function.

IV Approximate short range dependencies between different marginal posterior distributions; replace either of the Gaussian approximations by Gaussian mixtures or the selection Gaussian distribution of Arellano-Valle et al. (2006).

More on these extensions may be found in the discussion and concluding remarks sections in the respective papers.

Like most previous FIC procedures, those derived in Papers I-III rely on estimates of first order approximations (to the mean squared error). A natural extension of this is to consider second order approximations, as discussed in Tsai (2003); Hjort & Claeskens (2003b). The mean squared error formulae for both the nonparametric and parametric models would then get additional, possibly non-zero, squared bias and variance terms of size $O_p(n^{-2})$, which ought to be estimated in renewed FIC formulae. Such an extension should intentionally lead to better small samples properties. This *may* also turn the other way, however, as estimation of the additional quantities introduces more variability in the final FIC score, being particularly crucial for small sample sizes.

From a practical point of view, easing the access to the methodology and making it more readily available for practitioners, would be a vital contribution. When working with each of the four papers, fairly general functionality has been prepared in the statistical programming language R (R Core Team, 2015). To make the FIC methodology more accessible, I hope to gather the FIC-related functionality in an R-package to be put on 'CRAN'. For the methodology in Paper IV, a first step would rather be an implementation in a pre-compiled language which permits efficient use of large number of high performance cores for parallelisation, when such are available.

Finally, the generality of the method presented in Paper IV allows for applications, not only to CO_2 monitoring, but also in general reservoir characterisation and exploration. However, it ought to be worth investigating the possibilities for applying this methodology to more general Bayesian inversion problems, also outside the geophysical/rock physical inversion setting. In the absence of relevant expertise, it is difficult to list other fields to which our methodology would apply, but 'interpretation' of X-ray, magnetic resonance (MR), and related data appear to be potential candidates.

6.2 New vs. original FIC

Whereas essentially all previous developments of the FIC have been conducted inside local misspecification frameworks, those in Papers I, II and III are developed without such assumptions – as described in the above summary of Paper I. This, in addition to our inclusion of a non-parametric (or semiparametric) candidate model, is the main difference between the two FIC construction strategies which ought to be discussed and compared below.

The local misspecification framework underlying the original FIC and AFIC have received some critique. Raftery & Zheng (2003); Ishwaran & Rao (2003) indicate that the framework is unrealistic, and suggest it does not yield a valid environment for model selection. As pointed out in Hjort & Claeskens (2003b) and Claeskens & Hjort (2008a, Remark 5.3), the framework is not intended to be completely trusted per se, but should merely be perceived as a construction to derive sound asymptotic results, which then are estimated and interpreted in a usual finite sample fashion. The critique is however not completely uncalled for. Although one reaches neat asymptotic results which are perfectly valid within this framework, there is no guarantee that these will hold or be informative in practical applications. Further, even if one accepts that the results are informative when the true model is close to all candidates, they are hardly of much relevance if one or more models are far off. Nonetheless, similar frameworks *have* been utilised in other settings and seem to be acknowledged as informative; see e.g. Lehmann (1998, Ch. 3.3) for an application to a study of test power.

On another note, the local misspecification framework requires all competing models to be nested. That is, they all need to be special cases of a 'wide' model where all γ-parameters are being estimated, and, at the same time, extensions of a 'narrow' model without these γ-parameters. When none of the candidate models encompass all the others, the natural solution is to define such a 'wide' model. This strategy requires a number of extra 'nuisance' parameters to be defined, solely for the purpose of making the asymptotics work out. If several such extra parameters need to be estimated as part of the FIC apparatus, the estimation accuracy of the FIC scores may be significantly weakened. Differently constructed wide models may also give different FIC scores. A benefit of the local misspecification framework is that it provides asymptotics for quite general model average estimators (Hjort & Claeskens, 2003a). In particular, this allows for quantification of the post selection uncertainty, i.e. the uncertainty associated with the model selection step which is often ignored; cf. Breimans 'quiet scandal of statistics' (Breiman, 1992).

A conceptual advantage of our FIC approach is that we do not make any distributional assumptions relating the parametric candidate models to the (unknown) true model. Furthermore, since our approach does not rely on a local misspecification framework, the approach is immune to the aforementioned critique. The introduction of the nonparametric model also has several beneficial consequences. One of them is that it gives the criteria an insurance mechanism against poorly specified parametric models – a property virtually no other type of information criterion possess. Principally speaking, our criteria typically select fine-tuned parametrics when they are adequate, and trust nonparametrics when the parametric models are far off. Another benefit is that it, in contrast to the original FIC, gives model robust estimates of the bias involved in the parametric model. A possible drawback of trusting the nonparametric model and estimator so heavily, is that the FIC scores become more sensitive to changes in the data. Smoothing may help for that matter, though. Under appropriate and weak regularity conditions, Fernholz (1991) shows that the integrated kernel estimator (cf. (2.7)) has the same asymptotic properties as the empirical distribution function, allowing nonparametric estimation to be based upon the former instead.

Another possible drawback of our more flexible FIC approach is that it is rather difficult to extend the approach to the classical regression setup and density estimation – extensions which are fairly straightforward for the original FIC. The additional complexity for our approach is caused by the nonparametric estimator, whose convergence rate is slower in these situations. The joint limit distribution of parametric and nonparametric focus parameter estimators then takes a different form, having further complicating consequences for estimation of the approximated mean squared errors. These situations essentially require a slightly different approach, see Jullum & Hjort (2016, Sections 7.4 and 7.5).

In the supplementary material to Paper I, we compare the original FIC approach to our new proposed FIC on the former's 'home turf', i.e. in the local misspecification framework. Assuming the parametric models are indeed nested, we find precise limiting distributions for the two types of FIC formulae, allowing us to compare the two criteria in this local asymptotics framework. We find that the two schemes are asymptotically equivalent (in the local misspecification limit sense) when the widest parametric model has the same variance as the nonparametric estimator. In addition, the new parametric FIC scores may be seen as model robust versions of those in the original FIC.

Claeskens & Hjort (2003, Sec. 8) touch upon, but does not fully develop, a robust version of the FIC, denoted the focused robust information criterion (FRIC). This criterion uses a more general local misspecification framework, whose most general variant is completely unrelated to the candidate models. Although it is not emphasised by the authors themselves, such a framework should in principle work with non-nested candidate models. The FRIC has not received much attention in the consecutive literature, though.

6.3 Localising the inverse problem

The distinguishing feature of the procedure in Paper IV compared to most of those surveyed in Section 4.3, is the decomposition of the global problem into many local problems. That is, rather than dealing with the global posterior of the latent field, the procedure in Paper IV approximates the marginal posterior distribution in each of the cells in the latent field. With the variables involved being connected to certain locations, the data and model are usually designed such that short range dependencies dominate.[1] Hence, when concentrating on the marginal posterior distribution in a certain location, variables connected to locations far from the current cell would have a minimal impact. The dimensionality of the local problem is effectively reduced by ignoring these low-impact variables. This protects against the curse of dimensionality and is a key ingredient in Paper IV, as well as in Jullum & Kolbjørnsen (2015).

Localising the problem is indeed appropriate when local behaviour of the latent field is the focus of the study. In the geoscience setting this translates to the case where cell-wise inference about the rock properties are key – in contrast to drawing probabilistic statements about long range dependencies, performing flow calculations, and so on. The latter situation essentially requires a global solution. Note however that a straightforward extension of our approach allows our methodology to approximate the joint posterior distribution in smaller neighbourhoods. This enables also approximations of short range dependencies between the cells of the latent field, which may be of interest in certain applications.

The local inversion approach handles the local problems one by one. Under stationarity conditions, this allows for heavy parallelisation, ensuring that the methodology scales well. That is, each of the separate local inversion problems may be handled by a separate core on the computer when implemented in software allowing for parallelisation. This gives a significant and beneficial speed-up when ran on a laptop or desktop computer with say four or possibly eight cores. The real advantage comes into play when running the methodology on large computer clusters, which may have several thousands of cores working in parallel. This is a substan-

[1] If this is not the case, it is typically possible to apply resolution theory to focus the energy, see Appendix C in Paper IV.

tial advantage over non-parallelisable methodology which needs to handle one task at the time, especially for the large scale problems we are concerned with.

Although the localisation of the problem incurs an error, it seems to give fairly reasonable results in a situation where the number of appropriate alternatives are few. It is evident from the investigation in Paper IV that, in terms of predictive performance, our approach outperforms the frequently applied direct Gaussian inversion approach of Buland & Omre (2003). There is also an incredible speed-up compared to 'brute-force' MCMC. For a particular inversion problem of dimension 140, our approach gives results in seconds, while the MCMC implementation uses days to deliver reliable results.

Note finally that localisation of the geophysical/rock physical inversion problem is not an entirely new concept, being a key component in for instance Buland et al. (2008). In fact, a variant of the approach in Buland et al. (2008) may be seen as a special case of our approach.

6.4 The leitmotifs and frequentist vs. Bayesian statistics

> "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."
>
> *John Tukey*

This quote sort of sums up the underlying intention of the methodologies developed in this thesis. Asking the right question is indubitably essential whether one is faced with a model selection problem, an inverse problem, or any other kind of inference problem. From my point of view, focusing the energy on the precise question with which one is concerned, is a natural and indispensable principle which should be followed in all inferential problems.

Focus. The focused view is certainly glaring in the model selection methodology in Papers I-III. After forcing the data analyst to sharpen his or her question(s) by describing one or more focus parameters, the FIC and AFIC apparatuses concentrate unconditionally on finding the model with the best estimate(s) of the μ. Additionally, in the approximation Bayesian inversion methodology in Paper IV, the challenge is focused and reduced from a global inverse problem to several local inverse problems. The methodology is then tuned specifically towards approximations of the marginal posterior distribution in each of the cells, for the rock property chosen for scrutiny. It is irrelevant whether or not the various likelihood approximations are appropriate on a global scale, or for the neighbouring cell, as long as the variables related to the cell in focus are reasonable.

Approximation. Despite the technological developments in recent time, computational advances has not yet reached a level where inferential approximations are avoidable – and probably never will. In the model selection frameworks we employ the powerful toolbox of asymptotic theory to work around the finite sample distributional complexities. We particularly rely on finite sample approximations to the mean squared errors which generally hold only as the sample size tends to infinity. In Paper IV there are several layers of approximation, including discarding variables far from the cell under current inversion, and approximating the general likelihood by a Gaussian distribution.

Finally, working with both frequentist and Bayesian methodology, I am sometimes asked how I can 'play for both teams' and whether I am *really* a frequentist or a Bayesian. I then reply saying I am the 'glory-hunter' fan who cheers for the frequentists when they are doing well, and the Bayesians when they are doing well. Despite my efforts, I do not understand those claiming one needs to choose one or the other. The way in which the Bayesian approach enables quantifiable prior information to be incorporated into the inferential analysis really speaks to me. If information (for instance from past studies) is really there, it seems absurd not to utilise it. On the other hand, when such information is not available, I find the frequentist approach more appealing than the Bayesian solution of constructing non-informative (objective) priors. The use of flat priors, for instance, often carry with them unintentional side-effects (Simpson et al., 2014). I also find it hard to argue against the frequentist approach to hypothesis testing – except possibly for the artificial way in which the threshold value typically is set. Both approaches have their strengths and weaknesses, and I allow myself the luxury of using the best pieces of the two worlds. Like Irizarry (2014), I would be happy to see the Bayesian vs. frequentist debate to be declared over.

In summary, with our frequentist and Bayesian approaches, we obtain **approximate** answers to the precise **focused** questions being raised – just as John Tukey prefers.

References

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* **6**, 701–726.

AALEN, O., BORGAN, Ø. & GJESSING, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Heidelberg: Springer-Verlag.

AALEN, O. O. (1975). *Statistical inference for a family of counting processes*. Ph.D. thesis, Univ. of California, Berkeley.

AALEN, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907–925.

AITKEN, A. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* **55**, 42–48.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

ALDRICH, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science* **12**, 162–176.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Heidelberg: Springer-Verlag.

ANDERSON, T. W. & DARLING, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* **23**, 193–212.

ARELLANO-VALLE, R. B., BRANCO, M. D. & GENTON, M. G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics* **34**, 581–601.

AUSÍN, M. C. (2014). *Quadrature and Numerical Integration*. John Wiley & Sons, Ltd.

AVSETH, P., MUKERJI, T. & MAVKO, G. (2010). *Quantitative Seismic Interpretation: Applying Rock Physics Tools to Reduce Interpretation Risk*. Cambridge University Press.

BARRON, A. R. & SHEU, C.-H. (1991). Approximation of density functions by sequences of exponential families. *The Annals of Statistics* **19**, 1347–1369.

BARTOLUCCI, F. & LUPPARELLI, M. (2008). Focused information criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistics* **35**, 629–649.

BAYES, M. & PRICE, M. (1763). An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. *Philosophical Transactions* **53**, 370–418.

BEAL, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, University College London.

BEAUMONT, M. A., ZHANG, W. & BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

BEHL, P., CLAESKENS, G. & DETTE, H. (2014). Focused model selection in quantile regression. *Statistica Sinica* **24**, 601–624.

BEHL, P., DETTE, H., FRONDEL, M. & TAUCHMANN, H. (2012). Choice is suffering: A focused information criterion for model selection. *Economic Modelling* **29**, 817–822.

BERNOULLI, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis*. Thurneysen Brothers.

BÖHMER, P. (1912). Theorie der unabhängigen Warscheinlichkeiten. *Rapports, Mémoires et Procés-verbaux de Septiéme Congrés International d'Actuaries, Amsterdam* **2**, 327–343.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

BLANGIARDO, M. & CAMELETTI, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R - INLA*. Wiley.

BLUM, M. G. B., NUNES, M. A., PRANGLE, D. & SISSON, S. A. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science* **28**, 189–208.

BORGAN, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics* **11**, 1–16.

BOSCH, M., MUKERJI, T. & GONZALEZ, E. F. (2010). Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: A review. *Geophysics* **75**, 75A165–75A176.

BREIMAN, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

BRESLOW, N. (1972). Contribution to the discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society Series B* **34**, 216–217.

BRILLINGER, D. R. (1975). *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston.

BROCKWELL, P. & DAVIS, R. (1991). *Time Series: Theory and Methods*. Springer.

BROWN, L., CAI, T., ZHANG, R., ZHAO, L. & ZHOU, H. (2010). The root–unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields* **146**, 401–433.

BROWNLEES, C. T. & GALLO, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. *Journal of Financial Econometrics* **6**, 513–539.

BROWNLEES, C. T. & GALLO, G. M. (2011). Shrinkage estimation of semiparametric multiplicative error models. *International Journal of Forecasting* **27**, 365 – 378.

BULAND, A., KOLBJØRNSEN, O., HAUGE, R., SKJÆVELAND, Ø. & DUFFAUT, K. (2008). Bayesian lithology and fluid prediction from seismic prestack data. *Geophysics* **73**, C13–C21.

BULAND, A. & OMRE, H. (2003). Bayesian linearized AVO inversion. *Geophysics* **68**, 185–198.

CANTELLI, F. (1933). Sulla determinazione empirica di una legge di distribuzione (on the empirical determination of a probability law). *Giornale dell'Istituto Italiano degli Attuari* **4**, 421–424.

CHERNOFF, H. & LEHMANN, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics* **25**, 579–586.

CLAESKENS, G. & CARROLL, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249–265.

CLAESKENS, G., CROUX, C. & VAN KERCKHOVEN, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62**, 972–979.

CLAESKENS, G., CROUX, C. & VAN KERCKHOVEN, J. (2007). Prediction focussed model selection for autoregressive models. *The Australian and New Zealand Journal of Statistics* **49**, 359–379.

CLAESKENS, G. & HJORT, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98**, 900–916.

CLAESKENS, G. & HJORT, N. L. (2008a). *Model selection and model averaging*. Cambridge University Press.

CLAESKENS, G. & HJORT, N. L. (2008b). Minimising average risk in regression models. *Econometric Theory* **24**, 493–527.

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B* **34**, 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

CRAMÉR, H. (1928). On the composition of elementary errors. *Skandinavisk Aktuarietidskrift* **11**, 141–180.

DAHLHAUS, R. & WEFELMEYER, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Annals of Statistics* **24**, 952–973.

DE MOIVRE, A. (1733). Approximatio ad summam terminorum binomii $(a + b)^n$ in seriem expansi. Tech. rep., Printed for private circulation.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

DOYEN, P. (2007). *Seismic Reservoir Characterization: An Earth Modelling Perspective*. Netherlands: EAGE.

DURBIN, J. (1973a). *Distribution Theory for Tests Based on the Sample Distribution Function*. New York: SIAM.

DURBIN, J. (1973b). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics* **1**, 279–290.

EFRON, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association* **83**, 414–425.

EIDSVIK, J., AVSETH, P., OMRE, H., MUKERJI, T. & MAVKO, G. (2004). Stochastic reservoir characterization using prestack seismic data. *Geophysics* **69**, 978–993.

FERNHOLZ, L. T. (1991). Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics* **18**, 255–262.

FRANCIS, A. M. (2006a). Understanding stochastic inversion: Part 1. *First Break* **24**.

FRANCIS, A. M. (2006b). Understanding stochastic inversion: Part 2. *First Break* **24**.

GANDY, A. & HJORT, N. L. (2013). Focused information criteria for semiparametric linear hazard regression. Tech. rep., Department of Mathematics, University of Oslo.

GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

GLIVENKO, V. (1933). Sulla determinazione empirica di una legge di distribuzione (on the empirical determination of a probability law). *Giornale dell'Istituto Italiano degli Attuari* **4**, 92–99.

GRANA, D. & DELLA ROSSA, E. (2010). Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. *Geophysics* **75**, O21–O37.

HADAMARD, J. (1923). *Lectures on the Cauchy Problem in Linear Partial Differential Equations*. Yale University Press.

HAMMER, H., KOLBJØRNSEN, O., TJELMELAND, H. & BULAND, A. (2012). Lithology and fluid prediction from prestack seismic data using a Bayesian model with Markov process prior. *Geophysical Prospecting* **60**, 500–515.

HAND, D. J. & VINCIOTTI, V. (2003). Local versus global models for classification problems: fitting models where it matters. *The American Statistician* **57**, 124–131.

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

HERMANSEN, G. H., HJORT, N. L. & JULLUM, M. (2015). Parametric or nonparametric: The FIC approach for stationary time series. *Technical report, Department of Mathematics, University of Oslo* .

HERMANSEN, G. H., HJORT, N. L. & KJESBU, O. S. (2016). Recent advances in statistical methodology applied to the Hjort liver index time series (1859−2012) and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences* **73**, 279–295.

HJORT, N. L. (1992). On inference in parametric survival data models. *International Statistical Review* **60**, 355–387.

HJORT, N. L. & CLAESKENS, G. (2003a). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

HJORT, N. L. & CLAESKENS, G. (2003b). Rejoinder to 'The focused information criterion'. *Journal of the American Statistical Association* **98**, 938–945.

HJORT, N. L. & CLAESKENS, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association* **101**, 1449–1464.

HOERL, A. E. & KENNARD, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**, 69–82.

HOUCK, R. T. (2002). Quantifying the uncertainty in an AVO interpretation. *Geophysics* **67**, 117–125.

HÄRDLE, W. (1991). *Smoothing Techniques: With Implementation in S*. Springer, 1st ed.

HURVICH, C. M. & TSAI, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

IMBENS, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics* **20**, 493–506.

IRIZARRY, R. (2014). I declare the Bayesian vs frequentist debate over for data scientists. *Statslife, Opinion section – Royal statistical society*, 2014-10-15, `http://www.statslife.org.uk/opinion/1851`. Accessed: 2015-11-28.

ISHWARAN, H. & RAO, J. S. (2003). Discussion of 'The focused information criterion'. *Journal of the American Statistical Association* **98**, 922–925.

JONES, M. C., MARRON, J. S. & SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**, 401–407.

JULLUM, M. & HJORT, N. L. (2015a). What price semiparametric Cox regression? *Submitted for publication in Scandinavian Journal of Statistics* .

JULLUM, M. & HJORT, N. L. (2015b). Supplement to parametric or nonparametric: The FIC approach.

JULLUM, M. & HJORT, N. L. (2016). Parametric or nonparametric: The FIC approach. *Submitted for publication in Statistica Sinica* .

JULLUM, M. & KOLBJØRNSEN, O. (2015). An approximate Bayesian inversion framework based on local-Gaussian likelihoods. In *EAGE: Petroleum Geostatistics 2015 (extended abstracts)*.

JULLUM, M. & KOLBJØRNSEN, O. (2016). A Gaussian-based framework for local Bayesian inversion of geophysical data to rock properties. *Accepted for publication in Geophysics* .

KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

KHINCHIN, A. (1929). Sur la loi des grands nombres. *Comptes rendus de l'Académie des Sciences* **189**, 477–479.

KOLMOGOROV, A. (1933). Sulla determinazione empirica di una legge di distribuzione (on the empirical determination of a distribution law). *Giorn. Ist. Ital. Attuar.* **4**, 83–91.

KONISHI, S. & KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

KRISCH, A. (2011). *An introduction to Mathematical theory of inverse problems*. Springer, 2nd ed.

KRISENERGY.COM (2015). About oil and gas: Exploration. `https://krisenergy.com/company/about-oil-and-gas/exploration/`. Accessed: 2015-11-05.

KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

LARSEN, A. L., ULVMOEN, M., OMRE, H. & BULAND, A. (2006). Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model. *Geophysics* **71**, R69–R78.

LEHMANN, E. L. (1998). *Elements of Large-Sample Theory*. Berlin: Springer-Verlag.

LONGFORD, N. T. (2005). Editorial: Model selection and efficiency—is 'which model. . . ?'the right question? *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**, 469–472.

MALINVERNO, A. (2002). Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International* **151**, 675–688.

MARTINS, T. G., SIMPSON, D., LINDGREN, F. & RUE, H. (2013). Bayesian computing with inla: New features. *Computational Statistics & Data Analysis* **67**, 68 – 83.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.

MINKA, T. (2009). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology.

MOORE, D. S. & SPRUILL, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics* , 599–616.

MOSEGAARD, K. & TARANTOLA, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth* **100**, 12431–12447.

MOSEGAARD, K. & TARANTOLA, A. (2002). Probabilistic approach to inverse problems. In *In International Handbook of Earthquake & Engineering Seismology, Part A*. Academic Press.

NELSON, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1**, 27–52.

NELSON, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* **14**, 945–965.

NGUEFACK-TSAGUE, G. & BULLA, I. (2014). A focused Bayesian information criterion. *Advances in Statistics, ID 504325* **2014**, 1–8.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics* **24**, 2399–2430.

OWEN, A. B. (2001). *Empirical likelihood.* CRC press.

PEARSON, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A* **186**, 343–414.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5* **50**, 157–176.

PRENTICE, R. L. & SELF, S. G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *The Annals of Statistics* **11**, 804–813.

PRIESTLEY, M. B. (1981). *Spectral analysis and time series.* Academic press.

R CORE TEAM (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAFTERY, A. E. & ZHENG, Y. (2003). Discussion of 'The focused information criterion'. *Journal of the American Statistical Association* **98**, 931–938.

RIMSTAD, K. & OMRE, H. (2014a). Generalized Gaussian random fields using hidden selections. *arXiv preprint arXiv:1402.1144* .

RIMSTAD, K. & OMRE, H. (2014b). Skew-Gaussian random fields. *Spatial Statistics* **10**, 43 – 62.

ROBERT, C. & CASELLA, G. (2005). *Monte Carlo Statistical Methods.* Springer.

ROHAN, N. & RAMANATHAN, T. V. (2011). Order selection in arma models using the focused information criterion. *Australian & New Zealand Journal of Statistics* **53**, 217–231.

ROLLING, C. A. & YANG, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 749–769.

RUBINSTEIN, R. Y. & KROESE, D. P. (2008). *Simulation and the Monte Carlo method.* John Wiley & Sons Inc.

RUE, H., MARTINO, S. & CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* **71(2)**, 319–392.

SCHLAIFER, R. & RAIFFA, H. (1961). *Applied statistical decision theory.* Harvard University, MIT Press.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

SEN, M. K. & STOFFA, P. L. (2013). *Global Optimization Methods in Geophysical Inversion*. Cambridge University Press.

SHUMWAY, R. H. & STOFFER, D. S. (2011). *Time Series Analysis and Its Applications: With R Examples*. Springer, 3rd ed.

SILVERMAN, B. (1986). *Density Estimation for Statistics and Data analysis*. Chapman & Hall.

SIMPSON, D. P., MARTINS, T. G., RIEBLER, A., FUGLSTAD, G.-A., RUE, H. & SØRBYE, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *ArXiv preprint arXiv:1403.4630* .

SISSON, S. A., FAN, Y. & BEAUMONT, M. (2015). *Approximate Bayesian Computation: Likelihood-free Methods for Complex Models*. Chapman & Hall, 1st ed.

SMIRNOV, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics* **19**, 279–281.

STARK, P. (2000). Inverse problems in statistics. In *Surveys on solution methods for inverse problems*. Springer-Verlag, pp. 253–275.

SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods* **7**, 13–26.

SUN, Z., SU, Z. & MA, J. (2014). Focused vector information criterion model selection and model averaging regression with missing response. *Metrika* **77**, 415–432.

TAKEUCHI, T. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)* **153**, 12–18.

THISTED, R. A. (1988). *Elements of statistical computing: Numerical computation*. Chapman and Hall.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal statistical Society Series B* **58**, 267–288.

TSAI, C.-L. (2003). Discussion of 'The focused information criterion'. *Journal of the American Statistical Association* **98**, 928–930.

VAN DER VAART, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

VARIAN, H. (1975). A Bayesian approach to real estate assessment. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J.Savage*. Edward Elgar Publishing, pp. 195–208.

VON MISES, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*, vol. 3. Springer.

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer.

WATSON, G. S. & LEADBETTER, M. R. (1964). Hazard analysis II. *Sankhyā: The Indian Journal of Statistics, Series A* **26**, 101–116.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

WHITTLE, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)* **15**, 125–139.

ZELLNER, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* **81**, 446–451.

ZHANG, X. & LIANG, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics* **39**, 174–200.

ZHANG, X., WAN, A. T. K. & ZHOU, S. Z. (2012). Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business & Economic Statistics* **30**, 132–142.

I

III

PARAMETRIC OR NONPARAMETRIC: THE FIC APPROACH FOR STATIONARY TIME SERIES

Gudmund Hermansen, Nils Lid Hjort and Martin Jullum

Department of Mathematics, University of Oslo

Abstract. We seek to narrow the gap between parametric and nonparametric modelling of stationary time series processes. The approach is inspired by recent advances in focused inference and model selection techniques. The paper generalises and extends recent work by developing a new version of the focused information criterion (FIC), directly comparing the performance of parametric time series models with a nonparametric alternative. For a pre-specified focused parameter, for which scrutiny is considered valuable, this is achieved by comparing the mean squared error of the model-based estimators of this quantity. In particular, this yields FIC formulae for covariances or correlations at specified lags, for the probability of reaching a threshold, etc. Suitable weighted average versions, the AFIC, also lead to model selection strategies for finding the best model for the purpose of estimating e.g. a sequence of correlations.

Key words: focused inference, model selection, time series modelling, risk estimation

1. Introduction and summary

The focused information criterion (FIC) was introduced in Claeskens & Hjort (2003) and is based on estimating and comparing the accuracy of model-based estimators for a chosen focus parameter. This focus, say μ, ought to have a clear statistical interpretation across candidate models. For a given candidate model, μ is traditionally expressed as a function of this model's parameters. In general, the focus parameter can be any sufficiently smooth and regular function of the underlying model parameters, or more generally its spectral distribution. This includes quantiles, regression coefficients, a specified lagged correlation, but also various types of predictions and data dependent functions, to name some; see Hermansen & Hjort (2015) for a more complete list and discussion of valid focus parameters for time series models.

Suppose there are candidate models $M_1, \dots, M_k$, leading to focus parameter estimates $\widehat{\mu}_1, \dots, \widehat{\mu}_k$, respectively. The underlying idea leading to the FIC is to estimate the mean squared error (mse) of $\widehat{\mu}_j$ for each candidate model and then select the model that achieves the smallest value. The mse in question is

$$\mathrm{mse}_j = \mathrm{E}\,(\widehat{\mu}_j - \mu_{\mathrm{true}})^2 = \mathrm{bias}(\widehat{\mu}_j)^2 + \mathrm{Var}\,\widehat{\mu}_j,$$

Date: December 2015.

comprising the variance and the squared bias in relation to the true parameter value μ_{true}. Thus the FIC consists of finding ways of assessing, approximating and then estimating the mse_j for each candidate model. The winning model is the one with smallest $\widehat{\text{mse}}_j$. How this may be done depends on both the candidate models and the focus parameter, as well as on other characteristics of the underlying situation. The FIC apparatus hence leads to different types of formulae in different setups; see Claeskens & Hjort (2008, Ch. 5 & 6) for a fuller discussion and illustrations of such criteria for selection among parametric models.

Most FIC constructions have been derived by relying on a suitably defined local misspecification framework, see again Claeskens & Hjort (2008, Ch. 5 & 6). In such a framework the true model is assumed to gradually shrink with the sample size, starting from the biggest 'wide' model and hitting the simplest 'narrow' model in the limit. In addition, and all candidate models need to lie between these two model extremes. In the various data settings, such frameworks typically result in squared biases and variances of the same asymptotic order, motivating certain approximation formulae for the $\widehat{\text{mse}}_j$ in question. In Hermansen & Hjort (2015) such a framework is used to derive FIC machinery for choosing between parametric time series models within broad classes of time series models. See Section 7.5 for some further remarks.

The aim of the present paper is to derive FIC machinery which will justify comparison and selection among both parametric and nonparametric candidate models. The derivation will be somewhat different from that of Claeskens & Hjort (2003) and Hermansen & Hjort (2015) in that we do not rely on a certain local misspecification framework. We rather take a more direct approach following reasoning similar to the development of Jullum & Hjort (2015), where focused inference and model selection among parametric and nonparametric models are developed for independent observations. By including a nonparametric candidate among the parametric models, we will in particular be able to detect whether our parametric models are off-target. This FIC construction, with a nonparametric alternative, therefore has a built-in insurance mechanism against poorly specified parametric candidates. When one or more parametric models are adequate, such are selected as they typically have lower variance.

Though our methods will be extended to more general setups later, we start our developments with the class of zero-mean stationary Gaussian time series processes. Let $\{Y_t\}$ be such a process. Then the dependency structure, which in such cases determines the entire model, is completely specified by the corresponding covariance function $C(k) = \text{cov}(Y_t, Y_{t+k})$, defined for all lags $k = 0, 1, 2, \ldots$. Here we will, for mathematical convenience, work with the frequency representation, where the covariance function $C(k)$

can be represented by a unique spectral distribution G such that

$$C(k) = \int_{-\pi}^{\pi} e^{ik\omega}\,\mathrm{d}G(\omega) = 2\int_{0}^{\pi}\cos(k\omega)g(\omega)\,\mathrm{d}\omega, \tag{1.1}$$

provided the corresponding spectral distribution G has a continuous and symmetric density g. See among others Brillinger (1975), Priestley (1981) or Dzhaparidze (1986) for a general introduction to time series modelling in the frequency domain. When necessary, we will write C_g to indicate that this is the covariance indexed by the spectral density g. Note also that we can obtain the spectral density as the Fourier transform of the covariance function.

The types of parametric models we will consider are typically the classical autoregressive (AR), moving average (MA) and the mixture (ARMA), all of which have clear and well defined corresponding spectral densities; see e.g. Brockwell & Davis (1991) for an introduction to time series modelling with such models. Note that the theory developed here is general, and that there is nothing other than convenience that restricts us to these particular classes of parametric models. For an observed series $y_1, \ldots, y_n$, the raw periodogram

$$I_n(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} y_t \exp(i\omega t)\right|^2, \qquad \text{for } -\pi \le \omega < \pi, \tag{1.2}$$

will be our favourite nonparametric model for the underlying spectral density. The main reason for not considering variations of smoothed or tapered periodogram estimators is that we are interested in focus parameters that involves functions of the integrated spectrum, which essentially is a type of smoothing, rendering the pre-smoothing of the raw periodogram less critical and often unnecessary.

We will start out considering a class of focus functions of the type

$$\mu(G; h_0) = \int_{-\pi}^{\pi} h_0(\omega)\,\mathrm{d}G(\omega), \tag{1.3}$$

where h_0 is a piecewise continuous and bounded function on $[-\pi, \pi]$, with potentially a finite number of jump discontinuities. This class includes e.g. the covariance function, which is easily seen from (1.1) above, and allows studying specific parts of the spectral density by using indicator functions; see also Gray (2006) for further illustrations involving quantities of type (1.3).

Finding the best model to estimate the integrated spectrum (or total power/energy) over a specific region, may be an interesting and important applications in several areas of research; like pharmacology, astronomy, oceanography and in the interpretation of seismic data. The reason is that in all of these situations the observed time series is converted into the associated spectra, where the processed spectral density and especially the energy over certain regions of frequencies, have clear interpretations. For example, in

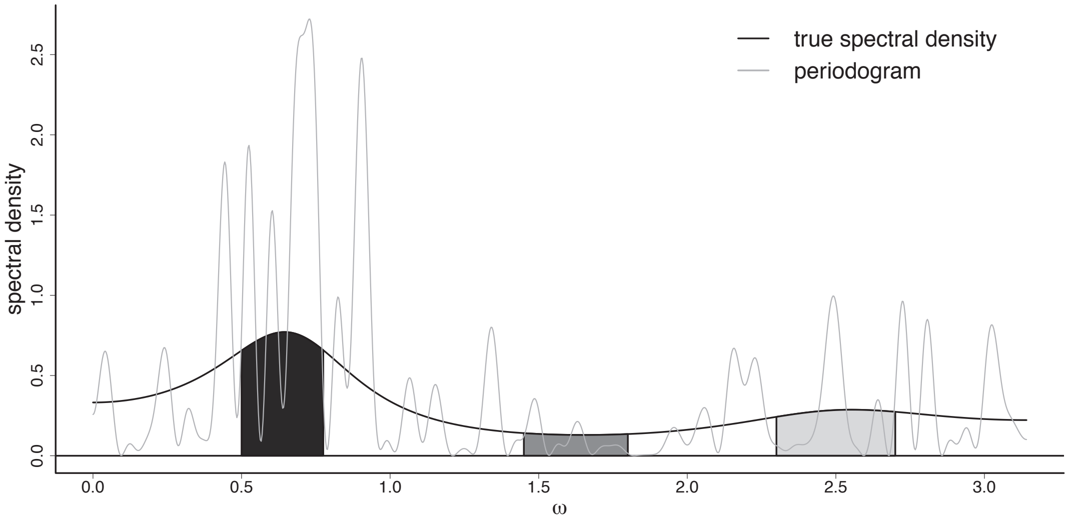

Figure 1.1. The true spectral density and the raw periodogram from a simulated autoregressive time series of order 4, with length $n = 100$ and parameters $\rho = (0.2, 0.2, -0.1, -0.2)$ and $\sigma = 1.30$. The shaded regions corresponds to three different focus parameters, namely, the integrated spectrum (or total energy) over that particular region.

pharmacology the spectrum of EEG/ERP signals may be used to quantify certain brain functions, indicating e.g. the effect of a potential drug. In such applications, the different models may not always have clear interpretations as time series, per se. The FIC is nevertheless able to rank the fitted models in terms of estimated precision of estimates, for the focus parameter in question. This general idea and particular usage of the FIC is illustrated in Figures 1.1 and 1.2 using simulated data from an autoregressive model of order 4, for focus parameters

$$\mu_j = \int_0^\pi I(a_j \leq \omega < b_j) g(\omega)\,\mathrm{d}\omega = G(b_j) - G(a_j),$$

for $j = 1, 2$ and 3, for the corresponding intervals $(a_j, b_j) \subset [0, \pi)$; which are marked by the shaded regions in Figure 1.1. The candidate models are the autoregressive models of order 0–4 and a nonparametric alternative based on integrating the raw periodogram (1.2). The AR-model of order 0 corresponds to the independence model. Here, the FIC works well: For each focus parameter it prefers models that all results in estimates that are reasonably close to the true value; which in terms if rmse (and absolute deviation from the truth) is not always the nonparametric or true model of order 4. Moreover, this example also illustrates a second and important concept, namely, that one and the same model is not necessarily best for all focus parameters. Note that the FIC prefers an AR(3), AR(4) and AR(1) for the respective regions 1, 2, 3.

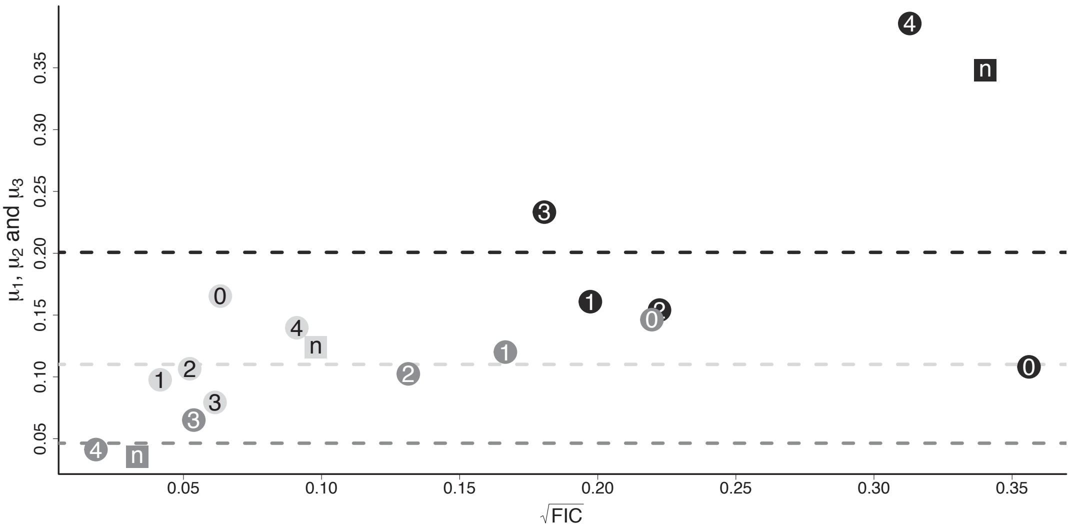

Figure 1.2. The horizontal lines indicate the true spectral density over the three shaded regions (of the same colour) shown in Figure 1.1; the three focus parameters μ_1, μ_2 and μ_3. The corresponding coloured dots show the performance, in terms of the root of the FIC score for the nonparamteric model based on the periodogram (n) and the autoregressive models of order 0–4, where 0 represent the model with independent.

A class of focus parameters wider than that of (1.3) takes focus parameters of the form

$$\begin{aligned}\mu(G; h, H) &= H(\mu(G; h_1), \ldots, \mu(G; h_k)) \\ &= H\Big(\int_{-\pi}^{\pi} h_1(\omega)\,\mathrm{d}G(\omega), \ldots, \int_{-\pi}^{\pi} h_k(\omega)\,\mathrm{d}G(\omega)\Big),\end{aligned} \tag{1.4}$$

for a k-dimensional vector function $h(\omega) = (h_1(\omega), \ldots, h_k(\omega))^{\mathrm{t}}$, where each of the h_j is of the above type, and $H(x_1, \ldots, x_k)$ a continuously differentiable function of the $x_j = \mu(G; h_j), j = 1, \ldots, k$. The direct correlations

$$\mathrm{corr}(Y_t, Y_{t+k}) = \frac{\mathrm{cov}(Y_t, Y_{t+k})}{\sigma^2} = \frac{C(k)}{C(0)} = \frac{\int_0^{\pi} \cos(k\omega)\,\mathrm{d}G(\omega)}{\int_0^{\pi} \mathrm{d}G(\omega)},$$

for example, are of type (1.4). Another class of estimands captured by (1.4) are conditional threshold probabilities, say $P\{Y_{n+1} \geq y \,|\, Y_n = y_n, \ldots, Y_{n-k} = y_{n-k}\}$, as these are functions of the $(k+1) \times (k+1)$ covariance matrix for $(Y_{n-k}, \ldots, Y_n, Y_{n+1})$. Later results will allow us to reach FIC formulae for this more general class.

In Section 2 we provide a brief overview of some standard results needed to obtain good estimates for various mean squared error quantities. Among other aspects we need properties of maximum likelihood- or Whittle approximated estimators outside the model, and some large-sample results regarding the periodogram. Then in Section 3 we motivate and develop such mean squared error estimators, leading to FIC formulae. In Section 4 we show that under certain conditions, a detrended time series may be handled by our FIC scheme as if it was the original time series. In Section 5 we extend the FIC methodology by deriving an average weighted focused information criterion which aims at selecting the best model for estimating a full set of focus parameters, possibly weighted to reflect their relative importance for the analysis. In Section 6 we discuss certain theoretical behavioural aspects of the derived FIC scheme, and present the results from a simulation study. Some concluding remarks, some of which pointing to future work, are finally provided in Section 7.

2. Estimation and approximations

We start out investigating the behaviour of the two most common parametric estimation procedures, those based on the maximum likelihood method and the associated Whittle approximation to the log-likelihood. We also give some basics for nonparametric modelling.

2.1. **Maximum likelihood estimation outside the model.** Let $\underline{y}_n = (y_1, \dots, y_n)^{\mathrm{t}}$ be a collection of n realisations from a zero mean stationary Gaussian time series process with spectral distribution function G and corresponding spectral density g. Furthermore, let the spectral distribution function F_θ and its corresponding spectral density $f_\theta = f(\cdot;\theta)$ index an arbitrary parametric candidate model, where θ belongs to some parameter space Θ of dimension say p. The corresponding full log-likelihood is

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_n(f_\theta)| - \frac{1}{2}\underline{y}_n^{\mathrm{t}}\Sigma_n(f_\theta)^{-1}\underline{y}_n, \tag{2.1}$$

where $\Sigma_n(f_\theta)$ is the covariance matrix with elements

$$C_{f_\theta}(|s-t|) = 2\int_0^\pi \cos(\omega|s-t|) f_\theta(\omega)\,\mathrm{d}\omega$$

for $s, t = 1, \dots, n$. Since the class of parametric candidate models is not assumed to necessarily include the true g, the maximum likelihood estimator does not converge to a 'true' parameter value. Instead it converges to the so-called least false parameter value, i.e. $\widetilde{\theta}_n = \mathrm{argmax}_\theta\{\ell_n(\theta)\} \to_p \mathrm{argmin}_\theta\{d(g, f_\theta)\} = \theta_0$, where

$$\begin{aligned} d(g, f_\theta) &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\Big\{\frac{g(\omega)}{f_\theta(\omega)} - 1 - \log\frac{g(\omega)}{f_\theta(\omega)}\Big\}\,\mathrm{d}\omega \\ &= -\frac{1}{4\pi}\int_{-\pi}^{\pi}\{\log g(\omega) + 1\}\,\mathrm{d}\omega - R(G,\theta), \end{aligned} \tag{2.2}$$

and where

$$R(G,\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi}\Big\{\log f_\theta(\omega) + \frac{g(\omega)}{f_\theta(\omega)}\Big\}\,\mathrm{d}\omega$$

may be referred to as the model specific part, see e.g. Dahlhaus & Wefelmeyer (1996) for details. Furthermore, it can be shown that

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d J_0^{-1}U \sim \mathrm{N}_p(0, J_0^{-1}K_0J_0^{-1}), \qquad \text{where } U \sim \mathrm{N}_p(0, K_0), \tag{2.3}$$

with J_0 and K_0 defined by

$$\begin{aligned} J_0 &= J(g, f_{\theta_0}) \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\Big[\nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}g(\omega) + \nabla^2\Psi_{\theta_0}(\omega)\{f_{\theta_0}(\omega) - g(\omega)\}\Big]\frac{1}{f_{\theta_0}(\omega)}\,\mathrm{d}\omega \end{aligned}$$

and

$$K_0 = K(g, f_{\theta_0}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}\Big\{\frac{g(\omega)}{f_{\theta_0}(\omega)}\Big\}^2\,\mathrm{d}\omega,$$

where $\Psi_\theta(\omega) = \log f_\theta(\omega)$. and $\nabla\Psi_\theta(\omega)$ and $\nabla^2\Psi_\theta(\omega)$ are respectively the vector of partial derivatives and matrix of second order partial derivatives with respect to θ, see Dahlhaus & Wefelmeyer (1996, Theorem 3.3). Note that $J_0 = K_0$ under model conditions.

2.2. **The Whittle approximation.** The Whittle pseudo-log-likelihood is an approximation to the full Gaussian log-likelihood ℓ_n of (2.1). It was originally suggested by P. Whittle in the 1950s (cf. Whittle (1953)), and is defined as

$$\widehat{\ell}_n(\theta) = -\tfrac{1}{2}n\Big[\log(2\pi) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\{2\pi f_\theta(\omega)\}\,\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_n(\omega)}{f_\theta(\omega)}\,\mathrm{d}\omega\Big], \tag{2.4}$$

where $I_n(\omega) = (2\pi n)^{-1}|\sum_{t\le n} y_t\exp(i\omega t)|^2$ is the periodogram. This approximation is close to the full Gaussian log-likelihood in the sense that $\ell_n(\theta) = \widehat{\ell}_n(\theta) + O_p(1)$ uniformly in f, see Coursol & Dacunha-Castelle (1982). More important here, however, is that (2.4) motivates an alternative estimation procedure, namely the Whittle estimator $\widehat{\theta}_n = \mathrm{argmax}_\theta\{\widehat{\ell}_n(f_\theta)\}$. This estimator is easier to work with in practice (both analytically and numerically) and shares several properties with the maximum likelihood estimator. In particular $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ achieves the same limit distribution as in (2.3), with the same least false parameter value θ_0 as defined in relation to (2.2); see Dahlhaus & Wefelmeyer (1996) for details. This means that in a large-sample perspective, the maximum likelihood estimator and the simpler Whittle estimator are equally efficient and essentially interchangeable.

2.3. **Nonparametric modelling.** As mentioned in the introduction, we shall use the periodogram in (1.2) for nonparametric modelling. Under appropriate short memory conditions, it follows from Brillinger (1975, Theorem 5.5.2) that $\mathrm{E}\{I_n(\omega)\} = g(\omega) + O(n^{-1})$, i.e. that the periodogram is asymptotically unbiased as an estimator of the spectral density. We shall thus use

$$\widehat{G}_n(\omega) = \int_{-\pi}^{\omega} I_n(u)\,\mathrm{d}u, \tag{2.5}$$

as a canonical estimator for the spectral distribution G; for which

$$\sqrt{n}(\widehat{G}_n(\omega) - G(\omega)) \to_d \mathrm{N}\Big(0, 4\pi \int_{-\pi}^{\omega} g(u)^2 \,\mathrm{d}u\Big),$$

see e.g. Taniguchi (1980).

3. Parametric versus nonparametric

We shall now obtain large-sample approximations for the focus parameter estimators. These shall then be used to construct approximate mse formulae for each model's estimator of the focus parameter. When estimated these mses then give the FIC formulae.

3.1. **How to compare parametric and nonparametric models?** In completely general terms, let $\mu(G)$ be a focus function, i.e. a functional mapping of the spectral distribution G to a scalar value. This may be estimated parametrically by estimators of the form $\widehat{\mu}_{\mathrm{pm}} = \mu(F_{\widehat{\theta}_n})$, or nonparametrically by $\widehat{\mu}_{\mathrm{np}} = \mu(\widehat{G}_n)$. Other estimators of θ and G may also be used, however. Typically, the collection of parametric candidate models does not include the true G. The question is then which model should we use – parametric or nonparametric – for estimating the focus parameter.

Assume for the nonparametric and each of the parametric candidate models that

$$\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}}) \to_d \mathrm{N}(0, v_{\mathrm{np}}) \quad \text{and} \quad \sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_0) \to_d \mathrm{N}(0, v_{\mathrm{pm}}),$$

where $\mu_{\mathrm{true}} = \mu(G)$ is the true value of the focus parameter and $\mu_0 = \mu(F_{\theta_0})$ is the focus function evaluated under the least false parametric model F_{θ_0} as discussed in relation to (2.2). Then, without going into details, the large-sample results above motivate the following first-order approximations for the mse of the estimated focus parameters:

$$\mathrm{mse}_{\mathrm{np}} = 0^2 + v_{\mathrm{np}}/n = v_{\mathrm{np}}/n \quad \text{and} \quad \mathrm{mse}_{\mathrm{pm}} = b^2 + v_{\mathrm{pm}}/n, \tag{3.1}$$

where $b = \mu_0 - \mu_{\mathrm{true}}$. The remainder of this section will be used to motivate and obtain good estimators for the mean squared errors in (3.1) with the class of focus paramters of the form $\mu(G; h_0)$ defined in (1.3), and the more general $\mu(G; h, H)$ in (1.4).

3.2. **Deriving unbiased risk estimates.** In the derivation below, the parametric candidates F_θ will be fitted using the Whittle estimator $\widehat{\theta}_n$ as defined in (2.4), while we will use the canonical periodogoram based estimator in (2.5) for nonparametric estimation of the spectral distribution G.

Using the Whittle estimator in collaboration with (2.5) results in a convenient simplification of the derivations below; extending the arguments to full ML estimation is relatively straightforward, using techniques in Dahlhaus & Wefelmeyer (1996). This motivates the following nonparametric and parametric estimators for focus parameters $\mu(G; h_0)$ on the form of (1.3):

$$\widehat{\mu}_{\rm np} = \int_{-\pi}^{\pi} h_0(\omega) I_n(\omega)\,{\rm d}\omega = \frac{1}{n} \underline{y}_n^{\rm t} \Sigma_n(h_0) \underline{y}_n \quad \text{and} \quad \widehat{\mu}_{\rm pm} = \int_{-\pi}^{\pi} h_0(\omega) f_{\widehat{\theta}_n}(\omega)\,{\rm d}\omega,$$

where $\Sigma_n(h_0)$ is a $n \times n$-dimensional symmetric Toeplitz matrix, having elements of the general form

$$\sigma_{n,s,t}(h_0) = \int_{-\pi}^{\pi} \cos(\omega|s-t|) h_0(\omega)\,{\rm d}\omega.$$

for $s, t = 1, \ldots, n$. The following proposition establishes the joint limit distribution for the estimators above (suitably normalised), which in turn will be used to obtain good approximations for their respective mean squared errors.

Proposition 1. *Let $y_1, \ldots, y_n$ be realisations from a stationary Gaussian time series model with spectral density g assumed to be uniformly bounded away from both zero and infinity. Suppose $|h_0|$ is bounded in ω, that f_θ is two times differentiable with respect to θ, and that f_θ and these derivatives, ∇f_θ and $\nabla^2 f_\theta$, are continuous and uniformly bounded in both ω and θ in a neighbourhood of the least false parameter value θ_0 as defined in (2.2) above. Then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\mu}_{\rm np} - \mu_{\rm true}) \\ \sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_0) \end{pmatrix} \to_d \begin{pmatrix} X_0 \\ c_0^{\rm t} J(g, f_{\theta_0})^{-1} U \end{pmatrix} \sim {\rm N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_{\rm np} & v_c \\ v_c & v_{\rm pm} \end{pmatrix} \right), \tag{3.2}$$

where

$$v_{\rm np} = 4\pi \int_{-\pi}^{\pi} \{h_0(\omega) g(\omega)\}^2\,{\rm d}\omega \quad \text{and} \quad v_{\rm pm} = c_0^{\rm t} J(g, f_{\theta_0})^{-1} K(g, f_{\theta_0}) J(g, f_{\theta_0})^{-1} c_0,$$

with J and K as defined below (2.3), and $v_c = c_0^{\rm t} J(g, f_{\theta_0})^{-1} d_0$, where the c_0 is the partial derivative of $\mu(F_{\theta_0}; h)$ with respect to θ, i.e. $c_0 = \nabla\mu(F_{\theta_0}; h) = \int_{-\pi}^{\pi} h_0(\omega) \nabla f_{\theta_0}(\omega)\,{\rm d}\omega$ and

$$d_0 = {\rm cov}(X, U) = \int_{-\pi}^{\pi} \frac{\nabla f_{\theta_0}(\omega) h_0(\omega) g(\omega)^2}{f_{\theta_0}(\omega)^2}\,{\rm d}\omega.$$

Proof. It follows from the results in (Dzhaparidze, 1986, Ch. 2) that $\widehat{\theta}_n - \theta_0 = J(g, f_{\theta_0})^{-1} U_n + o_p(1/\sqrt{n})$, where $U_n = \nabla \widehat{\ell}_n(f_{\theta_0})$ and

$$U_n = -\tfrac{1}{2}\{{\rm Tr}(\Sigma_n(\nabla\Psi_{\theta_0})) - \underline{y}_n^{\rm t} \Sigma_n(\nabla\Psi_{\theta_0}/f_{\theta_0}) \underline{y}_n\},$$

where $\Psi_{\theta_0} = \log f_{\theta_0}$ and $\nabla\Psi_{\theta_0}$ is the vector of its partial derivatives. As a consequence, a Taylor expansion motivated by the standard delta method gives $\widehat{\mu}_{\rm pm} - \mu_0 = c_0^{\rm t} J(g, f_{\theta_0})^{-1} U_n + o_p(1/\sqrt{n})$. Since $\sqrt{n} U_n \to_d U$ by the assumptions of the proposition (Dzhaparidze, 1986), the parametric part of the result holds. In addition

$$X_n = (\widehat{\mu}_{\rm np} - \mu_{\rm true}) = \frac{1}{n} \underline{y}_n^{\rm t} \Sigma_n(h_0) \underline{y}_n - \mu_{\rm true},$$

which can be shown, by a modified version of the argument leading to the limit distribution of U_n, to have the property that $\sqrt{n}X_n \to_d X_0 \sim \mathrm{N}(0, v_{\mathrm{np}})$. This proves the nonparametric part of the result. We finally need to show that these convergence results hold jointly. Since the two drivers in the derivation of the limit distribution, $\underline{y}_n^{\mathrm{t}}\Sigma_n(h_0)\underline{y}_n/n$ and U_n, are quadratic forms, the joint limit distribution is readily obtainable by a Cramér–Wold type of argument. To see how, let a be a vector in $\mathbb{R}^2$ to be used in the Cramér–Wold argument, and define

$$\Lambda_n = a_1\sqrt{n}X_n + a_2\sqrt{n}U_n = \frac{1}{\sqrt{n}}\underline{y}_n^{\mathrm{t}}\Sigma_n(a_1h_0 + a_2\nabla\Psi_{\theta_0}/f_{\theta_0})\underline{y}_n + \gamma_n$$

where $\gamma_n = \sqrt{n}\{a_1\mu_{\mathrm{true}} - a_2\mathrm{Tr}(\Sigma_n(\nabla\Psi_{\theta_0}))/2\}$. The γ_n cancels out the mean, here, such that Λ_n has mean zero. This is once again just a quadratic form, hence, Λ_n is normal under the assumptions of the proposition; see Dzhaparidze (1986) or Hermansen & Hjort (2014b) for derivations of a similar type. The proof is completed by observing that by Dahlhaus & Wefelmeyer (1996, Lemma A.5), the covariances take the relevant form

$$\mathrm{cov}(X_n, U_n) = \frac{2}{n}\mathrm{Tr}\{\Sigma_n(h_0)\Sigma_n(g)\Sigma_n(\nabla\Psi_\theta/f_\theta)\Sigma_n(g)\} \to \int_{-\pi}^{\pi}\frac{\nabla f_{\theta_0}(\omega)h_0(\omega)g(\omega)^2}{f_{\theta_0}(\omega)^2}\,\mathrm{d}\omega.$$

□

We next extend the above proposition to the more general class of We next extend the above proposition to the more general class of focus parameters $\mu(G; h, H)$ in (1.4), being a continuously differentiable function of a finite number of the $\mu(G; h_0)$ functions. The nonparametric and parametric estimators for this class take the form

$$\widehat{\mu}_{\mathrm{np}} = H\big(n^{-1}\underline{y}_n^{\mathrm{t}}\Sigma_n(h_1)\underline{y}_n, \ldots, n^{-1}\underline{y}_n^{\mathrm{t}}\Sigma_n(h_k)\underline{y}_n\big)$$

and

$$\widehat{\mu}_{\mathrm{pm}} = H\Big(\int_{-\pi}^{\pi} h_1(\omega)f(\omega;\widehat{\theta}_n)\,\mathrm{d}\omega, \ldots, \int_{-\pi}^{\pi} h_k(\omega)f(\omega;\widehat{\theta}_n)\,\mathrm{d}\omega\Big).$$

Proposition 2. *Under the conditions of Proposition 1 the focus parameters $\mu(G; h, H)$ in (1.4), with estimators and estimands as above, fulfils*

$$\begin{pmatrix}\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}})\\ \sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_0)\end{pmatrix} \to_d \begin{pmatrix}\nabla H_{\mathrm{np}}X\\ \nabla H_{\mathrm{pm}}c^{\mathrm{t}}J(g, f_{\theta_0})^{-1}U\end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}v_{\mathrm{np}} & v_c\\ v_c & v_{\mathrm{pm}}\end{pmatrix}\right), \tag{3.3}$$

where

$$v_{\mathrm{np}} = \nabla H_{\mathrm{np}}\{4\pi\int_{-\pi}^{\pi}\{h(\omega)g(\omega)\}^2\,\mathrm{d}\omega\}\nabla H_{\mathrm{np}}^{\mathrm{t}} \quad \text{and}$$
$$v_{\mathrm{pm}} = \nabla H_{\mathrm{pm}}c^{\mathrm{t}}J(g, f_{\theta_0})^{-1}K(g, f_{\theta_0})J(g, f_{\theta_0})^{-1}c\nabla H_{\mathrm{pm}}^{\mathrm{t}},$$

and $v_c = \nabla H_{\mathrm{pm}}c^{\mathrm{t}}J(g, f_{\theta_0})^{-1}d\,\nabla H_{\mathrm{np}}^{\mathrm{t}}$, where ∇H_{np} and ∇H_{pm} are the gradients of H evaluated at respectively $(\mu(G; h_1), \ldots, \mu(G; h_k))$ and $(\mu(F_{\theta_0}; h_1), \ldots, \mu(F_{\theta_0}; h_k))$, c is the

$k \times p$-dimensional matrix with rows given by $\nabla\mu(F_{\theta_0}; h_j), j = 1, \ldots, k$ and

$$d = \mathrm{cov}(X, U) = \int_{-\pi}^{\pi} \frac{\nabla f_{\theta_0}(\omega) h(\omega) g(\omega)^2}{f_{\theta_0}(\omega)^2}\, \mathrm{d}\omega.$$

Proof. By Propostion 1, we see that (3.2) holds for each $\mu(G; h_j)$. Let now $X_{n,j} = \frac{1}{n}\underline{y}_n^{\mathrm{t}} \Sigma_n(h_j) \underline{y}_n - \mu_{\mathrm{true}}$ for $j = 1, \ldots, k$. By extending the Cramér–Wold argument in Propostion 1 to all of $X_{n,1}, \ldots, X_{n,k}, U_n$, we see that there is joint convergence for all these. The standard (multivariate) delta method then completes the proof. □

Remark 1. *From the underlying structure of the proof of Propositions 1 and 2, and the arguments (of e.g. Dahlhaus & Wefelmeyer (1996) or Dzhaparidze (1986)) used to show that the Whittle estimator has the same large-sample properties as the maximum likelihood estimator, it is clear that the conclusions of the two propositions stays true if we replace Whittle with full maximum likelihood estimation.*

The nonparametric estimator is by construction unbiased in the limit; an estimate for the risk is therefore easily obtained from the variance formula above. For the parametric candidate, we need in addition an unbiased estimate for the squared bias. Following Jullum & Hjort (2015) we start with $\widehat{b} = \widehat{\mu}_{\mathrm{pm}} - \widehat{\mu}_{\mathrm{np}}$ as an initial estimate for $b = \mu_0 - \mu_{\mathrm{true}}$. Since it follows from (3.2) that $\sqrt{n}(\widehat{b} - b) \to_d c^t J^{-1} U - X \sim \mathrm{N}(0, \kappa)$, where $\kappa = v_{\mathrm{pm}} + v_{\mathrm{np}} - 2v_c$, we have $\mathrm{E}\,\widehat{b}^2 \approx b^2 + \kappa/n + o(1/n)$. This leads to mse estimators of the form

$$\begin{aligned} \mathrm{FIC}_{\mathrm{np}} &= \widehat{\mathrm{mse}}_{\mathrm{np}} = \widehat{v}_{\mathrm{np}}/n, \\ \mathrm{FIC}_{\mathrm{pm}} &= \widehat{\mathrm{mse}}_{\mathrm{pm}} = \widehat{\mathrm{bsq}} + \widehat{v}_{\mathrm{pm}}/n = \max(0, \widehat{b}^2 - \widehat{\kappa}/n) + \widehat{v}_{\mathrm{pm}}/n. \end{aligned} \tag{3.4}$$

For the most general focus parameter formulation in (1.4), the variance and covariance estimators take the form

$$\begin{aligned} \widehat{v}_{\mathrm{np}} &= \nabla\widehat{H}_{\mathrm{np}}\{2\pi \int_{-\pi}^{\pi} h(\omega)^2 I_n(\omega)^2\, \mathrm{d}\omega\}\nabla\widehat{H}_{\mathrm{np}}^{\mathrm{t}}, \text{ and} \\ \widehat{v}_{\mathrm{pm}} &= \nabla\widehat{H}_{\mathrm{pm}}\widehat{c}^{\,\mathrm{t}} J(I_n, f_{\widehat{\theta}_n})^{-1} K(I_n/\sqrt{2}, f_{\widehat{\theta}_n}) J(I_n, f_{\widehat{\theta}_n})^{-1}\widehat{c}\, \nabla\widehat{H}_{\mathrm{pm}}, \end{aligned}$$

where $\widehat{c} = (\nabla\mu(F_{\widehat{\theta}_n}; h_k), \ldots, \nabla\mu(F_{\widehat{\theta}_n}; h_k))^{\mathrm{t}}$, $\nabla\widehat{H}_{\mathrm{np}}$ and $\nabla\widehat{H}_{\mathrm{pm}}$ are the gradients of H evaluated at respectively $(\mu(\widehat{G}_n; h_1), \ldots, \mu(\widehat{G}_n; h_k))$ and $(\mu(F_{\widehat{\theta}_n}; h_1), \ldots, \mu(F_{\widehat{\theta}_n}; h_k))$, and J and K are as defined in relation to (2.3) – using $I_n(w)^2/2$ as the canonical nonparametric unbiased estimator for $g(w)^2$. These are all consistent according to Taniguchi (1980); Deo & Chen (2000).

With FIC scores as above, representing clear-cut estimates of the risk of the nonparametric and parametric models' estimators of μ, our model selection strategy turns out as follows: Compute the FIC score for each candidate model, rank them accordingly, and select the model and estimator associated with the smallest FIC score. The same $\mathrm{FIC}_{\mathrm{pm}}$

formula (with different estimates and quantities) is used for all of the possibly m different parametric candidate models for simultaneous selection among the $m+1$ models. This is perfectly fine as the $\mathrm{FIC}_{\mathrm{pm}}$ formula does not depend on the other parametric models.

Although we have concentrated on focus functions $\mu(G; h)$ and $\mu(G; h, H)$ given by (1.3-1.4), our focused model selection strategy applies also to more general focus parameters, as long as joint limit distributions like (3.2) and (3.3) may be proven. In completely general terms, our results may be generalised to focus parameters of the form $\mu = T(G)$ for well-behaved functionals T mapping the spectral distribution G to a scalar value. The type of smoothness required for T is in fact that the functional is so-called Hadamard differentiable at G and F_{θ_0}, see e.g. van der Vaart (2000, Theorem 20.8) for further details. This allows us, for instance, to handle focus parameters involving quantiles of the spectral distribution G. It is also possible to extend the arguments to other parametric estimation procedures, especially if they are derived as minimisers of the empirical analogue of $\mathrm{argmin}\{R(G, \theta)\}$ for R the model specific part of possibly different divergence measure than in (2.2), see Dahlhaus & Wefelmeyer (1996) and Taniguchi (1980) for alternatives.

4. Models with trends

So far we have only considered stationary time series with mean zero. In real applications, this is often an unrealistic assumption to make. Even if the series is stationary, the underlying mean is rarely exactly zero; the common solution in such cases is to detrend the series. In time series modelling, detrending usually refers to the act of removing an estimated or deterministic trend from the observed series before the main analysis. This may be a complex function of time and covariates including seasonal effects, or be as simple as subtracting the arithmetic mean. A common approach is to work with the detrended series, which we will denote by $\widehat{y}_t$, and then analyse this series using models for stationary time series, without factoring in the extra estimation uncertainty involved in the detrending. This is often unproblematic, but even the innocent action of subtracting the mean may have unforeseen consequences (typically for the so-called second order properties). Hermansen & Hjort (2014b) shows that such a simple operation alter the underlying motivation and interpretation of the AIC for stationary Gaussian time series. Thus, special care is required for such an operation.

Suppose the observed series is generated by the model

$$Y_t = m(x_t, \beta) + \varepsilon_t, \tag{4.1}$$

where the x_t are p-dimensional covariates, the m is of known parametric structure, and $\{\varepsilon_t\}$ is a zero mean stationary Gaussian time series process with spectral distribution function G and corresponding density g. Assume further that we are able to estimate β by a suitable $\widehat{\beta}_n$ with reasonable precision. The question is then whether the results of

Section 3 are still valid also with detrended data, such that we may still use the same FIC formulae.

Proposition 3. *Suppose the spectral densities g and f_θ and function h satisfy the conditions of Proposition 1, and that the assumed trend m and corresponding estimator $\widehat{\beta}$ for the unknown β are such that $\sqrt{n}(\widehat{\beta}_n - \beta) = O_p(1)$. Assume further that in a neighbourhood of β we have*

$$m(x, \widehat{\beta}_n) = m(x, \beta) + \nabla m(x, \beta)^{\mathrm{t}}(\widehat{\beta}_n - \beta) + r_n(x),$$

with $\max_i |r_n(x_i)| = o_p(1/\sqrt{n})$ and $|\nabla m(x, \beta)|$ bounded in x. Then the conclusions of Proposition 1 are still true if we replace y_t with the detrended $\widehat{y}_t = y_t - m(x_t, \widehat{\beta}_n)$.

Proof. We will show that the result follows as a corollary from certain general results regarding limit behaviour of quadratic forms from Hermansen & Hjort (2014a, Section 3).

The argument is structured similarly to that of Proposition 1 and is built around a Cramér–Wold type of argument. Observe that if we replace y_t with the detrended $\widehat{y}_t = y_t - m(x_t, \widehat{\beta}_n)$, we now have $\widehat{X}_n = (\underline{\widehat{y}}_n^{\mathrm{t}} \Sigma_n(h_0) \underline{\widehat{y}}_n - \mu_{\mathrm{true}})$ and similarly

$$\widehat{U}_n = -\tfrac{1}{2}\{\mathrm{Tr}(\Sigma_n(\nabla\Psi_{\theta_0})) - \underline{\widehat{y}}_n^{\mathrm{t}} \Sigma_n(\nabla\Psi_{\theta_0}/f_{\theta_0}) \underline{\widehat{y}}_n\},$$

where $\underline{\widehat{y}}_n = (\widehat{y}_1, \ldots, \widehat{y}_n)^{\mathrm{t}}$. Again, for any $a = (a_1, a_2)$ in $\mathbb{R}^2$, we now have

$$\widehat{\Lambda}_n = a_1\sqrt{n}\widehat{X}_n + a_2\sqrt{n}\widehat{U}_n = \underline{\widehat{y}}_n^{\mathrm{t}} \Sigma_n(a_1 h_0 + a_2 \nabla\Psi_{\theta_0}/f_{\theta_0}) \underline{\widehat{y}}_n/\sqrt{n} + \gamma_n,$$

with γ_n as in the proof of Proposition 1. Then, according to Proposition 3.1 of Hermansen & Hjort (2014a),

$$\widehat{\Lambda}_n - \Lambda_n = o_p(n^{-1/2})$$

where $\Lambda_n = \underline{\varepsilon}_n^{\mathrm{t}} \Sigma_n(a_1 h_0 + a_2 \nabla\Psi_{\theta_0}/f_{\theta_0}) \underline{\varepsilon}_n/\sqrt{n} + \gamma_n$, where $\underline{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{t}}$ has elements corresponding to (4.1). Since the limit behaviour of Λ_n is what defines the limit distribution in Proposition 1, the argument is essentially complete. □

The above proposition may also be extended to the focus parameter in (1.4), as handled in Proposition 2. Traditionally, the least squares estimator has been the canonical method for estimating β in models of the form of (4.1). As an illustration, consider the linear regression model with dependent errors where $Y_t = x_t^{\mathrm{t}}\beta + \varepsilon_t$, for p-dimensional covariates x_t, and where $\{\varepsilon_t\}$ is a zero mean stationary Gaussian time series process with spectral density g. On matrix form this yields $\underline{y}_n = X\beta + \underline{\varepsilon}_n$, where X is the related $n \times p$-dimensional design matrix. The ordinary least squares estimate for β is then given by $\widehat{\beta}_n = (X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}\underline{y}_n$. Then, in order for $\widehat{\beta}_n$ to satisfy the conditions of Proposition 3, it is sufficient that $n\mathrm{Var}(\widehat{\beta}_n) = n(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}\Sigma_n(g)X(X^{\mathrm{t}}X)^{-1} = o(1)$, which is clearly satisfied if $X^{\mathrm{t}}X/n \to_p Q_1$ and $X^{\mathrm{t}}\Sigma(g)X/n \to_p Q_2$, as n approaches infinity, where Q_1 and Q_2 are both finite positive definite matrices. These are the standard assumptions needed to ensure consistency of both standard and generalised least squares for models with correlated errors.

5. Average focused information criterion

We have so far concentrated on inference for a single focus parameter μ. A natural generalisation of this is to consider several focus parameters joinly, say correlations of orders 1 to 5. The FIC machinery can easily be lifted to such a situation, involving a weighted average of FIC scores, the AFIC, with weights reflecting importance dictated by the statistician.

Suppose in general terms that estimands $\mu(u)$ are under consideration, for u in some index set. For each of these we have the nonparametric $\widehat{\mu}_{\rm np}(u)$ and one or more parametric estimators $\widehat{\mu}_{\rm pm}(u)$. These typically have versions of Propositions 1 or 2, leading as per (3.1) to

$$\mathrm{mse}_{\rm np}(u) = 0^2 + v_{\rm np}(u) \quad \text{and} \quad \mathrm{mse}_{\rm pm}(u) = b(u)^2 + v_{\rm pm}(u),$$

with $b(u) = \mu_0(u) - \mu_{\rm true}(u)$. These mean squared errors can then be combined, via some suitable cumulative weight function $W(u)$, to

$$\mathrm{risk}_{\rm np} = \int v_{\rm np}(u)\,\mathrm{d}W(u) \quad \text{and} \quad \mathrm{risk}_{\rm pm} = \int \{b(u)^2 + v_{\rm pm}(u)\}\,\mathrm{d}W(u)$$

Here $\mathrm{d}W(\cdot)$ is meant to reflect the relative importance of the different $\mu(u)$, and should stem from the statistician's judgement and the actual context. Based on the data we may now form the following natural estimates of these risk quantities:

$$\begin{aligned} \mathrm{AFIC}_{\rm np} &= \int \widehat{v}_{\rm pm}(u)\,\mathrm{d}W(u), \\ \mathrm{AFIC}_{\rm pm} &= \int \big[\max\{\widehat{b}(u)^2 - \widehat{\kappa}(u)/n\} + \widehat{v}_{\rm pm}(u)\big]\,\mathrm{d}W(u). \end{aligned} \tag{5.1}$$

This operation also needs the covariances $v_c(u)$, as $\widehat{\kappa}(u)$ is to be constructed as the natural estimator of $\kappa(u) = v_{\rm pm}(u) + v_{\rm pm}(u) - 2v_c(u)$.

The AFIC scheme (5.1) can be used in a variety of circumstances. A typical application may involve assessing models for estimating a threshold probability $P\{Y_{n+1} \geq a\}$ over a set of many a, again with a weight function $w(a)$ indicating relative importance. Another attractive application is for the task of estimating correlations $\mathrm{corr}(h)$ for lags $h = 1, 2, 3, \ldots$, perhaps with a decreasing $w(h)$. The AFIC method may similarly be applied for comparing the popular autorcorrelation function, such as `acf` in the statistical software package `R` (R Core Team, 2015), with potentially more accurate parametric alternatives.

6. Performance

In the present section we will discuss some behavioural aspects of the derived FIC methodology. First we present some theoretical consequences of using our new FIC construction for model selection. Then we discuss some issues related to the more practical performance of this criterion, and illustrate some of these in a simulation study. The goal is not to conduct a broad simulation based investigation, but rather show the potential of having a criterion for selecting among parametric models and a nonparametric alternative in a simple proof of concept type of illustration.

6.1. **FIC under model conditions.** Although we have been working outside specific parametric model conditions when deriving the FIC (and AFIC) above, it is natural to ask how the criteria selects when a parametric model is indeed correct. Consider however first the case where a specific parametric candidate model is incorrect and have bias $b \neq 0$. From the structure of the FIC formulae in (3.4) and the consistency of the involved variance and covariance estimators, we see that $\mathrm{FIC}_{\mathrm{np}} = o_p(1)$, while $\mathrm{FIC}_{\mathrm{pm}} = O_p(1) + o_p(1) = O_p(1)$. I.e. the squared bias term dominates completely, and the probability that the FIC will select this particular parametric model will tend to 0 as $n \to \infty$. If all the parametric candidate models are biased in this sense, then the FIC will eventually prefer the nonparametric model when the sample size increases.

Going more into detail, it is seen from the FIC formulae in (3.4) that the FIC prefers a specific parametric model over the nonparametric whenever

$$\max(\widehat{b}^2 - \widehat{\kappa}/n, 0) + n^{-1}\widehat{v}_{\mathrm{pm}} \leq n^{-1}\widehat{v}_{\mathrm{np}}.$$

Whenever $\widehat{v}_{\mathrm{np}} \geq \widehat{v}_{\mathrm{pm}}$, this is seen to be equivalent to

$$Z_n \leq 2,$$

where $Z_n = (n\widehat{b}^2)/(\widehat{v}_{\mathrm{np}} - \widehat{v}_c)$.

It turns out that under model conditions, we have $v_c = v_{\mathrm{pm}}$. This is rather straightforward to see by investigating the forms of v_c and v_{pm} involved in Proposition 2, in addition to the forms of K_0 and J_0. Inserting $g = f_{\theta_0}$ in these formulae reveals that $K_0 = J_0$, $\nabla H_{\mathrm{np}} = \nabla H_{\mathrm{pm}}$ and $c = d$ and thereby $v_c = v_{\mathrm{pm}}$. Now, due to the consistency, we have $\widehat{v}_{\mathrm{np}} - \widehat{v}_c \to_p v_{\mathrm{np}} - v_{\mathrm{pm}}$. Further, the limit distribution result of $\sqrt{n}(\widehat{b} - b)$ given above (3.4) ensures that $Z_n \to_d \chi_1^2$, with χ_1^2 a chi-squared distributed variable with one degree of freedom. That is, the limiting probability that the parametric model will be selected over the nonparametric when it is indeed true is $P\{Z_n \leq 2\} \to P\{\chi_1^2 \leq 2\} \approx 0.843$. Thus, if one of the parametric candidate models is correct, and the others have biases $b \neq 0$, then, for sufficiently large samples, the first parametric model and estimator will be selected with a probability tending to 84.3%, while the nonparametric will be selected in the other 15.7% proportion.

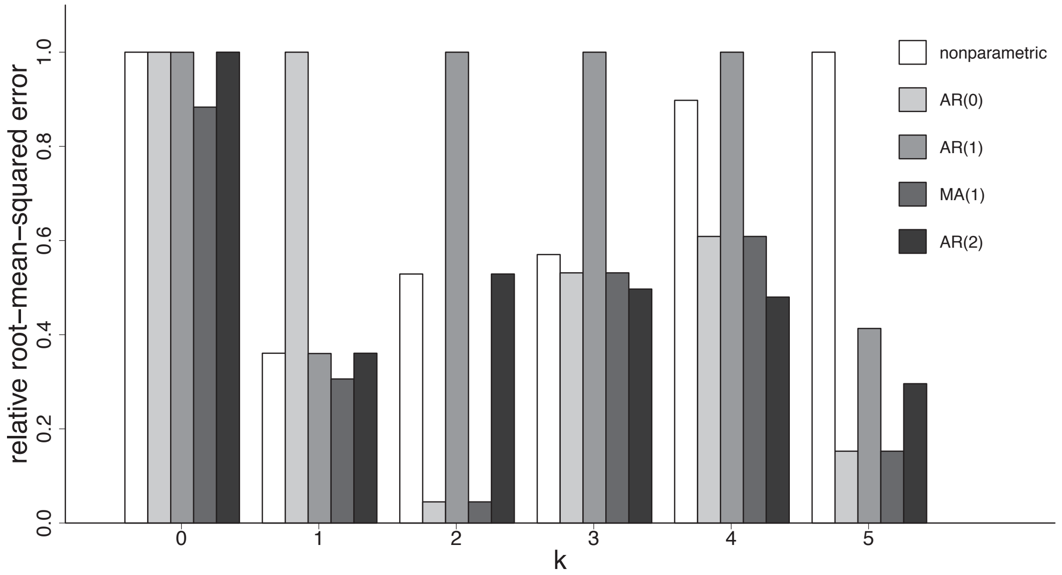

Figure 6.1. Relative root-mse for each candidate model fitted to the six focus parameters $\mu_k = C(k)$, for $k = 0, \ldots, 5$. The root-mse is computed based on 5000 simulated AR(2) series of length $n = 100$, with $\sigma = 1.0$ and $\rho = (0.7, -0.6)$, For ease of comparison we have scaled the root-mse to the unit interval.

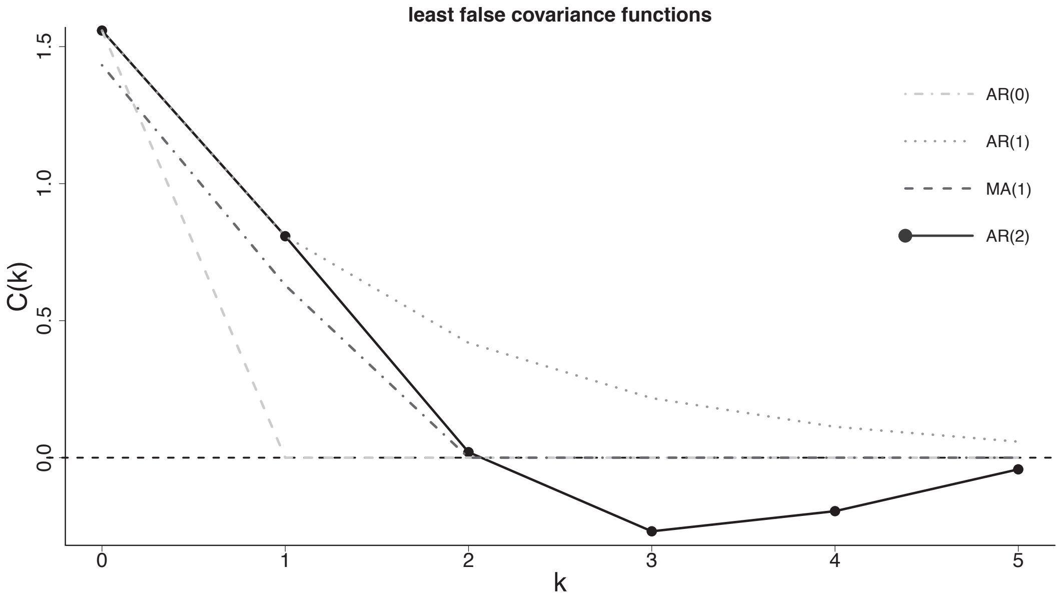

Figure 6.2. The five least false covariance functions under the assumption that the true model is an autoregressive model specified by the parameters $\sigma = 1.0$ and $\rho = (0.7, -0.6)$.

6.2. **FIC in practice.** Figure 6.1 shows the relative root-mse for estimating the focus parameter

$$\mu_k = \mu(G; h_k) = \int_{-\pi}^{\pi} \cos(\omega k) g(\omega)\, \mathrm{d}\omega = C_g(k), \qquad \text{for } k = 1, \ldots, 5, \tag{6.1}$$

based on the following five candidates models: the independence model (autoregressive of order zero); the autoregressive of orders one and two; the moving average of order one; and finally the nonparametric one, where nothing more is assumed than saying that the series is stationary with a finite variance. The true model is an autoregressive model specified by the parameters $\rho = (0.7, -0.6)$ and $\sigma = 1.0$. This means that all but two, the autoregressive model of order two and the nonparametric model, are misspecified. The corresponding least false covariance estimates are plotted in Figure 6.2. In the simulation study, we have used $B = 5000$ repetitions of length $n = 100$ to compute the actual relative root-mse values for each candidate. Note that since we have included the true model among our candidates, nonparametric estimation is never the optimal choice; it is however often close and it is the second best choice for lags 1 and 3. For lags 2 and 5, where the true values are close to zero, the simpler models, like AR(0) and MA(1), are highly successful, achieving reasonably low bias and also low variance.

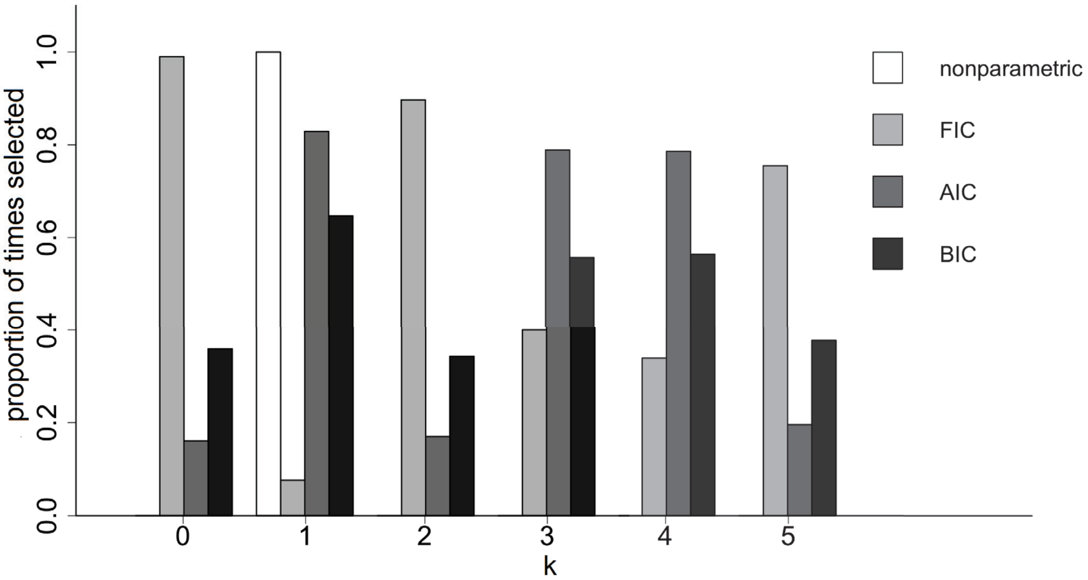

Figure 6.3. The proportion for which the different criteria selects the model with the theoretical lowest root-mean-squared error. The model-selectors are always nonparametric, FIC, AIC and BIC. The results are based on 5000 simulated series.

In Figure 6.3 and 6.4 we further investigate the performance of the FIC. Here, we compare our FIC machinery with three other model selection strategies, (i) to always use the nonparametric model, (ii) select the best parametric model according to the AIC and (iii) the parametric model selected by the BIC. Note that the AIC and BIC tools do not work for the nonparametric model, since there is no likelihood function. In Figure 6.3 we have counted how many times each criterion selects the model that obtains the smallest root-mse value, for each focus parameter μ_k as defined in (6.1). Figure 6.4 contains the corresponding attained root-mse values. Note that for lag 1 the theoretical root-mse for the autoregressive models are, for all practical purposes, equal to that obtained by the nonparametric model. In all other cases, the nonparametric model has a root-mse larger than the optimal model.

In this illustration, the FIC behaves more or less as intended, by selecting (on average) the models that produces the smallest risk. The amount of evidence is by no means conclusive, but it indicates that the FIC machinery has a real potential.

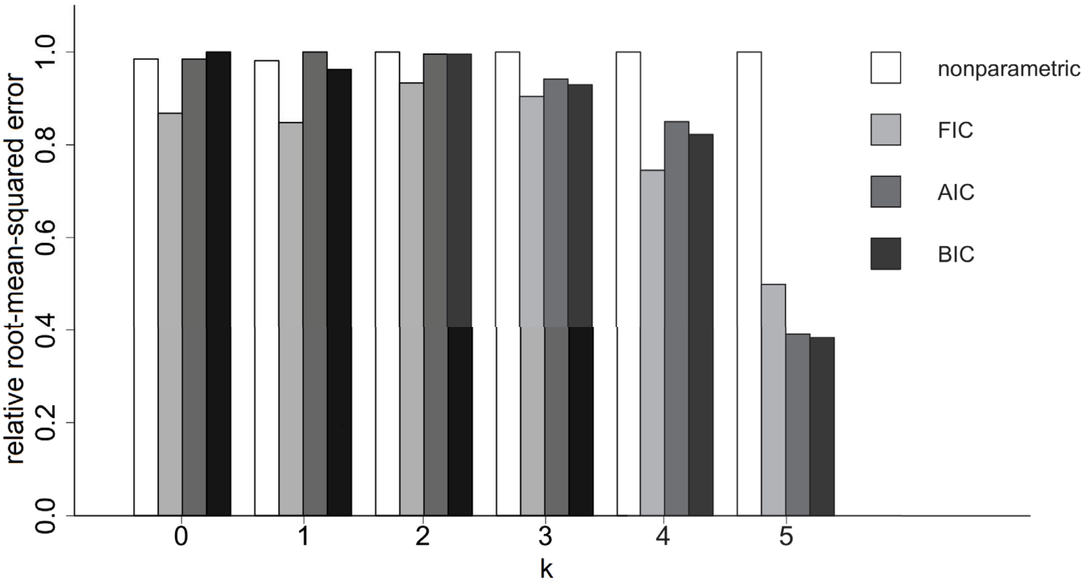

Figure 6.4. The relative root-mean-squared (computed in the same simulations) for the models selected by FIC, AIC and BIC, and by always using the nonparametric model.

7. Concluding remarks

Here we offer a list of conclucing comments, some pointing to further relevant research.

7.1. **Model averaging.** The FIC scores may also be used to combine the most promising estimators into a model averaged estimator, say $\widehat{\mu}^* = \sum_j c(M_j)\widehat{\mu}_j$, with $c(M_j)$ given higher values for models M_j with higher FIC scores; as discussed in Hjort & Claeskens (2003).

7.2. **The conditional FIC.** For time series processes, several interesting and important focus parameters are naturally related to predictions, are sample size dependent or otherwise formulated conditional on past observations. The classical example is k-step ahead predictions. A class of such estimands could take the form

$$\mu(\alpha, \gamma, y_1, \ldots, y_m) = P\{Y_{n+1} > \alpha \text{ and } Y_{n+2} > \gamma \,|\, y_1, \ldots, y_m\}$$

for a suitable choice of α. The dependency on previous data requires a new and extended modelling framework, which in Hermansen & Hjort (2015, Sections 5 & 6) led to generalisations and also motivated a conditional focused information criterion (cFIC). In completing the FIC-framework for selecting among parametric and nonparametric time series models, such considerations should also be taken into account.

7.3. **Linear time series processes.** Building on Walker (1964); Hannan (1973); Brillinger (1975), the main results of Section 3 can be extended to more general types of time series processes, like the generalised linear processes (cf. Priestley (1981)); also without the assumption of Gaussian innovation terms.

7.4. **Trends and covariates.** In the presented work, our focus was on the dependency structure only. However, the methods and results of our paper may be generalised to select simultaneously among models with different trends and dependency structures, like $Y_t = m(x_t, \beta) + \varepsilon_t$, with ε_t a stationary Gaussian time process. These issues, leading to a larger repertoire of FIC formulae, will be returned to in later work. Since it is generally hard to estimate both the trend and dependency structure using a full nonparametric framework, the two main challenges is to extend the existing work to handle the case with various parametric candidates for the trend $m(x_t, \beta)$ and both parametric models and a nonparametric candidate for the dependency, i.e. the spectral distribution (since we are working under the Gaussian assumption). Alternatively, we may assume that the ε_t belongs to an appropriate width family of parametric stationary time series processes, such as the autoregressive AR, the moving average MA or the mixture ARMA (cf. Brockwell & Davis (1991)) and instead compare a nonparametric method for estimating the trend part of the model, perhaps extending this to functions of the type $m(t, x_i, \beta)$, against a class of parametric alternatives.

7.5. **The local large-sample framework.** As mentioned in the introduction, Hermansen & Hjort (2015) derives FIC for selecting among parametric time series models using a local asymptotics framework. The parametric candidate models then have spectral densities belonging to a parametric family $f(\cdot; \theta, \gamma)$, with a p-dimensional protected θ and a q-dimensional open γ. This constitutes a set of 2^q potential parametric candidate models. The full (or wide) model is represented by the spectral density $f(\cdot; \theta, \gamma)$. At the other end of the spectrum, the narrow model corresponds to fixating $\gamma = \gamma_0$, a known value, with the resulting $f(\cdot;\theta) = f(\cdot;\theta,\gamma_0)$. The local misspecification framework then assumes that the true spectral density takes the form $f(\cdot;\theta_0,\gamma_0+\delta/\sqrt{n})$, for some unknown q-dimensional δ describing the distance to the wide model. This framework causes variances and squared biases to become of the same order of magnitude $O(1/n)$. Those lead to approximation formulae for the mean squared error and FIC formulae for nested parametric models, which are different from those obtained in this paper.

The introduction of the 'asymptotically correct' nonparametric model of the present paper allowed us to derive FIC formulae even when sidestepping the above local misspecification assumption. An alternative approach is to retain the local asymptotics framework and work with spectral densities of the type $f_r(\omega) = f_{\theta_0}(\omega) + r(\omega)/\sqrt{n}$, where f_{θ_0} is a standard type of parametric model. Such structures have already been worked with in Dzhaparidze (1986), making the extension potentially less cumbersome. This will not be dealt with here, however.

References

Brillinger, D. (1975). *Time Series: Data Analysis and Theory.* Holt, Rinehart and Winston.

Brockwell, P. & Davis, R. (1991). *Time Series: Theory and Methods.* Springer.

Claeskens, G. & Hjort, N. L. (2003). The focused information criterion [with discussion and a rejoinder]. *Journal of the American Statistical Association* **98**, 900–916.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

Coursol, J. & Dacunha-Castelle, D. (1982). Remarks on the approximation of the likelihood function of a stationary Gaussian process. *Theory of Probability and its Applications* **27**, 162–167.

Dahlhaus, R. & Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Annals of Statistics* **24**, 952–973.

Deo, R. S. & Chen, W. W. (2000). On the integral of the squared periodogram. *Stochastic processes and their applications* **85**, 159–176.

Dzhaparidze, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.* Berlin: Springer.

Gray, R. (2006). *Toeplitz and Circulant Matrices: A Review.* Now publishers Inc.

Hannan, E. J. (1973). The asymptotic theory of linear time-series models. *Journal of Applied Probability* **10**, 130–145.

Hermansen, G. & Hjort, N. L. (2014a). Limiting normality of quadratic forms with applications to time series analysis. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hermansen, G. & Hjort, N. L. (2014b). A new approach to Akaike's information criterion and model selection issues in stationary Gaussian time series. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hermansen, G. & Hjort, N. L. (2015). Focused information criteria for time series. *Submitted for publication* .

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators [with discussion and rejoinder]. *Journal of the American Statistical Association* **98**, 879–899.

Jullum, M. & Hjort, N. L. (2015). Parametric or nonparametric: The FIC approach. *Submitted for publication* .

Priestley, M. (1981). *Spectral Analysis and Time Series.* Academic Press.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Taniguchi, M. (1980). On estimation of the integrals of certain functions of spectral density. *Journal of Applied Probability* **17**, 73–80.

van der Vaart, A. (2000). *Asymptotic Statistics.* Cambridge University Press.

Walker, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *Journal of the Australian Mathematical Society* **4**, 363–384.

Whittle, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society Series B* **15**, 125–139.

GEOPHYSICS, VOL. 81, NO. 3 (MAY-JUNE 2016); P. 1–13, 11 FIGS., 2 TABLES.
10.1190/GEO2015-0314.1

A Gaussian-based framework for local Bayesian inversion of geophysical data to rock properties

Martin Jullum[1] and Odd Kolbjørnsen[2]

ABSTRACT

Working in a Bayesian framework, we have derived a procedure for inverting rock properties based on geophysical data. The purpose was to arrive at a widely applicable and general procedure in which few and weak assumptions are required for application to various inverse problems within the geophysical industry. Our Bayesian statistical approach combines sampling-based techniques and Gaussian approximations to assess local approximations to quantities related to the posterior distribution of rock properties. These approximated quantities define the Bayesian inversion. A conceptual advantage of our approach is that there are few restrictions on the initial model, allowing realistic statistical models to be approximated directly. The methodology is easily parallelized and offers a range of procedures, which gives a trade-off between inversion speed and accuracy. We have tested the approach in a monitoring setting using seismic amplitudes by evaluating a synthetic case and real data from the Sleipner CO_2 injection project. For the synthetic case, the inversion results correspond well with the rock properties used to generate the data and the posterior distribution derived using an MCMC approach. We also found improved accuracy compared with a frequently used Gaussian inversion approach. In the real data case, we clearly identified high-saturation layers present in previous qualitative interpretations.

INTRODUCTION

In the geophysical industry, there is a great need for solutions to various types of inverse problems. Most of these problems are however ill posed and seldom have a unique or well-defined solution. The final objective of many of the inverse problems is to predict rock properties such as porosity, lithology, saturation, permeability etc. from geophysical data such as seismic amplitudes (Mukerji et al., 2001; Doyen, 2007; Gunning and Glinsky, 2007; Avseth et al., 2010). The Bayesian approach (Tarantola and Valette, 1982) is a popular framework for solving inverse problems (Bosch et al., 2010). The main advantage of the Bayesian approach is the possibility to incorporate additional knowledge of the problem and assess the uncertainty after accounting for the data. In such a setting, computation or approximation of a posterior distribution, here corresponding to the probability distribution of the rock properties conditioned on the observed geophysical data, determines the inversion.

Unfortunately, full analytical evaluation of the posterior distribution is only possible for highly restricted classes of distributions. For instance, Buland and Omre (2003) perform inversion from seismic amplitude versus offset (AVO) data to elastic parameters (but not further to actual rock properties) by assuming that the two model components are jointly Gaussian, resulting in an analytic closed-form Gaussian posterior distribution for the elastic parameters. Because realistic models seldom fit such a formulation, the methodology may merely be viewed as an approximation, possibly far from the actual posterior distribution (Rimstad and Omre, 2014a). Rimstad and Omre (2014a, 2014b) relax the Gaussian model assumptions, but to evaluate the posterior distribution, they need to use Markov Chain Monte Carlo (MCMC; Robert and Casella, 2005) sampling procedures, which may be very time consuming in high dimensions. The Gaussian mixture approach of Grana and Della Rossa (2010) also relaxes the Gaussian assumption, and even goes all the way to rock properties, but it requires modeling approximations elsewhere and thereby also operates approximately.

Several authors have restricted their attention to discrete facies as the rock property of interest. Larsen et al. (2006) introduce Markov property dependencies to describe a vertical profile based on seismic AVO data. Buland et al. (2008), Ulvmoen and Omre (2010), and others follow along similar lines. These approaches include an initial step in which a distribution claimed to be multimodal is approximated by a unimodal Gaussian distribution based on Buland

Manuscript received by the Editor 13 June 2015; revised manuscript received 14 October 2015.
[1]University of Oslo, Department of Mathematics, Oslo, Norway. E-mail: martinju@math.uio.no.
[2]Formerly Norwegian Computing Center, Oslo, Norway; presently Lundin Norway, Oslo, Norway. E-mail: odd.kolbjornsen@lundin-norway.no.

and Omre (2003). Even if the multimodality is subsequently corrected for, and the results appear reasonable, the conflicting mode assumptions have unclear implications. Further, even if the used discrete Markov property is suitable in some situations, it generally restricts the dependence structure. Finally, because these methodologies rely on the discrete nature of the facies, the approaches cannot be directly transferred to situations with continuously distributed rock properties.

Although there is a wide range of techniques in the statistical literature for approximating Bayesian posterior distributions (for a review, see e.g., Green et al., 2015), there are few examples of such techniques being directly applied to geophysical types of inverse problems. This is possibly caused by a gap between the computational efficiency of such techniques when being applied to large geophysical types of inverse problems, and what is acceptable for the industry (Mosegaard and Tarantola, 2002). Even though they are time consuming, attempts have been made to solve such inverse problems based on the MCMC approach (Mosegaard and Tarantola, 1995; Malinverno, 2002; Hammer et al., 2012).

It is evident that there is a need for a computationally feasible large-scale inversion methodology which can deal with more general model formulations than the existing ones. Short of computational advances, attempts to make statistical methods computationally tractable involve some kind of simplification, approximation, or intelligent elimination of redundancy. For these types of inverse problems, it often suffices to obtain inversion results for each individual cell in a grid of the region of interest. In a Bayesian framework, such local inversion corresponds to computing or approximating the marginal posterior distribution in each cell and using predictors and uncertainty measures based on those distributions as the inversion result.

Our methodology uses a local inversion approach that simplifies the problem. Inversion of a larger region then reduces to a large number of local inversions, which we handle individually. The efficiency of the local inversion approach lies partly in including only the variables and data most relevant for the current local inversion. This reduces the dimensionality of the problem to a magnitude that we can handle efficiently, while still taking the most relevant spatial dependence into account. The local inversion is carried out by relying on a certain Gaussian likelihood approximation and a weighted Monte Carlo sampling routine. Handling the local inversions individually allows us to parallelize the full-inversion problem, leading to heavy algorithmic speed-up compared with MCMC-type procedures. There are also few modeling limitations underlying our approach. In particular, our approach can handle any type of rock-physics model and any type of prior distribution for the rock properties. It is not limited by specific restrictions on the spatial structure typically present in other approaches (such as e.g., certain Markov-type dependence structures), and it can in principle be used with any rock property: continuous, discrete, or a combination of the two.

METHODOLOGY

The full inverse problem is usually decomposed into two inversions: geophysical inversion (Buland and Omre, 2003) and rock-physics inversion (Avseth et al., 2010). The joint global problem may be described by the simple hierarchical formulation shown in Figure 1. Here, $\mathbf{r}$ denotes a rock property, such as lithology, porosity, or saturation that we are interested in. The geophysical properties of $\mathbf{m}$ typically consist of density and two elastic parameters (either P- and S-wave velocities, or acoustic and shear impedance) or any triplets of these. Finally, $\mathbf{d}$ denotes the geophysical data often consisting of seismic AVO data at a few different offsets. The geophysical data may however also represent data sources such as root-mean-square (rms) velocity (Buland et al., 2011) or gravimetrics data (Hauge and Kolbjørnsen, 2015). Hence, the left arrow of Figure 1 represents the rock-physical relation, whereas the right arrow represents the geophysical relation. That is, all impact from the rock properties to the geophysical data goes through the geophysical properties.

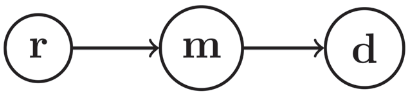

Figure 1. Forward model hierarchy.

We shall need a fair amount of notation when building our inversion framework. We shall assume that the global quantities in Figure 1 all operate on the same grid of the region of interest. (They may in principle work on different grids of the region, but we exclude that case for presentational simplicity.) Uppercase letters in calligraphic font ($\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$) are used to denote subsets of cells of the gridded region. A variable written in boldface roman font with a calligraphic subscript (e.g., $\mathbf{d}_{\mathcal{D}}$) refers to the subvector corresponding to the subset of that subscript. A boldface roman variable with no calligraphic subscript (e.g., $\mathbf{d}$) contains the individual variable(s) of the complete gridded region under consideration. For other quantities, we will use fairly standard statistical notation: We use a superscript roman T for the matrix transpose, $p(\cdot)$ as a generic notation for probability distributions, $\sim$ for "distributed as," $\boldsymbol{\mu}$ for the mean (vector), and $\boldsymbol{\Sigma}$ for the covariance matrix. The Gaussian distribution of a variable $\mathbf{x}$ (with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$) is denoted by $\mathrm{N}_{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Noncalligraphic subscripts will be used to distinguish variables of similar types. A superscript asterisk will be specifically used to denote approximate quantities used in the framework, like e.g., $p^*(\mathbf{x})$ and $\boldsymbol{\mu}^*_{\mathbf{x}}$. The most important quantities are also given in Table 1.

The general framework

The overall goal of the inversion is to predict the rock properties $\mathbf{r}$ from the obtained geophysical data $\mathbf{d}$ over a gridded interest region $\mathbb{A}$. We will handle this by focusing on one grid cell at a time, and hence carry out predictions and uncertainty measures on grid cell level. Because the methodology will be the same for each cell in $\mathbb{A}$, we will present the methodology by considering a single cell $\mathcal{A}$ — the extension to $\mathbb{A}$ amounts to repeating the procedure for each $\mathcal{A} \in \mathbb{A}$. Working in the Bayesian framework, carrying out the inversion to rock properties in cell $\mathcal{A}$ amounts to evaluating approximated quantities or measures related to the marginal posterior distribution $p(\mathbf{r}_{\mathcal{A}}|\mathbf{d})$ of the target variable $\mathbf{r}_{\mathcal{A}}$. Any preferred measure of central tendency of this posterior distribution may be used as a predictor for the true rock property in cell $\mathcal{A}$. Common selections are the mean, mode, and median. The uncertainty may be quantified by a measure of spread, such as the standard deviation, or through one or more suitably chosen credibility intervals. Local probability statements, such as the probability that the porosity in $\mathcal{A}$ is more than 0.15, may also be computed. Our method may also produce an approximation to the complete marginal posterior distribution, should that be of interest.

By using Bayes' formula, the posterior distribution of the target variable is given by

$$p(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{r}_{\mathcal{A}})p(\mathbf{r}_{\mathcal{A}}). \tag{1}$$

Here, $p(\mathbf{r}_{\mathcal{A}})$ is the prior probability distribution for the target variable, whereas $p(\mathbf{d}|\mathbf{r}_{\mathcal{A}})$ is the likelihood of the geophysical data, conditioned only on the target variable. Hence, in $p(\mathbf{d}|\mathbf{r}_{\mathcal{A}})$, the geophysical properties $\mathbf{m}$ are marginalized out along with the rock properties in other cells than $\mathcal{A}$. In terms of Figure 1, this model setup corresponds to a direct arrow from $\mathbf{r}_{\mathcal{A}}$ to $\mathbf{d}$, reducing the global two stage problem to a local one stage problem. This is beneficial because two-step approaches do not fully account for the dependence in Bayesian models (Bosch et al., 2010). It is not generally possible to give a closed-form expression for $p(\mathbf{d}|\mathbf{r}_{\mathcal{A}})$ based solely on the global geophysical and rock-physical likelihoods $p(\mathbf{d}|\mathbf{m})$ and $p(\mathbf{m}|\mathbf{r})$. Essentially, one would have to go through the following rewrite of equation 1:

$$\begin{aligned} p(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) &\propto \iint p(\mathbf{d}|\mathbf{m})p(\mathbf{m}|\mathbf{r})p(\mathbf{r}_{\mathbb{A}\setminus\mathcal{A}}|r_{\mathcal{A}})\,\mathrm{d}\mathbf{m}\,\mathrm{d}\mathbf{r}_{\mathbb{A}\setminus\mathcal{A}}p(\mathbf{r}_{\mathcal{A}}) \\ &= \iint p(\mathbf{d}|\mathbf{m})p(\mathbf{m}|\mathbf{r})p(\mathbf{r})\,\mathrm{d}\mathbf{m}\,\mathrm{d}\mathbf{r}_{\mathbb{A}\setminus\mathcal{A}}, \end{aligned} \tag{2}$$

where $\mathbb{A}\setminus\mathcal{A}$ denotes all cells in $\mathbb{A}$ except $\mathcal{A}$. The dimensions of these integrals depend on the number of cells in $\mathbb{A}$ and whether the geophysical and rock-physical models possess independence between certain cells. For realistic problems, these are usually at least 100-dimensional, and in the densest cases, they might be of dimension 10^6 or more. Thus, we cannot tackle this problem directly via equations 1 and 2.

Instead of attempting to work with the global geophysical and rock-physical likelihoods, our approach aims at modeling only the part that is most relevant for the target variable $\mathbf{r}_{\mathcal{A}}$. Let us thus introduce the local subsets $\mathcal{B}, \mathcal{C}$, and $\mathcal{D}$, which are sets of region cells reflecting the modeled part of respectively the rock properties $\mathbf{r}$, the geophysical properties $\mathbf{m}$, and the geophysical data $\mathbf{d}$ in the local inversion for cell $\mathcal{A}$. Their corresponding variable sets $\mathbf{r}_{\mathcal{B}}$, $\mathbf{m}_{\mathcal{C}}$, and $\mathbf{d}_{\mathcal{D}}$ are named, respectively, the neighborhood variable, influence variable, and local data. To set the idea of the local subsets straight, the bulleted list below and illustration in Figure 2 provide basic guidelines for how these may be specified in an AVO data setting with vertical dependence:

- All local subset variables should be centered in cell $\mathcal{A}$.
- $\mathcal{D}$ should include the cells for which the data $\mathbf{d}$ are influenced by $\mathbf{m}_{\mathcal{A}}$, that is, half a wavelet length above and below $\mathcal{A}$.
- $\mathcal{C}$ should include the cells for which the geophysical properties $\mathbf{m}$ influence the local data $\mathbf{d}_{\mathcal{D}}$, i.e., one wavelet length above and below $\mathcal{A}$.
- $\mathcal{B}$ should have size at least in order of the tuning thickness. This ensures that $\mathbf{r}_{\mathcal{B}}$ is the main source of variability for the data interfering with the contribution from $\mathcal{A}$.

Note, however, that our approach is not restricted to vertically defined local subsets; that is, lateral dependence may in principle also be modeled by our approach. The task of selecting the local subsets will be discussed in more depth later on.

Referring to Bayes' formula as in equation 1, our approximation approach is based on the relation

$$p(\mathbf{r}_{\mathcal{A}}|\mathbf{d}_{\mathcal{D}}) = \int p(\mathbf{r}_{\mathcal{B}}|\mathbf{d}_{\mathcal{D}})\,\mathrm{d}\mathbf{r}_{\mathcal{B}\setminus\mathcal{A}} \propto \int p(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}})p(\mathbf{r}_{\mathcal{B}})\,\mathrm{d}\mathbf{r}_{\mathcal{B}\setminus\mathcal{A}}, \tag{3}$$

where $\mathcal{B}\setminus\mathcal{A}$ denotes all cells in $\mathcal{B}$ except $\mathcal{A}$. Replacing $p(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}})$ (henceforth referred to as the local likelihood) by a Gaussian approximation yields the following integral form approximation:

$$p(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) \approx p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}_{\mathcal{D}}) \propto \int p^*(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}})p(\mathbf{r}_{\mathcal{B}})\,\mathrm{d}\mathbf{r}_{\mathcal{B}\setminus\mathcal{A}}, \tag{4}$$

where the Gaussian approximation is given by

$$p^*(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}}) = \mathrm{N}_{\mathbf{d}_{\mathcal{D}}}\left(\boldsymbol{\mu}^*_{\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}}}(\mathbf{r}_{\mathcal{B}}), \boldsymbol{\Sigma}^*_{\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}}}(\mathbf{r}_{\mathcal{B}})\right). \tag{5}$$

Table 1. Important quantities.

Symbol	Description
$\mathbf{d}$	Vector of geophysical data for all cells in the gridded interest region
$\mathbf{m}$	Vector of geophysical properties for all cells in the gridded interest region
$\mathbf{r}$	Vector of the rock property of interest for all cells in the gridded interest region
$\mathbb{A}$	Gridded interest region
$\mathcal{A}$	Cell in $\mathbb{A}$ under consideration
$\mathcal{B}$	Subset of region cells for which the rock properties are modeled in the local inversion for $\mathbf{r}_{\mathcal{A}}$
$\mathcal{C}$	Subset of region cells for which the geophysical properties are modeled in the local inversion for $\mathbf{r}_{\mathcal{A}}$
$\mathcal{D}$	Subset of region cells for which the geophysical data are included in the local inversion for $\mathbf{r}_{\mathcal{A}}$
$\mathbf{r}_{\mathcal{A}}$	Target variable
$\mathbf{r}_{\mathcal{B}}$	Neighborhood variable
$\mathbf{m}_{\mathcal{C}}$	Influence variable
$\mathbf{d}_{\mathcal{D}}$	Local data
$\mathbf{G}$	Matrix representing the impact $\mathbf{m}$ has on the mean of $\mathbf{d}$
$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$	Covariance matrix of the geophysical likelihood

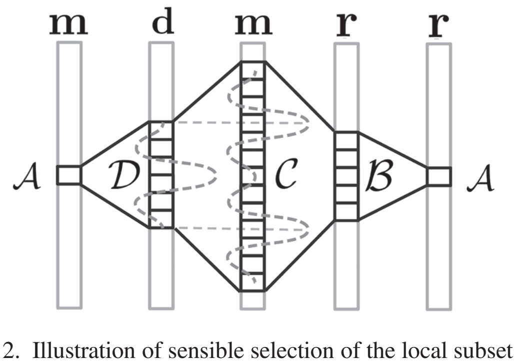

Figure 2. Illustration of sensible selection of the local subsets $\mathcal{B}, \mathcal{C}$, and $\mathcal{D}$ for seismic AVO data with vertical dependence.

As will become clear in the following subsections, the Gaussian distribution in equation 5 will be established by merging a Gaussian approximation to the local rock-physical likelihood $p(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$ with a Gaussian approximation to the local geophysical likelihood $p(\mathbf{d}_\mathcal{D}|\mathbf{m}_\mathcal{C})$. Based on the integral form in equation 4, we define a weighted Monte Carlo routine enabling us to approximate in principle any quantities of interest related to $p(\mathbf{r}_\mathcal{A}|\mathbf{d})$ (such as the mean, variance, probability of a certain event, or the complete density) by properly aggregating weighted samples from the prior of the target variable. The next three subsections present the details of the approximations for the local rock-physical and geophysical likelihoods, in addition to the proposed weighted Monte Carlo routine.

Local rock-physical likelihood

There is in general no simple description of the spatial distribution of the rock-physical likelihood $p(\mathbf{m}|\mathbf{r})$. To obtain a Gaussian approximation to the required local rock-physical likelihood $p(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$, we therefore take a flexible sampling-based approach. This approach only requires that we are able to sample "pairs" $(\mathbf{m}_\mathcal{C}, \mathbf{r}_\mathcal{B})$ from their joint distribution. The objective is to use these samples to fit the best possible mean and covariance matrix functions in a Gaussian approximation of the form

$$p^*(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B}) = \mathrm{N}_{\mathbf{m}_\mathcal{C}}(\boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B}}(\mathbf{r}_\mathcal{B}), \boldsymbol{\Sigma}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B}}(\mathbf{r}_\mathcal{B})). \tag{6}$$

Note that a Gaussian approximation to $p(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$ is less restrictive than the Gaussian approximation to the unconditional distribution of the geophysical properties $p(\mathbf{m})$ used in Buland and Omre (2003), Larsen et al. (2006), and related work. In principle, the mean function $\boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B}}(\mathbf{r}_\mathcal{B})$ and covariance function $\boldsymbol{\Sigma}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B}}(\mathbf{r}_\mathcal{B})$ could behave in completely unrestricted ways. In order not to make the fitting procedure too complicated, we do, however, suggest to divide the sampled pairs into K different nonoverlapping classes according to some specified criterion on $\mathbf{r}_\mathcal{B}$. Within each such class k, separate mean functions $\boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}(\mathbf{r}_\mathcal{B})$ are fitted by a regression procedure, and the resulting residuals $\boldsymbol{\varepsilon}' = \mathbf{m}_\mathcal{C} - \boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}(\mathbf{r}_\mathcal{B})$ are used to estimate a fixed covariance matrix $\boldsymbol{\Sigma}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}$ for that class. Using this procedure, the mean function has an unrestricted dependence on the neighborhood variable $\mathbf{r}_\mathcal{B}$, whereas the covariance matrix depends categorically on the class k of $\mathbf{r}_\mathcal{B}$. Hence, the criterion that divides the samples into different classes should be chosen such that the dependence structure within the influence variable $\mathbf{m}_\mathcal{C}$ is fairly stable.

Although a simple linear regression (least squares) method may be used to fit $\boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}(\mathbf{r}_\mathcal{B})$ in the above procedure, we suggest using a technique that allows for increased fidelity in more complex situations. Examples are multivariate adaptive regression splines (MARS; Friedman, 1991), projection pursuit regression (Friedman and Stuetzle, 1981), neural networks (Cheng and Titterington, 1994), and generalized additive models (Hastie and Tibshirani, 1986). Flexibility is essential here because it allows the approximated dependence on $\mathbf{r}_\mathcal{B}$ to match that of the true model more closely. Also, the larger the sample size the more stable the approximations become.

One strength of the sampling-based approach is that the distribution fit may be checked by standard multivariate normality tests (Henze, 2002). If tests deem the Gaussian model acceptable, the approximations are guaranteed to be good. If not, one may attempt to correct for the non-Gaussianity by using a more flexible regression procedure, increase the number of classes, redefine the local subsets, or reduce the influence of the outliers. The last option may for instance be done by using a range spanning covariance estimation routine to estimate $\boldsymbol{\Sigma}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}$ as opposed to using the standard sample covariance. Such a routine stretches the tails of the Gaussian model by using an estimate of the covariance that spans broadly enough for no sampled pairs $(\mathbf{m}_\mathcal{C}, \mathbf{r}_\mathcal{B})$ to be very unlikely under the fitted model. This method weakens the impact of deviations from the Gaussian model's dependence structure, and it reduces to the standard sample covariance if the fit is already good. The suggested routine is outlined in Appendix A.

Local geophysical likelihood

The global geophysical likelihood model typically takes the form $p(\mathbf{d}|\mathbf{m}) = \mathrm{N}_\mathbf{d}(\mathbf{Gm}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, where $\mathbf{G}$ is a matrix of the appropriate dimension representing the linear dependence of the geophysical data $\mathbf{d}$ on the geophysical properties $\mathbf{m}$. That is

$$\mathbf{d} = \mathbf{Gm} + \boldsymbol{\varepsilon}, \tag{7}$$

with $\boldsymbol{\varepsilon}$ some error term with distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. We seek an approximation for the local geophysical likelihood $p(\mathbf{d}_\mathcal{D}|\mathbf{m}_\mathcal{C})$, where

$$\mathbf{d}_\mathcal{D} \approx \tilde{\mathbf{G}}\mathbf{m}_\mathcal{C} + \tilde{\boldsymbol{\varepsilon}}, \tag{8}$$

with $\tilde{\mathbf{G}}$ corresponding to $\mathbf{G}$ above, and $\tilde{\boldsymbol{\varepsilon}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}})$. However, extracting the local part of equation 7 gives $\mathbf{d}_\mathcal{D} = \mathbf{G}_\mathcal{D}\mathbf{m} + \boldsymbol{\varepsilon}_\mathcal{D}$, where $\mathbf{G}_\mathcal{D}$ is the submatrix of $\mathbf{G}$ containing only the rows corresponding to $\mathcal{D}$. As illustrated upon introduction of the local subsets, $\mathcal{C}$ is chosen as the region whose geophysical properties influence the local data (the most). Hence, it is reasonable to approximate $\mathbf{G}_\mathcal{D}\mathbf{m}$ by $\mathbf{G}_{\mathcal{D},\mathcal{C}}\mathbf{m}_\mathcal{C}$, where $\mathbf{G}_{\mathcal{D},\mathcal{C}}$ contains the columns of $\mathbf{G}_\mathcal{D}$ corresponding to the cells in $\mathcal{C}$. The sought-after approximation is consequently obtained by letting $\tilde{\mathbf{G}} = \mathbf{G}_{\mathcal{D},\mathcal{C}}$ and $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_\mathcal{D}}$ in relation 8.

By a result in Appendix B, the established local rock-physical and geophysical likelihood approximations give a fully specified local likelihood approximation $p^*(\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B})$ as in equation 5 with

$$\begin{aligned}\boldsymbol{\mu}^*_{\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B}}(\mathbf{r}_\mathcal{B}) &= \mathbf{G}_{\mathcal{D},\mathcal{C}}\boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}(\mathbf{r}_\mathcal{B}),\\ \boldsymbol{\Sigma}^*_{\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B}}(\mathbf{r}_\mathcal{B}) &= \mathbf{G}_{\mathcal{D},\mathcal{C}}\boldsymbol{\Sigma}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}\mathbf{G}^\mathrm{T}_{\mathcal{D},\mathcal{C}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_\mathcal{D}},\end{aligned} \tag{9}$$

for each $\mathbf{r}_\mathcal{B}$ in class k.

Weighted Monte Carlo routine

The final part of the framework concerns the weighted Monte Carlo routine, which approximates inversion quantities of interest. The routine relies on the integral in relation 4 with $p^*(\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B})$ as specified by equations 5 and 9. The routine goes as follows:

1) Sample a large number L of $\mathbf{r}_\mathcal{B}$-variables from its prior $p(\mathbf{r}_\mathcal{B})$.
2) For each sample $\mathbf{r}^{(l)}_\mathcal{B}$, compute the mean $\boldsymbol{\mu}^*_{\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B}}(\mathbf{r}^{(l)}_\mathcal{B})$ and covariance matrix $\boldsymbol{\Sigma}^*_{\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B}}(\mathbf{r}^{(l)}_\mathcal{B})$ of the local likelihood approximation from the formulae in equation 9.
3) For each sample $\mathbf{r}^{(l)}_\mathcal{B}$, use the computed $\boldsymbol{\mu}^*_{\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B}}(\mathbf{r}^{(l)}_\mathcal{B})$ and $\boldsymbol{\Sigma}^*_{\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B}}(\mathbf{r}^{(l)}_\mathcal{B})$ to evaluate the approximate local likelihood $p^*(\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B} = \mathbf{r}^{(l)}_\mathcal{B})$.

4) For each sample $\mathbf{r}_{\mathcal{B}}^{(l)}$, extract $\mathbf{r}_{\mathcal{A}}^{(l)}$ and define unnormalized and normalized weights by respectively

$$v^{(l)} = p^*(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}} = \mathbf{r}_{\mathcal{B}}^{(l)}) \;\text{ and }\; w^{(l)} = \frac{v^{(l)}}{\sum_{j=1}^{L} v^{(j)}}. \qquad (10)$$

Each pair $(\mathbf{r}_{\mathcal{A}}^{(l)}, w^{(l)}), l = 1, \ldots, L$ may then be used to approximate in principle any quantity related to $p(\mathbf{r}_{\mathcal{A}}|\mathbf{d})$. Some examples are:

1) $\boldsymbol{\mu}^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) = \sum_{l=1}^{L} w^{(l)} \mathbf{r}_{\mathcal{A}}^{(l)}$.
2) $p^*(\mathbf{r}_{\mathcal{A}} \in S|\mathbf{d}) = \sum_{l=1}^{L} w^{(l)} \mathbf{1}_{\{\mathbf{r}_{\mathcal{A}}^{(l)} \in S\}}$ for some set $S \in (-\infty, \infty)$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.
3) For $\mathbf{r}_{\mathcal{A}}$ with discrete prior distribution $p(\mathbf{r}_{\mathcal{A}})$: $p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) = \sum_{l=1}^{L} w^{(l)} \mathbf{1}_{\{\mathbf{r}_{\mathcal{A}}^{(l)} = \mathbf{r}_{\mathcal{A}}\}}$.
4) For $\mathbf{r}_{\mathcal{A}}$ with continuous prior distribution $p(\mathbf{r}_{\mathcal{A}})$: $p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) = \sum_{l=1}^{L} w^{(l)} K_h(\mathbf{r}_{\mathcal{A}} - \mathbf{r}_{\mathcal{A}}^{(l)})$, where K_h is a scaled kernel density function with bandwidth h (see e.g., Silverman, 1986).

In some cases, it is fruitful to sample from the prior conditioned on some criterion, rather than directly. This is especially the case when the interesting parts of the sample space are a priori unlikely or naturally separated in e.g., a discrete and continuous part. Appendix C gives more details on this subject.

Full-region inversion

Two requirements must be met to use our approximation framework. First, relation 8 must hold with a Gaussian error term $\tilde{\boldsymbol{\varepsilon}}$, at least approximately. Second, it must be possible to sample from $p(\mathbf{m}_{\mathcal{C}}, \mathbf{r}_{\mathcal{B}})$. Sampling from $p(\mathbf{r}_{\mathcal{B}})$ is ensured by the latter requirement.

A major advantage of this framework first becomes apparent when considering inversion of a larger region. As mentioned earlier, inversion over a gridded region $\mathbb{A}$ is carried out by applying the presented technique to each grid cell $\mathcal{A} \in \mathbb{A}$. This allows for parallelization which under the following additional stationarity assumptions, results in a computationally efficient inversion procedure:

- $\mathcal{B} = \mathcal{B}(\mathcal{A}), \mathcal{C} = \mathcal{C}(\mathcal{A}), \mathcal{D} = \mathcal{D}(\mathcal{A})$ are all specified with relation to $\mathcal{A}$ only.
- $p(\mathbf{d}_{\mathcal{D}(\mathcal{A})}, \mathbf{m}_{\mathcal{C}(\mathcal{A})}, \mathbf{r}_{\mathcal{B}(\mathcal{A})})$ is stationary with respect to $\mathcal{A}$.

These assumptions ensure that the local subsets follow $\mathcal{A}$ when it shifts from one cell to another and that the joint distribution of the local subsets is independent of this shift. The mean and covariance matrix functions of the local likelihood approximation in equation 9 thus hold for all $\mathcal{A} \in \mathbb{A}$ and need to be computed only once. The only part of the procedure that changes from one position to another is the local geophysical data $\mathbf{d}_{\mathcal{D}}$. Hence, global inversion of $\mathbb{A}$ may be carried out by simply repeating steps 3) and 4) in the above Monte Carlo routine for each cell $\mathcal{A} \in \mathbb{A}$. The stationarity assumptions are not strictly required for our method to work out but are introduced for computational speed-up. There is speed-up also if the assumptions hold only for certain parts of $\mathbb{A}$.

Selecting local subsets

Selecting appropriate local subsets $\mathcal{B}, \mathcal{C}$, and $\mathcal{D}$ is essential in this approximation framework. The dimension of these is an issue of approximation accuracy versus computation speed, but it is not necessarily so that choosing them to be larger independently of each other will lead to a better approximation. A sensible and efficient approximation framework thus requires careful selection of these subsets.

By narrowing $\mathbf{d}$ to the local data $\mathbf{d}_{\mathcal{D}}$, we limit the information used for the local likelihood evaluation. Using too little data gives a considerable loss of information, but using too much data leads to infeasible computation. Also, including data with minor relation to the target variable $\mathbf{r}_{\mathcal{A}}$ does not bring anything new. Selecting $\mathcal{D}$ to be the region directly influenced by the geophysical properties in $\mathcal{A}$ gives a reasonable trade off. For other geophysical models, this region is not as easily determined as for seismic amplitude data. That situation is discussed in Appendix D.

As mentioned earlier, $\mathcal{C}$ should be the region for which the geophysical properties influence the local data $\mathbf{d}_{\mathcal{D}}$. Expanding $\mathcal{C}$ beyond this would not improve the fit, but only make the task of fitting $p^*(\mathbf{m}_{\mathcal{C}}|\mathbf{r}_{\mathcal{B}})$ more complex.

A common assumption within the geosciences is that all that could be learned about local geophysical properties from rock properties are found in the local variables; that is, $p(\mathbf{m}_{\mathcal{X}}|\mathbf{r}) = p(\mathbf{m}_{\mathcal{X}}|\mathbf{r}_{\mathcal{X}})$ for any set of region cells $\mathcal{X}$. Based on this relation, it is clear that it would be optimal to set $\mathcal{B}$ equal to whatever $\mathcal{C}$ is set to. Because the dimension of $\mathcal{B}$ is exactly the dimension of the integral that the weighted Monte Carlo routine relies on, practicalities usually make such a choice impossible. The reason is that to maintain the accuracy of the Monte Carlo routine when the dimension of the integrand increases, a larger number of prior samples, and hence likelihood evaluations, is required. However, assuming that the Monte Carlo accuracy is maintained (at a higher computational cost), it is likely that expanding $\mathcal{B}$ would lead to a better approximation of the true posterior and, consequently, more accurate predictions of the rock properties and their uncertainties. If $\mathcal{B}$ is too small, important features of the local data $\mathbf{d}_{\mathcal{D}}$, which are "transferred" to the influence variable $\mathbf{m}_{\mathcal{C}}$, may be caused by characteristics of rock properties outside $\mathcal{B}$ and thereby not be associated with appropriate values of the target variable $\mathbf{r}_{\mathcal{A}}$.

The local subsets may also be selected by comparing the properties of samples from the local likelihood approximation $p^*(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}})$ obtained for different choices of $\mathcal{B}, \mathcal{C}$ and $\mathcal{D}$ to those of the true local likelihood $p(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}})$. Alternatively, confidence in certain local subsets may be built based on application to a synthetic case in which the resulting approximated posterior distribution can be compared with the true rock properties. We rely on this latter approach for selection of $\mathcal{B}$ in the upcoming data illustrations.

SYNTHETIC DATA TEST

In the next section, we will consider a 4D survey from the Sleipner CO_2 injection project as a real data example. We shall use that case as a motivation in this synthetic example. The synthetic data shall reflect a region similar to the Utsira Formation in the Sleipner field offshore Norway, where CO_2 has been injected for storage and seismic base and monitor surveys have been conducted. The CO_2 is typically trapped underneath thin layers of shale within the formation or under the formation top. The main objective is to "map" the CO_2 based on the saturation in the region. Hence, we define the rock property of interest $\mathbf{r}$ as the CO_2 saturation at the time point of one of the monitor surveys (herein for simplicity referred to as the saturation), assuming no injection prior to the base survey. The geophysical properties $\mathbf{m}$ are defined as the change in the logarithm of the P- and S-wave velocity and density from base to monitor time. Finally, **d** denotes the change in properly aligned seismic angle gathers from base to monitor time.

In this synthetic example, we shall be content working with models and data specified on a gridded 2D region $\mathbb{A}$ of size 3500 m × 280 ms (length × two-way traveltime [TWT]), where each of the 19,600 grid cells have size 25 m × 2 ms. We shall perform inversion of synthetic geophysical data to saturation for the complete region $\mathbb{A}$. The synthetic case is constructed by specifying global stochastic rock-physics and geophysical models in this region. Note in particular that the stochastic rock-physics model involves also other rock properties than the saturation **r**.

Let us first consider the rock-physics model. There are few reliable measurements of velocity and density from the injection well (Rabben and Ursin, 2011). The problems with the logging are probably due to issues with the very loose sand in Utsira. The P-wave velocity in brine-filled sand is slightly more than 2000 m/s, and the rock matrix in this region corresponds to loose sand. The rock-physics model we use for the Utsira sand is consistent with these data. In our model, we use a Reuss mix of the mineral point and a high-porosity member constructed using Walton's model with 45% porosity (see e.g., Mavko et al., 2009). The parameters and corresponding uncertainty model for this rock-physics model are given in Table 2, which also provides the temperature and pressures used to derive the fluid properties. The rock-physics model is similar to the model used by Arts et al. (2004), but a notable difference is that we use the velocity and density for brine, which are compatible with the most recent pressure and temperature measurements in the formation (Batzle and Wang, 1992; Alnes et al., 2011). When matching the observed velocity in Utsira, this gives softer sand than is used in Arts et al. (2004). The effect of the saturation on geophysical properties is computed using fluid substitution, that is, Gassmann's equations (Mavko et al., 2009) where a homogeneous fluid mix (Reuss) is assumed. Due to the nature of CO_2 and the soft rock at the Sleipner injection site, saturation is the main cause of variability for the change in geophysical properties from base to monitor time. This is illustrated in Figure 3, which displays samples of brine- and CO_2-saturated rocks from the rock-physics model, overlaid on a rock-physics template.

Table 2. Rock-physics model parameters. The mineral parameters have correlation 0.99 between properties. Details on the two and four parameter beta distributions are given in Appendix B.

Property	Distribution
Mineral bulk modulus (GPa)	N(35.4, 3.2)
Mineral shear modulus (GPa)	N(27.3, 7.4)
Mineral density (g/ccm)	N(2.647, 0.008)
Brine bulk modulus (GPa)	Fixed = 2.538
Brine density (g/ccm)	Fixed = 1.027
CO_2 bulk modulus (GPa)	Fixed = 0.065
CO_2 density (g/ccm)	Fixed = 0.686
Coordination number	Fixed = 7.3
Friction factor	Beta(5.0, 0.8)
Porosity	Beta_4(2.0, 2.0, 0.27, 0.42)
Pore pressure (MPa)	Fixed = 10
Effective pressure (MPa)	Fixed = 10
Temperature (°C)	Fixed = 36

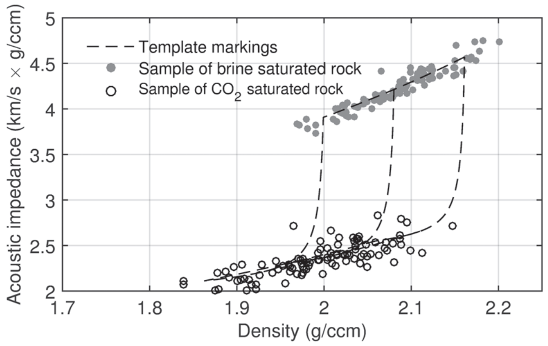

Figure 3. Rock-physics template: The boundaries of the template correspond to porosity of 40% and 30% and, CO_2 saturation of 0 and 1. The center lines of the template correspond to porosity and CO_2 saturation of, respectively, 35% and 0.5. For saturation, that line is partly hidden close to the lower boundary of the template.

The geophysical likelihood model is of the form described by equation 7, i.e., Gaussian with linear mean function. The linear multiplicand here is $\mathbf{G} = \mathbf{WAD}$; **W** is a block diagonal matrix representing the smoothing with a 25 Hz Ricker wavelet; **A** is the matrix of weak-contrast coefficients for each of the offset angles 5° (near), 20° (mid), and 35° (far) as defined by Aki and Richards (1980); and **D** is a differential matrix producing contrasts of the geophysical properties. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ for the error term is block diagonal with independence between the different offsets, and the standard deviations are 0.04 for near, 0.05 for mid, and 0.06 for far offset.

The synthetic saturation for our region of interest $\mathbb{A}$ is constructed to mimic the high-saturation layers in the Utsira Formation. It contains multiple layers with a flat top and variable thickness. In particular, 5% of the region cells possess saturation greater than 0.6. The synthetic saturation and noisy geophysical data sampled from the rock model are displayed in Figure 4. Note in particular that the peak amplitude of the seismic data does not follow a flat top.

Let us now turn to the inversion, in which we will limit our approximations to include vertical dependence. We shall define a stochastic prior model for the saturation, and otherwise we use vertical analogs of the rock-physics and geophysical models described above as the basis for our approximations. The prior model for saturation is defined as a transformation of a Gaussian copula (Joe, 1997), which models the effect that adjacent cells are more correlated than distant ones. The marginal distribution in each cell has a point mass at zero and a continuous distribution from 0 to 1. In each cell, the prior probability for the saturation being zero is 99%, whereas the saturation is distributed as Beta(6, 1.5), as shown in Figure 5, when it is strictly positive (because the distribution of saturation has a discrete and a positive part, this example illustrates the applicability of our approach to both these types of rock properties). Consult Appendix B for details on the beta distribution.

The Gaussian copula has a stationary exponential covariance function with range parameter $R = 50$ ms in the parametrization $C(h) = \exp(-3|h|/R)$, where h is the vertical "distance" (in ms) between two locations.

Because we limit our approximations to only include vertical dependence, the local subsets $\mathcal{B}, \mathcal{C}$, and $\mathcal{D}$ only include variables within the same vertical profile as $\mathcal{A}$. They are also all centered in $\mathcal{A}$. Specifically, we let $\mathcal{D}$ include 20 ms above and below $\mathcal{A}$, having dimension 21, whereas $\mathcal{C}$ includes 44 ms above and below $\mathcal{A}$, having dimension 45. As we shall see shortly, $\mathcal{B}$ is selected on a trail and error basis through comparison with the true synthetic saturation.

In our approximations, we estimate $p^*(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$ by splitting the sampled pairs $(\mathbf{m}_\mathcal{C}, \mathbf{r}_\mathcal{B})$ into $K = 4$ classes depending on whether the saturation in the two boundaries of $\mathcal{B}$ (i.e., the shallowest and deepest cells in $\mathcal{B}$) are zero or strictly positive. The mean functions $\boldsymbol{\mu}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}(\mathbf{r}_\mathcal{B})$ for each class k are estimated by a MARS procedure, where generalized cross validation is used to specify the tuning parameters in the regression. The covariance matrices $\boldsymbol{\Sigma}^*_{\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B},k}$, $k = 1, \ldots, 4$ are estimated by the range spanning covariance estimation routine in Appendix A.

Between 45,000 and 100,000 samples are used for each of the $K = 4$ classes to fit the approximate local rock-physical likelihood $p^*(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$. Because nonzero saturation occurs in only 1 out of 100 cells by direct sampling, it is beneficial to oversample positive $\mathbf{r}_\mathcal{B}$. This will give more robust approximations for the whole sample space of the target variable $\mathbf{r}_\mathcal{A}$. Hence, two sets of $\mathbf{r}_\mathcal{B}$ samples (one conditioned on $\mathbf{r}_\mathcal{A} = 0$ and one conditioned on $\mathbf{r}_\mathcal{A} > 0$) are used in the weighted Monte Carlo routine (see Appendix C for further details). This sampling procedure, with each of the two sets being of size 10^5, will be used throughout the paper.

To fully specify the inversion, we must choose the size of $\mathcal{B}$ for this synthetic case. (The chosen size will also be passed forward to the upcoming real case.) We do this by pointwise comparing predictors derived from the local inversions with the true synthetic saturation. For that comparison, and for the remainder of this paper, we will use the approximated marginal posterior mean saturation in each cell as a predictor for the true unknown saturation. Increasing the dimension of $\mathcal{B}$ should in theory (on average) result in a closer match between the truth and the approximation, represented by the bias. At the same time, it increases the variability between the performances of different sets of prior samples, represented by the (Monte Carlo) variance. We define the optimal $\mathcal{B}$ as the one minimizing the mean squared error (MSE), which decomposes nicely into squared bias plus variance. Empirical versions of squared bias, variance, and MSE averaged over the cells in the full 2D region are shown in Figure 6 for varying size of $\mathcal{B}$ and constant sample size — when using the posterior mean as predictor. Note that the MSE is computed on a cell-by-cell basis; hence, it penalizes any minor misalignment in the predicted saturation severely. In terms of the bias versus variance trade-off (i.e., MSE minimization), we deem $\mathcal{B}$ of dimension 17 the optimal for the current sampling regime. That is, the optimal $\mathcal{B}$ includes 16 ms above and below $\mathcal{A}$. As seen from the figure, the squared bias is fairly flat to the right of $\dim(\mathcal{B}) = 17$. This indicates that increasing the sample size further would not improve the fit considerably. The slight increase in bias for the largest dimensions is an indirect effect of the relatively small sample size.

Figure 7 shows the marginal posterior means for the chosen $\mathcal{B}$ and explained sampling regime, along with its difference from the synthetic truth. As seen from the figure, the posterior mean matches the true saturation well over the complete 2D region, at thicker and thinner layers of high saturation. Note the good positioning of the top and base of high-saturation layers also when the layer thickness is below the tuning thickness. As expected, the positioning of thin layers far below the tuning thickness results in some misalignment of the high-saturation layers. They are, however, still detected, and vertical profile averages are generally well preserved. Compared with using the prior directly without support of the data (MSE = 0.038), our framework reduces the empirical MSE almost an order of magnitude (MSE = 0.0050). Also, the prior average saturation in

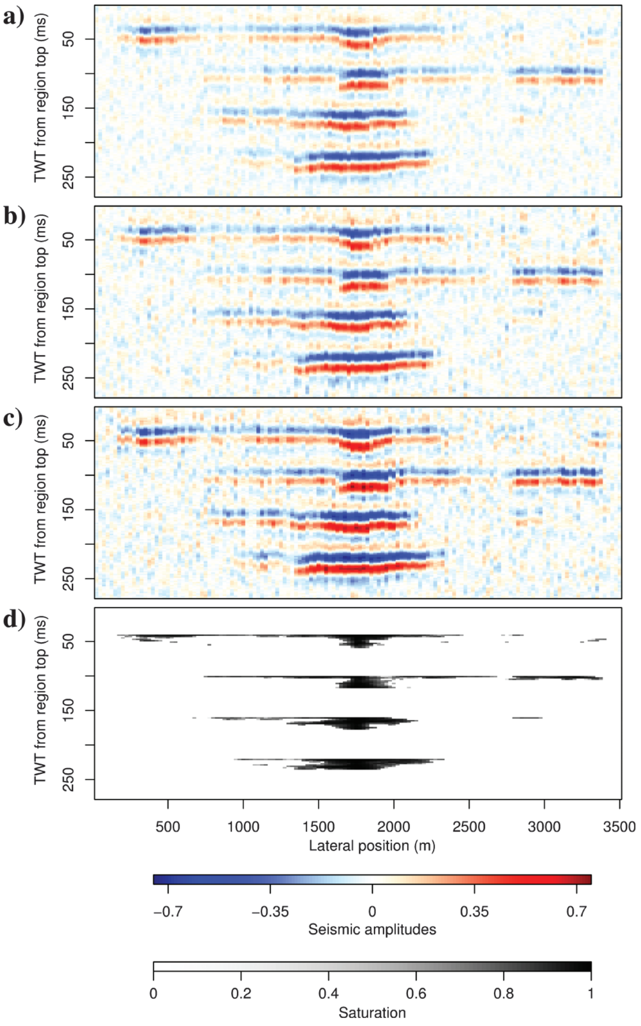

Figure 4. (a-c) Synthetic seismic difference data (near, mid, and far) relevant for the 2D region of interest. (d) Generated synthetic saturation at monitor time $\mathbf{r}$ for the complete synthetic 2D region.

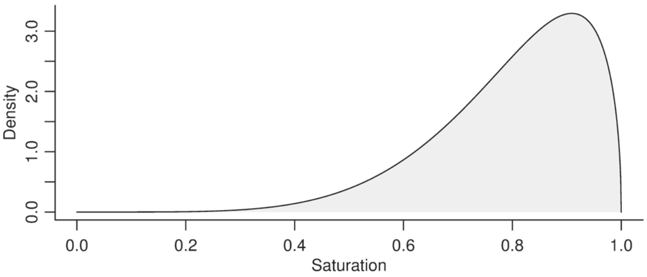

Figure 5. Marginal prior distribution of positive saturation: A beta distribution with shape parameters $a = 6$ and $b = 1.5$, having mean 0.8 and variance approximately 0.02.

the region is 0.0080, whereas the posterior regionwise average is 0.0455. These should be compared with the regionwise average of the true synthetic saturation, which is 0.0449. Thus, even if the approach operates locally, it still gives information about bulk properties. These results are not only an outcome of our framework and procedure, but they also depend on the statistical model and grid being used. Because the same grid is used to generate the synthetic case and perform inversion, potential bias caused by grid cells being misaligned with high-saturation layers are not present in this synthetic case. Also, the selection bias is not accounted for when the same data are used to tune a parameter and present performance results. However, the flat behavior of the MSE curve around the minimum point in Figure 6 indicates that such selection bias is insignificant here.

To properly evaluate the accuracy of our approximation method, the approximated marginal posterior distributions should be compared with the true posterior distribution. Because "exact" methods for computing the posterior distribution are computationally extremely costly, this is not feasible for the full 2D region. To accompany the above evaluation and comparison with the actual synthetic saturation, we do however evaluate the true posterior distribution in a single vertical profile (positioned at 1625 m) for comparison with our procedure. The true posterior for the vertical trace is obtained by running a blockwise Metropolis Hastings MCMC scheme (Bolstad [2009], chapter 6.3) with an independence sampler corresponding to the conditional prior distribution. This "brute force" MCMC procedure required several days of CPU running time to provide reliable results — whereas our approach produced approximate results within seconds. For this vertical profile, Figure 8 shows pointwise 80% credibility intervals (CI) (the range between P10 and P90), posterior means and medians for the true posterior and our local approximation procedure with a few different sizes of $\mathcal{B}$. Also plotted are the seismic difference data and true synthetic saturation. The figure illustrates the typical behavior for our method when varying the size of $\mathcal{B}$, while the sample size is kept constant. When $\mathcal{B}$ is too small, the approximated posteriors are simplified too much. Increasing the size of $\mathcal{B}$ increases the level of detail, but a too-large $\mathcal{B}$ generates unwanted noise with thin wrongly predicted high-saturation layers, and gives unstable results due to the relatively small sample size. For the MSE-optimized procedure, where $\dim(\mathcal{B}) = 17$, the posterior mean, median, and the pointwise 80% credibility intervals match those of the true posterior very well. This indicates that the approximation method works as intended also on smaller scales.

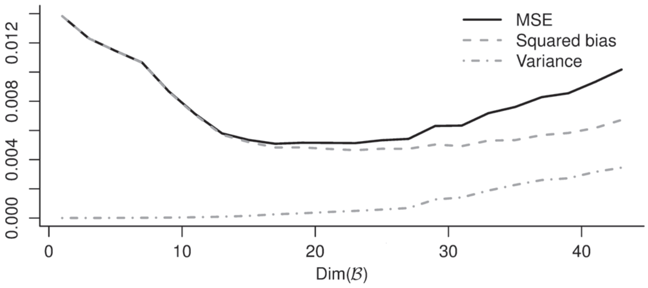

Figure 6. Empirical estimates of the MSE, squared bias, and Monte Carlo variance are plotted for different dimensions of $\mathcal{B}$ in the synthetic 2D case. These are computed based on five different prior sampling seeds (with totally 2×10^5 samples) for each $\mathcal{B}$ centered in $\mathcal{A}$.

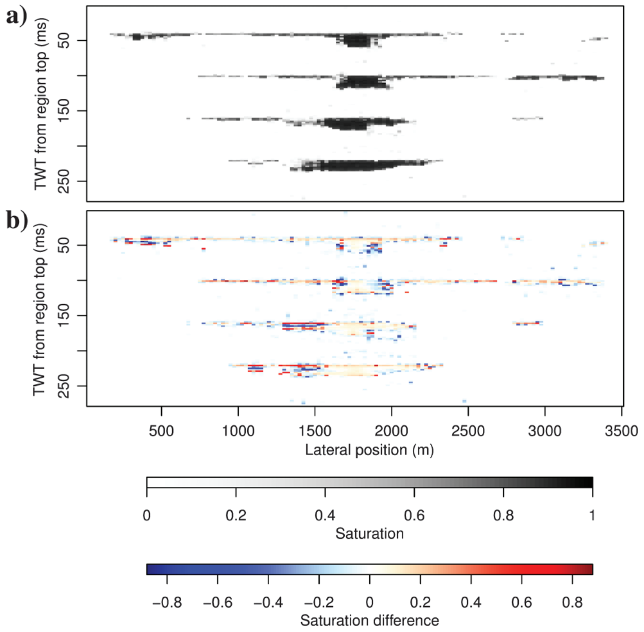

Figure 7. (a) Inversion results for the synthetic 2D region shown through approximated marginal posterior mean saturations in each cell. (b) The difference between the true synthetic saturation and the prediction in panel (a).

Finally, we compare predictions from our procedure with the corresponding ones based on the frequently used Gaussian inversion approach of Buland and Omre (2003). Because the approach of Buland and Omre (2003) actually is a geophysical inversion approach, we temporarily change our focus to prediction of the change in the logarithm of the density ρ, instead of saturation. Figure 9 shows the posterior means for the 2D region discussed above for both approximation methods along with the synthetic true change in log density. As seen from the figure, the predictions based on our method are more distinct and clear compared with the vaguer predictions provided by the Gaussian inversion. The latter perform poorly on thin layers of reduced density, and also predict areas with a positive change in log density not present in the synthetic data. Hence, our method improves substantially upon Buland and Omre (2003) in terms of prediction accuracy.

REAL DATA CASE

The Sleipner CO_2 injection project aims at storing CO_2 captured from the gas production in the Sleipner field offshore Norway by leading compressed CO_2 down to the Utsira Formation through an injection well. We consider geophysical data from a seismic 4D survey of this formation and aim at monitoring or mapping the CO_2 based on the saturation. We concentrate on the changes from a base survey in 1994 (before injection) and until a monitor survey in 2006. Because the saturation is effectively zero everywhere prior to injection, the changes in saturation correspond to the amount at monitor time. The geophysical data consist of changes in seismic AVO data from base to monitor time for three different offsets. The data are aligned using rms and pushdown data prior to difference computation. We concentrate on a west–east-directed 2D region intersecting the injection well. The region spans more than 2900 m × 334 ms, has a seismic sampling resolution of 25 m × 2 ms, and is positioned approximately 800 m below sea level.

The setup of the model to be approximated and its various parameters are essentially the same as for the synthetic data test, with the exception of some parameter specifications for the geophysical likelihood $p(\mathbf{d}|\mathbf{m})$. In the present case, $\mathbf{W}$ consists of 35, 30, and 25 Hz Ricker wavelets for the near (12.5°), mid (25°), and far offsets (40°), respectively. The frequency content of the wavelets was set to match those in the seismic data. For near and far stacks, this was done based on the common part of the base and monitor surveys in a region directly above the reservoir. This part has a slightly lower frequency content than the individual parts, indicating a lower signal-to-noise ratio at higher frequencies. Processing issues made such detailed analysis impossible for the far stack. Hence, the far wavelet was set only from the frequency content in the base survey, accounting also for a high end frequency loss. The near and mid wavelets are scaled by a factor two compared with the far offset. This was determined by analysis of the base data, which gave twice as strong a signal for the near and mid stack than for the far — most likely caused by survey and processing effects. The standard deviation of the error term in the geophysical likelihood model is 0.5 for near and mid offsets and 0.2 for far offset. This is larger than observed directly above the reservoir, reflecting that there are larger alignment errors and stronger amplitude effects in the target region than directly above.

All other model parameters, local subsets, and other parts of the inversion setup are the same as for the synthetic case. The optimal size of the neighborhood variable established for the synthetic case is used in the sampling. The seismic AVO data relevant for the 2D region and the resulting marginal posterior means are shown in Figure 10. The results indicate several wide sections of increased saturation, which seem to generally match well with the reflections from the seismic data.

Our approach provides more than the best estimate. Figure 11 shows more detailed inversion results for a vertical profile near the injection well, together with the seismic amplitudes of that profile.

The figure indicates that there are five main high-saturation layers with tops approximately at vertical positions 60, 90, 110, 145–160, and 190 ms. This matches well with a straightforward visual interpretation of the reflections of the seismic amplitudes. The approximate posterior uncertainty is small at the high-saturation areas corresponding to the three shallowest and the deepest of these positions, indicating that the presence of these layers is fairly certain. On the other hand, the contiguously wide pointwise 80% credibility intervals reaching all the way down to zero for TWTs of 145–180 ms,

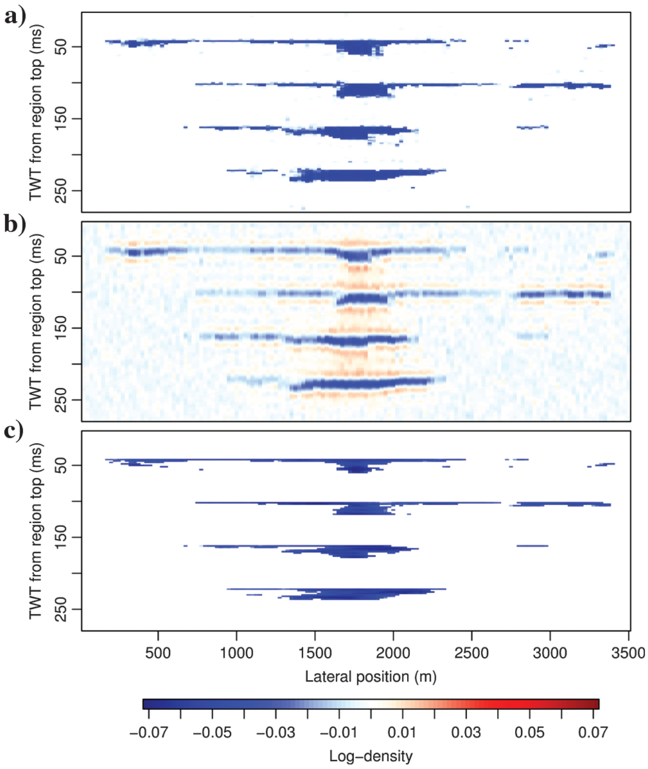

Figure 9. Comparison of predictions of changes in log density for the synthetic 2D region. (a) Predictions (means) based on our method with local subsets as above, (b) predictions (means) based on the Gaussian inversion approach (Buland and Omre, 2003), and (c) true synthetic change in log density.

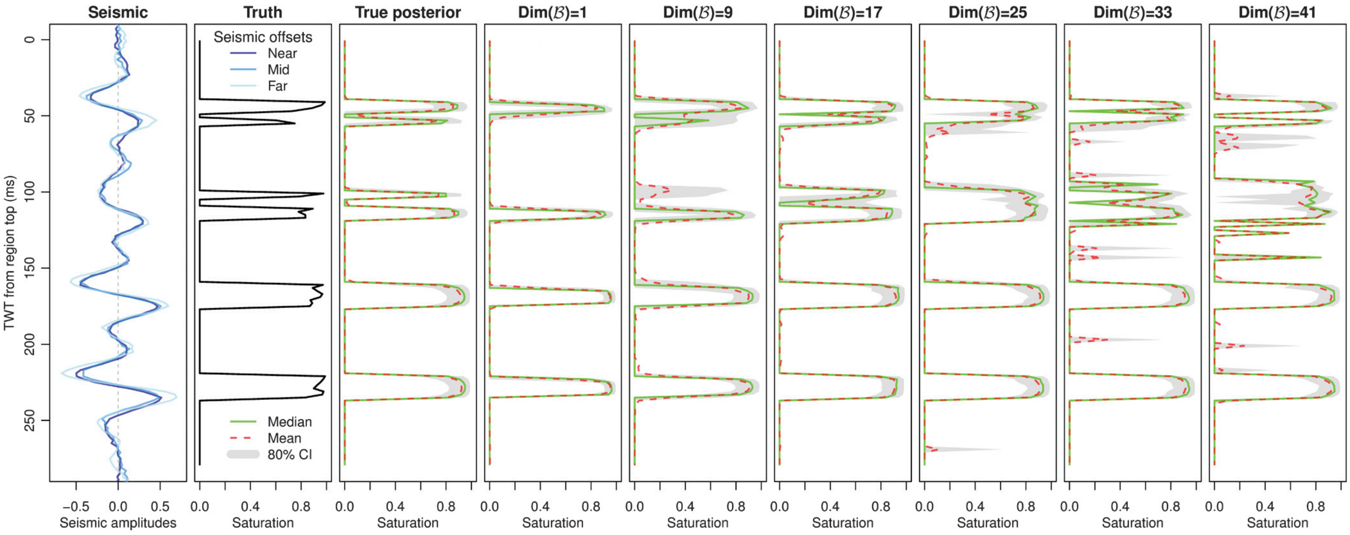

Figure 8. Seismic amplitudes, true saturation, and details of the true posterior and our approximation using different sizes of $\mathcal{B}$ in a vertical profile of the synthetic data positioned at 1625 m. The 80% CI shows the pointwise range between the P10 and P90.

indicate larger uncertainty for the levels of saturation. Together with the disagreement of the mean and median predictors, this suggests that the presence and depth of one or two high-saturation layers are highly uncertain within this range.

Boait et al. (2012) study the CO_2 migration in the Utsira Formation after injection using seismic time-lapse data with several monitor surveys, and they present a qualitative interpretation of the horizons in a 2D region close to the injection point. The qualitative interpretations match our profile inversion results very well by indicating the same five layers. On a larger scale, the qualitative interpretations also match our inversion results well. Some deviations are, however, seen in an area slightly east of the injection well, and in the deepest areas where the signals in our data are weak.

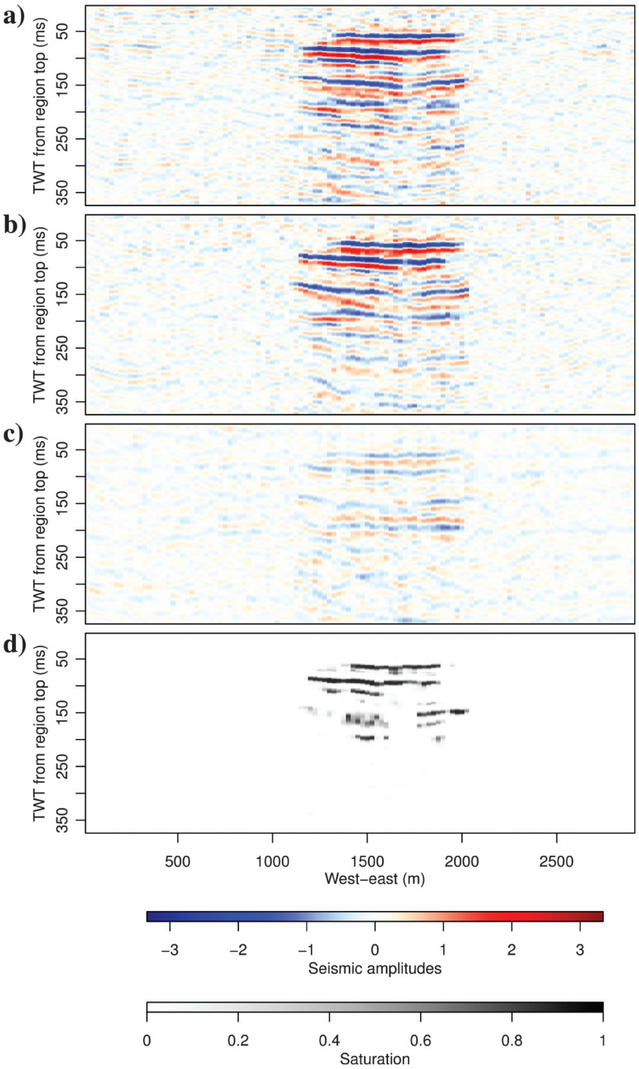

Figure 10. (a-c) The three different offsets (near, mid, and far) of the seismic AVO differences between base and monitor time relevant for the real case 2D region. The near and mid offsets are scaled by a factor two compared with the far offset. (d) Inversion results shown through marginal posterior mean saturations for each cell in the gridded 2D region of the real data case.

DISCUSSION

Our approach has connections to other general statistics-based and geophysical-motivated approximation techniques. Similar to our approach, the integrated nested Laplace approximation (INLA) method of Rue et al. (2009) approximates parts of the model by Gaussian distributions, uses their convenient properties, and solves a lower dimensional final integral by a numerical routine. The method does not involve local subset parameters, however. Those Gaussian approximations are also applied to different parts of the model than ours, and therefore require other types of assumptions — typically being unrealistic within geoscience applications. Of the existing geophysics types of techniques, the connection to Buland et al. (2008) is perhaps the closest. Like us, Buland et al. (2008) perform global inversion by focusing on one cell at a time and rely on Gaussian approximations. However, they only consider the direct behavior of the rock and geophysical properties in the cell under current consideration, and not the spatial neighbors, which is key in our framework. They do not require a Gaussian rock-physical likelihood (neither globally nor locally), but this also restricts the applicability of their technique to discrete rock properties, such as facies or lithology classes.

The overall procedure used in our framework can also be applied when $p(\mathbf{d}_{\mathcal{D}}|\mathbf{r}_{\mathcal{B}})$ is approximated by a non-Gaussian distribution. There are, however, several benefits of relying on Gaussianity. Among them are the simple way that conditional distributions are handled, the computational savings available when preevaluating larger parts of the Gaussian distribution (under stationarity assumptions), and usage of the common assumption that the geophysical likelihood $p(\mathbf{d}|\mathbf{m})$ is Gaussian with a linear mean function. The

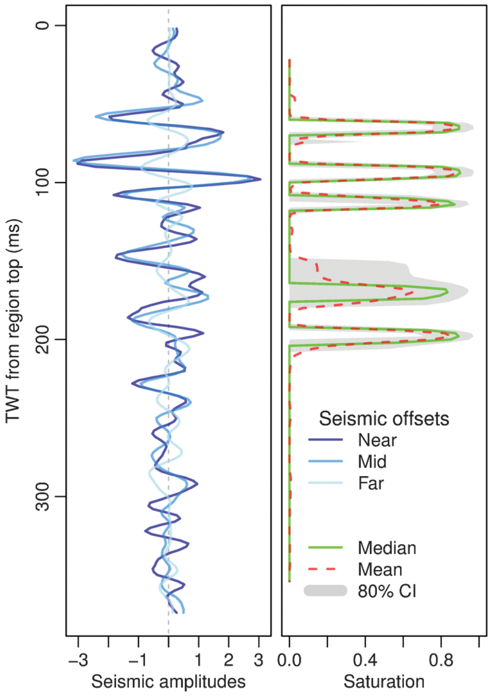

Figure 11. Seismic AVO data compared with the approximate marginal posterior distributions of saturation in a vertical profile close to the injection well (west–east position 1475 m in the real case 2D region).

Gaussian restriction can, however, be relaxed without losing all the beneficial properties. This is achieved by allowing the local geophysical likelihood $p(\mathbf{d}_\mathcal{D}|\mathbf{m}_\mathcal{C})$ and/or the local rock-physical likelihood $p(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$ to rather be approximated by Gaussian mixture distributions. This increases the flexibility because it allows for skewness and/or multimodal conditional distributions, which may improve the accuracy of the procedure if Gaussianity is inappropriate. This extension has the consequence that the local likelihood approximation $p^*(\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B})$ also becomes a Gaussian mixture. The total number of mixture components is the product of the number of mixture components of the two likelihoods. Because evaluation of a Gaussian mixture likelihood with q mixture components is computationally q times more expensive than evaluating a regular Gaussian likelihood, there should, however, be substantial reasons for including many extra levels of complexity. Another approach would be to approximate e.g., $p(\mathbf{m}_\mathcal{C}|\mathbf{r}_\mathcal{B})$ by the selection Gaussian distribution of Rimstad and Omre (2014b), with the consequence that $p^*(\mathbf{d}_\mathcal{D}|\mathbf{r}_\mathcal{B})$ also becomes selection Gaussian distributed. However, further investigations are required for such an extension.

Although the methodology of the framework is motivated by inversion of single cells, the framework is not strictly restricted to this situation. All theory and methodology work out even if $\mathcal{A}$ refers to several cells or some weighted sum of these, and also if $\mathbf{r}$ corresponds to two or more rock properties. The latter is handled simply by sampling, say, the porosity and saturation jointly with the geophysical properties when approximating the local rock-physics likelihood in equation 6, and when sampling rock properties in the Monte Carlo routine. Further, the local subsets $\mathcal{B}, \mathcal{C}$, and $\mathcal{D}$ are not restricted to be contiguous and centered around cell $\mathcal{A}$ as illustrated in Figure 2, even though that is the most natural choice. For some type of applications (e.g., in seismic tomography with complex raypaths), it may be more appropriate to include scattered cells in $\mathcal{B}$. This would, however, call for a more comprehensive local subset selection process.

Our approach does not require closed form expressions for the rock properties' prior distribution and the rock-physical likelihood, neither globally nor locally. We only require that samples from the local distributions are obtainable. This is advantageous in the fairly common setting where it is hard to specify closed form expressed models possessing the desired randomness features, but where sampling-based model specifications are easier to put up.

As mentioned, our framework is well suited for parallelization on multicore computers or graphics processing unit (GPU) accelerators. For our data application, the methodology was straightforwardly implemented in the statistical programming language R (R Core Team, 2014). The complete inversion procedure for the synthetic data test was run in parallel on a Windows-based laptop with an Intel i7 2.6GhZ processor and four cores in less than half an hour. This included the initial fitting procedure and the weighted Monte Carlo routine for all the 19,600 individual cells. Because no attempt was made on boosting this performance, there is still room for a significant reduction in the computing time. The main contributions would be a devoted implementation in a precompiled language and massive parallelization on high-performance cores.

The mixing of the CO_2 and brine is a frequently discussed topic. As an alternative to the homogeneous mix that we use in our applications, patchy mixtures of fluids have been suggested (Ghaderi and Landrø, 2009). For the real case, we also tested a simple model for a patchy fluid mix. The patchy model was obtained by substituting the Reuss average with a Voigt average when deriving the properties of the effective pore fluid. For the vertical profile of Figure 11, this reduced the uncertainty in the high-saturation layers and introduced the possibility of a low-saturated layer below the top CO_2 layer. Apart from this, the inversion results were very similar.

CONCLUSION

We have presented a general framework for approximate Bayesian inversion of geophysical data into rock properties. The methodology approximates quantities directly related to the marginal posterior distributions of the rock properties defining the Bayesian inversion. The framework is well suited for parallelization, producing potentially very fast inversion results. The generality, parallelization properties, and mild assumptions of the approach should make it attractive for a broad range of geophysical challenges in reservoir characterization, monitoring, and exploration.

The issue of the dimensionality in Bayesian inversion was reduced by considering local behavior for all model components (rock properties, geophysical properties, and geophysical data). This was done by merging a Gaussian approximation to a local rock-physical likelihood with a Gaussian approximation to a local geophysical likelihood, and then applying a weighted Monte Carlo routine to compute the approximate quantities relevant for the inversion.

The synthetic data test showed that our local approach delivers acceptable errors compared to the much slower MCMC solution; and at least in this case, it is substantially more precise than a frequently used Gaussian inversion approach. For the real data case, the inversion results matched well with a previously published qualitative interpretation of the region.

ACKNOWLEDGMENTS

Thanks go to A.K. Furre and E.H. Jensen for providing the data and details of the rock-physics model for the Utsira Formation. We also thank the Sleipner license partners Statoil, ExxonMobil, and Total for permission to use the Sleipner data.

APPENDIX A

RANGE SPANNING COVARIANCE MATRIX

Consider estimation of a range spanning covariance matrix based on n residuals $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n$ of dimension q, with both n and q fairly large. Let $\delta_i = \boldsymbol{\varepsilon}_i^\mathrm{T}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\varepsilon}_i$ (with $\hat{\boldsymbol{\Sigma}} = 1/(n-1)\sum_{i=1}^n \boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\mathrm{T}$ being the standard sample covariance matrix) be a measure of how likely $\boldsymbol{\varepsilon}_i$ is. When the residuals are Gaussian, it follows from arguments involving the central limit theorem that the distribution of δ_i is approximately Gaussian with mean q and variance $2q$. Clever use of Jensen's inequality and the moment generating function of the Gaussian distribution then shows that the maximum of n such Gaussian variables is bounded by $q + 2\sqrt{q\,\log(n)}$. The range spanning estimation method consists of checking whether any of the δ_i exceed this bound. If some do, we reestimate a covariance matrix $\hat{\boldsymbol{\Sigma}}_\mathrm{max}$ representative for the most extreme residuals, and we cleverly merge that with $\hat{\boldsymbol{\Sigma}}$. $\hat{\boldsymbol{\Sigma}}_\mathrm{max}$ is computed by taking the sample covariance of the most extreme residuals only, while still making sure sufficiently many are included to secure stability. The merged covariance matrix is denoted by $\hat{\boldsymbol{\Sigma}}_\mathrm{span}$ and is required to span both $\hat{\boldsymbol{\Sigma}}_\mathrm{max}$ and $\hat{\boldsymbol{\Sigma}}$ in the sense that both $\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}_\mathrm{span}^{-1}$ and

$\hat{\boldsymbol{\Sigma}}_{\max}^{-1} - \hat{\boldsymbol{\Sigma}}_{\text{span}}^{-1}$ are positive semidefinite, i.e., that $\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\text{span}}^{-1}\mathbf{x} \leq \max\{\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}, \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\max}^{-1}\mathbf{x}\}$ for all $\mathbf{x}$. This is obtained by ensuring that the inequality holds for all generalized eigenvectors of $\hat{\boldsymbol{\Sigma}}_{\max}$ with respect to $\hat{\boldsymbol{\Sigma}}$. After reestimation and merging, $\hat{\boldsymbol{\Sigma}}$ is set equal to $\hat{\boldsymbol{\Sigma}}_{\text{span}}$ and δ_i is recomputed. The procedure is repeated until no residuals are very large compared with the bound.

APPENDIX B

DISTRIBUTIONS

Result: Merging two Gaussian distributions

Let $\mathbf{x}|\mathbf{y} \sim \mathrm{N}_{\mathbf{x}}(\mathbf{Hy}, \boldsymbol{\Sigma}_0)$ for some nonrandom matrix $\mathbf{H}$ and $\mathbf{y} \sim \mathrm{N}_{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, $\mathbf{x} \sim \mathrm{N}_{\mathbf{x}}(\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_0)$.

Proof: The two relations are equivalent to $\mathbf{x} = \mathbf{Hy} + \boldsymbol{\varepsilon}_0$, $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}_0 \sim \mathrm{N}_{\boldsymbol{\varepsilon}_0}(0, \boldsymbol{\Sigma}_0)$ independently of $\boldsymbol{\varepsilon} \sim \mathrm{N}_{\boldsymbol{\varepsilon}}(0, \boldsymbol{\Sigma})$. Hence, $\mathbf{x} = \mathbf{H}(\boldsymbol{\mu} + \boldsymbol{\varepsilon}) + \boldsymbol{\varepsilon}_0 = \mathbf{H}\boldsymbol{\mu} + \boldsymbol{\varepsilon}'$, where $\boldsymbol{\varepsilon}' = \mathbf{H}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}_0 \sim \mathrm{N}_{\boldsymbol{\varepsilon}_0}(0, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_0)$, which again is equivalent to $\mathbf{x} \sim \mathrm{N}_{\mathbf{x}}(\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_0)$.

Two- and four-parameter beta distributions

The (two-parameter) beta distribution $\mathrm{Beta}(a, b)$ has shape parameters a, b and continuous probability distribution function $f_2(y; a, b) \propto y^{a-1}(1 - y)^{b-1}, y \in [0, 1]$. It has mean $a/(a + b)$ and variance $ab/[(a + b)^2(a + b + 1)]$.

The four-parameter beta distribution $\mathrm{Beta}_4(a, b, c, d)$ is a Beta(a, b)-distribution scaled and shifted to match the support $[c, d]$. Its probability distribution function is scaled by $d - c$ and shifted by c: $f_4(y;a,b,c,d) \propto f_2((y-c)/(d-c);a,b), y \in [c,d]$; its mean is $[a(d - c)]/(a + b) + c$; and its variance is $[ab(d - c)^2]/[(a + b)^2(a + b + 1)]$.

APPENDIX C

WEIGHTED MONTE CARLO WITH CONDITIONAL SAMPLING

In some cases, one may wish to apply the weighted Monte Carlo routine when conditioning on the event that $\mathbf{r}_{\mathcal{A}}$ is contained in some part or block of the sample space. That may in particular be the case when the prior distribution for $\mathbf{r}_{\mathcal{A}}$ has discrete and continuous parts because such parts are best treated separately. The technique may also be used to oversample a priori unlikely parts of the sample space of $p(\mathbf{r}_{\mathcal{A}})$ to increase the overall accuracy of the weighted Monte Carlo routine.

Choose the conditions, say, $\mathcal{E}_j, j = 1, \ldots, J$, such that their corresponding blocks on the sample space form a partition, i.e., that the blocks are nonempty disjoint sets whose union is the sample space itself. Assume further that the enumerated routine defined in the weighted Monte Carlo section is carried out J times, once with prior samples conditioned on each of the $\mathcal{E}_j$. Denote, respectively, the extracted $\mathbf{r}_{\mathcal{A}}$ value and corresponding unnormalized and normalized importance weights for the l-th prior sample in the j-th run by $\mathbf{r}_{\mathcal{A}}^{(l,j)}$, $v^{(l,j)}$ and $w^{(l,j)}$, $l = 1, \ldots, L_j$, $j = 1, \ldots, J$. Quantities conditioned on $\mathcal{E}_j$ may be computed analogous to the unconditional ones with weights and samples replaced by those for the j-th condition. For instance, $p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}, \mathcal{E}_j) = \sum_{l=1}^{L_j} w^{(l,j)} \mathbf{1}_{\{\mathbf{r}_{\mathcal{A}}^{(l,j)} = \mathbf{r}_{\mathcal{A}}\}}$ for discrete distributions, and $p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}, \mathcal{E}_j) = \sum_{l=1}^{L_j} w^{(l,j)} K_h(\mathbf{r}_{\mathcal{A}} - \mathbf{r}_{\mathcal{A}}^{(l,j)})$ for continuous distributions. Further, the posterior probability of each condition $\mathcal{E}_j$ may be computed by

$$p^*(\mathcal{E}_j|\mathbf{d}) = \frac{p(\mathcal{E}_j)\frac{1}{L_j}\sum_{l=1}^{L_j} v^{(l,j)}}{\sum_{i=1}^{J} p(\mathcal{E}_i)\frac{1}{L_i}\sum_{l=1}^{L_i} v^{(l,i)}} \qquad \text{(C-1)}$$

Finally, the unconditioned approximated posterior distribution may be computed by$p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}) = \sum_{j=1}^{J} p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d}, \mathcal{E}_j) p^*(\mathcal{E}_j|\mathbf{d})$. Other unconditioned quantities may be computed similarly using the derived quantities and posterior probabilities for each condition $p^*(\mathcal{E}_j|\mathbf{d})$, $j = 1, \ldots, J$, or directly via $p^*(\mathbf{r}_{\mathcal{A}}|\mathbf{d})$.

APPENDIX D

RESOLUTION THEORY FOR SELECTING $\mathcal{D}$ IN GENERAL GEOPHYSICAL PROBLEMS

The formulation we have used in the paper considers the geophysical relation of equation 7, that is, $\mathbf{d} = \mathbf{Gm} + \boldsymbol{\varepsilon}$, with the matrix $\mathbf{G}$ given by the geophysical relations. This is indeed intended because these types of problems are what we aim to solve. However, this general form may be limiting when we consider selection of the local subset $\mathcal{D}$ defining the local data $\mathbf{d}_{\mathcal{D}}$. For AVO data, the influence region may be bounded fairly easily, but when working with e.g., the rms velocities of Buland et al. (2011), the influence region typically span all data below the position considered. For these cases, it is useful to consider resolution theory. Left multiplying both sides of equation 7 by $\mathbf{G}^{\mathrm{T}}(\mathbf{GG}^{\mathrm{T}})^{-1}$ gives $\mathbf{d}' = \mathbf{G}'\mathbf{m} + \boldsymbol{\varepsilon}'$, where $\mathbf{G}' = \mathbf{G}^{\mathrm{T}}(\mathbf{GG}^{\mathrm{T}})^{-1}\mathbf{G}$ is the standard resolution kernel, and $\mathbf{d}'$ and $\boldsymbol{\varepsilon}'$ are, respectively, rescaled data and noise. The resolution kernel will in general be much more locally focused. If the matrix $(\mathbf{GG}^{\mathrm{T}})^{-1}$ is not invertible, one could include a ridge term before inverting, or use some other type of pseudo inversion method to ensure stability of the inverse. It is also possible to perform a preinversion to focus the energy in the problem. This is done by left multiplying both sides of equation 7 by $\hat{\boldsymbol{\Sigma}}_{\mathbf{m}}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\hat{\boldsymbol{\Sigma}}_{\mathbf{m}}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\varepsilon})^{-1}$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{m}}$ is an estimate of the covariance matrix of the geophysical properties and $\boldsymbol{\Sigma}_{\varepsilon}$ is the covariance matrix of the errors.

REFERENCES

Aki, K., and P. G. Richards, 1980, Quantitative seismology: W. H. Freeman & Co.

Alnes, H., O. Eiken, S. Nooner, G. Sasagawa, T. Stenvold, and M. Zumberge, 2011, Results from Sleipner gravity monitoring: Updated density and temperature distribution of the CO_2 plume: Energy Procedia, **4**, 5504–5511, doi: 10.1016/j.egypro.2011.02.536..

Arts, R., O. Eiken, A. Chadwick, P. Zweigel, B. van der Meer, and G. Kirby, 2004, Seismic monitoring at the Sleipner underground CO_2 storage site (North Sea): Geological Society, London, Special Publications, **233**, 181–191.

Avseth, P., T. Mukerji, and G. Mavko, 2010, Quantitative seismic interpretation: Applying rock physics tools to reduce interpretation risk: Cambridge University Press.

Batzle, M. L., and Z. Wang, 1992, Seismic properties of pore fluids: Geophysics, **57**, 1396–1408, doi: 10.1190/1.1443207.

Boait, F. C., N. J. White, M. J. Bickle, R. A. Chadwick, J. A. Neufeld, and H. E. Huppert, 2012, Spatial and temporal evolution of injected CO_2 at the Sleipner field, North Sea: Journal of Geophysical Research: Solid Earth, **117**, 1–21, doi: 10.1029/2011JB008603.

Bolstad, W. M., 2009, Understanding computational Bayesian statistics: Wiley.

Bosch, M., T. Mukerji, and E. F. Gonzalez, 2010, Seismic inversion for reservoir properties combining statistical rock physics and geostatistics:

A review: Geophysics, **75**, no. 5, 75A165–75A176, doi: 10.1190/1.3478209.

Buland, A., O. Kolbjørnsen, and A. Carter, 2011, Bayesian Dix inversion: Geophysics, **76**, no. 2, R15–R22, doi: 10.1190/1.3552596.

Buland, A., O. Kolbjørnsen, R. Hauge, Ø. Skjæveland, and K. Duffaut, 2008, Bayesian lithology and fluid prediction from seismic prestack data: Geophysics, **73**, no. 3, C13–C21, doi: 10.1190/1.2842150.

Buland, A., and H. Omre, 2003, Bayesian linearized AVO inversion: Geophysics, **68**, 185–198, doi: 10.1190/1.1543206.

Cheng, B., and D. M. Titterington, 1994, Neural networks: A review from a statistical perspective: Statistical Science, **9**, 2–30, doi: 10.1214/ss/1177010638.

Doyen, P., 2007, Seismic reservoir characterization: An earth modeling perspective: EAGE.

Friedman, J. H., 1991, Multivariate adaptive regression splines: The Annals of Statistics, **19**, 1–67, doi: 10.1214/aos/1176347963.

Friedman, J. H., and W. Stuetzle, 1981, Projection pursuit regression: Journal of the American Statistical Association, **76**, 817–823, doi: 10.1080/01621459.1981.10477729.

Ghaderi, A., and M. Landrø, 2009, Estimation of thickness and velocity changes of injected carbon dioxide layers from prestack time-lapse seismic data: Geophysics, **74**, no. 2, O17–O28, doi: 10.1190/1.3054659.

Grana, D., and E. Della Rossa, 2010, Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion: Geophysics, **75**, no. 3, O21–O37, doi: 10.1190/1.3386676.

Green, P., K. Atuszyski, M. Pereyra, and C. Robert, 2015, Bayesian computation: A summary of the current state, and samples backwards and forwards: Statistics and Computing, **25**, 835–862, doi: 10.1007/s11222-015-9574-5.

Gunning, J., and M. E. Glinsky, 2007, Detection of reservoir quality using Bayesian seismic inversion: Geophysics, **72**, no. 3, R37–R49, doi: 10.1190/1.2713043.

Hammer, H., O. Kolbjørnsen, H. Tjelmeland, and A. Buland, 2012, Lithology and fluid prediction from prestack seismic data using a Bayesian model with Markov process prior: Geophysical Prospecting, **60**, 500–515, doi: 10.1111/j.1365-2478.2011.01012.x.

Hastie, T., and R. Tibshirani, 1986, Generalized additive models: Statistical Science, **1**, 297–310, doi: 10.1214/ss/1177013604.

Hauge, V. L., and O. Kolbjørnsen, 2015, Bayesian inversion of gravimetric data and assessment of CO_2 dissolution in the Utsira Formation: Interpretation, **3**, no. 2, SP1–SP10, doi: 10.1190/INT-2014-0193.1.

Henze, N., 2002, Invariant tests for multivariate normality: A critical review: Statistical Papers, **43**, 467–506, doi: 10.1007/s00362-002-0119-6.

Joe, H., 1997, Multivariate models and dependence concepts: Chapman & Hall.

Larsen, A. L., M. Ulvmoen, H. Omre, and A. Buland, 2006, Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model: Geophysics, **71**, no. 5, R69–R78, doi: 10.1190/1.2245469.

Malinverno, A., 2002, Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem: Geophysical Journal International, **151**, 675–688, doi: 10.1046/j.1365-246X.2002.01847.x.

Mavko, G., T. Mukerji, and J. Dvorkin, 2009, The rock physics handbook: Tools for seismic analysis of porous media: Cambridge University Press.

Mosegaard, K., and A. Tarantola, 1995, Monte Carlo sampling of solutions to inverse problems: Journal of Geophysical Research: Solid Earth, **100**, 12431–12447.

Mosegaard, K., and A. Tarantola, 2002, Probabilistic approach to inverse problems, *in* W. H. K. Lee, P. Jennings, C. Kisslinger, and H. Kanamori, eds., International handbook of earthquake & engineering seismology, Part 1: Academic Press, 237–265.

Mukerji, T., A. Jørstad, P. Avseth, G. Mavko, and J. R. Granli, 2001, Mapping lithofacies and pore-fluid probabilities in a North Sea reservoir: Seismic inversions and statistical rock physics: Geophysics, **66**, 988–1001, doi: 10.1190/1.1487078.

Rabben, T. E., and B. Ursin, 2011, AVA inversion of the top Utsira Sand reflection at the Sleipner field: Geophysics, **76**, no. 3, C53–C63, doi: 10.1190/1.3567951.

R Core Team, 2014, R: A language and environment for statistical computing: R Foundation for Statistical Computing.

Rimstad, K., and H. Omre, 2014a, Skew-Gaussian random fields: Spatial Statistics, **10**, 43–62, doi: 10.1016/j.spasta.2014.08.001.

Rimstad, K., and H. Omre, 2014b, Generalized Gaussian random fields using hidden selections: ArXiv e-prints.

Robert, C., and G. Casella, 2005, Monte Carlo statistical methods: Springer.

Rue, H., S. Martino, and N. Chopin, 2009, Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion): Journal of the Royal Statistical Society, Series B, **71**, 319–392, doi: 10.1111/j.1467-9868.2008.00700.x.

Silverman, B., 1986, Density estimation for statistics and data analysis: Taylor & Francis.

Tarantola, A., and B. Valette, 1982, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics and Space Physics, **20**, 219–232, doi: 10.1029/RG020i002p00219.

Ulvmoen, M., and H. Omre, 2010, Improved resolution in Bayesian lithology/fluid inversion from prestack seismic data and well observations. Part 1: Methodology: Geophysics, **75**, no. 2, R21–R35, doi: 10.1190/1.3294570.