Bayesian Approaches to Multi-Sensor Data Fusion

A dissertation submitted to the University of Cambridge

for the degree of Master of Philosophy

Olena Punska, St. John's College

August 31, 1999

Signal Processing and Communications Laboratory

Department of Engineering

University of Cambridge

Declaration

I hereby declare that my thesis is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University. I further state that no part of my thesis has already been or is being concurrently submitted for any such degree, diploma or other qualification.

I hereby declare that my thesis does not exceed the limit of the length prescribed in the Special Regulations of the M.Phil. examination for which I am a candidate. The length of my thesis is less than 14000 words.

Acknowledgments

I am most grateful to my supervisor Dr. Bill Fitzgerald for his advice, support and constant willingness to help during the past year. I am also indebted to Dr. Christophe Andrieu and Dr. Arnaud Doucet for their endless support, encouragement and kindness in answering the questions; from them, through the numerous fruitful discussions and helpful comments, I benefited immensely. My gratitude goes to Mike Hazas and, again, Dr. Christophe Andrieu and Dr. Arnaud Doucet for their companionship, useful comments and proof-reading the sections of this dissertation, and to Roger Wareham and Paul Walmsley for software and hardware support.

I am thankful to my parents for their ever present love and all kinds of support. Without the tremendous sacrifices they have made for me, I would not have had a chance to come to Cambridge. At last, but not at least, I would like to thank my husband for his tolerance and for always being near ready to help, and my daughter Anastasia for making life such great fun, for her patience and understanding.

Keywords

Multi-sensor data fusion; Bayesian inference; General linear model; Markov chain Monte Carlo methods; Model selection; Retrospective changepoint detection.

NOTATION

z	scalar
$\mathbf{z}$	column vector
z_i	ith element of $\mathbf{z}$
$\mathbf{z}_{0:n}$	vector $\mathbf{z}_{0:n} \triangleq (z_0, z_1, ..., z_n)^{\mathtt{T}}$

$\mathbf{I}_n$	identity matrix of dimension $n \times n$
$\mathbf{A}$	matrix
$\mathbf{A}^{\mathtt{T}}$	transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse of matrix $\mathbf{A}$
$\lvert\mathbf{A}\rvert$	determinant of matrix $\mathbf{A}$
$\mathbb{I}_E(\mathbf{z})$	indicator function of the set E (1 if $\mathbf{z} \in E$, 0 otherwise)
$\mathbf{z} \sim p(\mathbf{z})$	$\mathbf{z}$ is distributed according to distribution $p(\mathbf{z})$
$\mathbf{z}\rvert\, \mathbf{y} \sim p(\mathbf{z})$	the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ is $p(\mathbf{z})$

Probability distribution	$\mathcal{F}$	$f_{\mathcal{F}}(\cdot)$
Inverse Gamma	$\mathcal{IG}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp(-\beta/z)\, \mathbb{I}_{(0,+\infty)}(z)$, $\alpha > 0, \beta > 0.$
Gamma	$\mathcal{G}a(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} \exp(-\beta z)\, \mathbb{I}_{(0,+\infty)}(z)$,
Gaussian	$\mathcal{N}(\mathbf{m}, \Sigma)$	$\lvert 2\pi\boldsymbol{\Sigma}\rvert^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{m})^t \boldsymbol{\Sigma}^{-1} (\mathbf{z}-\mathbf{m})\right)$. $\alpha > 0, \beta > 0.$
Beta	$\mathcal{B}e(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1} (1-z)^{\beta-1}\, \mathbb{I}_{(0,1)}(z)$ $\alpha > 0, \beta > 0.$
Uniform	$\mathcal{U}_A$	$\left[\int_A d\mathbf{z}\right]^{-1} \mathbb{I}_A(\mathbf{z})$.
Binomial	$\mathcal{B}i(\lambda, n)$	$\binom{n}{z} \lambda^z (1-\lambda)^{n-z} \mathbb{I}_{\mathbb{N}}(z)$, $0 < \lambda < 1$, $n \in \mathbb{N}$.

Contents

1 INTRODUCTION

1.1 Overview

The field of multi-sensor data fusion is fairly young and has only recently been recognised as a separate branch of research. It has been considered from widely different perspectives by scientists of various theoretical backgrounds and interests. In fact, data fusion is a multi-disciplinary subject that draws from such areas as statistical estimation, signal processing, computer science, artificial intelligence, weapon systems, etc. The general problem arising in all these cases is one of how to combine, in the best possible manner, diverse and uncertain measurements and other information available in a multi-sensor system. The ultimate aim is to enable the system to estimate or make inference concerning a certain state of nature [42].

Traditionally the type of applications to which data fusion has been applied have been military in nature (for example, automatic target recognition). However, more recently the need for data fusion (and more generally data processing) has been recognised in many areas including remote sensing, finance, retail, automated manufacture. In the last twenty years there has been a significant increase in the number of real problems concerned with monitoring problems such as fault detection and diagnosis, safety of complex systems (aircrafts, rockets, nuclear power plants), quality control, plant monitoring and monitoring in biomedicine. These problems result from the increasing complexity of most technological processes, the availability of sophisticated sensors and the existence of sophisticated information processing systems, which are widely used. Solutions to these problems is of crucial interest for safety, ecological, and economical reasons [7].

Many practical problems arising in monitoring can be modelled with the aid of parametric models in which the parameters are subject to abrupt changes at unknown time instants. These changes are normally associated with some sort of *disorder*, which is highly undesirable and should be quickly detected with as few *false alarms* as possible. Multiple sensors are used in these systems in order to reduce uncertainty and obtain more complete knowledge of the state. Thus, the application of data fusion to changepoint detection problem is an extremely important task and this problem is addressed in the dissertation.

It is a strong belief that the issue of sensor measurements ultimately remains best handled within the framework of statistical inference. The Bayesian methodology [12] provides an elegant and consistent method of dealing with uncertainty associated with sensor measurements. However, it tends to require the evaluation of high-dimensional integrals that do not admit any closed form analytical expression. If one wants to perform Bayesian inference in these important cases, it is necessary to numerically approximate these integrals. Conventional numerical integration techniques are of limited use when the dimension of the integrand is large. An alternative approach is to use Markov chain Monte Carlo (MCMC) methods [50], which have been recently rediscovered by the Bayesian statisticians as a means to perform this integrals.

In this thesis an original algorithm for retrospective changepoint detection based on a reversible jump MCMC method [29] is proposed. It allows the estimation of the number of changepoints in the data, which can be described in terms of a general linear model, as well as the number of parameters, their values and noise variances for each segment. The main difficulty is that since both the number of changepoints and the number of parameters are assumed random, the posterior distribution to be evaluated is defined on a finite disconnected union of subspaces of various dimensions. Each subspace corresponds to a model with some fixed number of changepoints and some fixed number of model parameters. To the best of our knowledge, this joint detection/estimation problem of so called "double" model selection has never been addressed before and in the dissertation a new approach to solve it is proposed.

First, the case of one information source available (piecewise constant AR process) is considered. The proposed algorithm is applied to synthetic and real data (speech signal examined in the literature before [1], [6], [7] and [32]) and the results confirm the good performance of both the model and the algorithm when put into practice. The flexibility of the method allows the generalisation of the algorithm to the case of centralized fusion of several signals from different sources. The thesis concludes with an analysis of synthetic data "obtained" from three different sources (multiple simple steps, ramps and piecewise constant AR process), to illustrate the proposed technique. In addition, a failure of one sensor is simulated in order to demonstrate the efficiency of this approach.

1.2 Structure of the thesis

The dissertation is organised as follows.

Chapter 2 introduces the idea of using multiple sensors as a way of reducing uncertainty and obtaining more complete knowledge of the state of nature and describes the main issues which are pervasive in multi-sensor data fusion.

Chapter 3 derives a Bayesian probabilistic model of the observation process for a sensor and the subsequent inference of the state. A compact matrix formulation of a very common class of signal model, known as a general linear model is presented. Finally, the problem of combining probabilistic information from several sources is considered.

Chapter 4 is a review of Markov chain Monte Carlo methods. A few definitions concerning Markov chains are recalled, and the classical MCMC algorithms such as the Gibbs sampler, Metropolis-Hastings, Metropolis-Hastings one-at-a-time and a reversible jump MCMC methods are described.

Chapter 5 applies the methods described in chapters 2, 3 and 4 for the problem of retrospective changepoint detection. First, the problem of segmentation of piecewise constant AR processes is addressed. An original algorithm based on a reversible jump MCMC method is proposed and an extensive study of it on the synthetic and real data is carried out. The algorithm allows the estimation of the number of changepoints as well as the model orders, parameters and noise variances for each of the segments, and is then generalised for any signal which might be described in terms of a general linear model. Finally, the centralized data fusion of the signals “obtained” from three different sources (multiple simple steps, ramps and piecewise constant AR process) is considered and the case of a failure of one sensor is simulated.

2 MULTI-SENSOR DATA FUSION

One of the most fundamental problems in the history of mankind is the question of satisfying the need for knowledge concerning the external world. Human beings, as well as all living organisms, were given a special mechanism of gaining this knowledge, known as sense perception. In the information age we find ourselves in, different autonomous systems must carry out a similar function of obtaining an internal description of the external environment; and various sensing techniques are extensively employed to tackle this task.

2.1 Multi-sensor systems

Sensors are the devices used to make observations or measurements of physical quantities such as temperature, range, angle, etc. A certain relationship of mapping exists between this measured quantity and the state of nature, and thus necessary information is provided. In this regard, the interpretation of sensor measurements and sensor environment is extremely important. However, physical descriptions of sensors (sensor models) are unavoidably only approximations owning to incomplete knowledge and understanding of the environment. This, coupled with the varying degrees of uncertainty inherent in a system itself and the practical reality of occasional sensor failure, results in the lack of confidence in sensor measurements. The fact is that despite any advances in sensor technologies, no single sensor is capable of obtaining all the required information reliably, at all times, in often dynamic environments. The obvious solution in this case is to employ several sensors thus extracting as much information as possible.

In multi-sensor systems these sensors can be used to measure the same quantities, which is especially helpful in the case of sensor failure. Alternatively, different quantities associated with the same state of nature can be measured by different sensors. In addition, various sensor technologies can be employed. In all these cases the uncertainty is significantly reduced thus making the system more reliable.

2.2 A taxonomy of issues

In sensing often the problem is not one of shortage of information but rather one of how to combine the diverse and sometimes conflicting amounts of it in the best possible way.

The whole process involves different steps. First of all, the sensor model is developed.

It entails understanding of the sense environment, the nature of the measurements, the limitations of the sensor and, most importantly, probabilistic understanding of the sensor in terms of measurement uncertainty and informativeness.

Second, all the available relevant information is combined in a certain consistent and coherent manner and a single estimate of the state of the feature, given the inherent uncertainty in sensor measurements, is obtained.

Finally, if there are several sensing options or configurations, the one making the best use of sensor resources must be chosen.

Thus, irrespective of the specifics of given applications, the three main issues which are pervasive in sensor data fusion may be summarized as follows (see also Fig. 1):

- Interpretation and Representation
- Fusion, Inference and Estimation
- Sensor Management

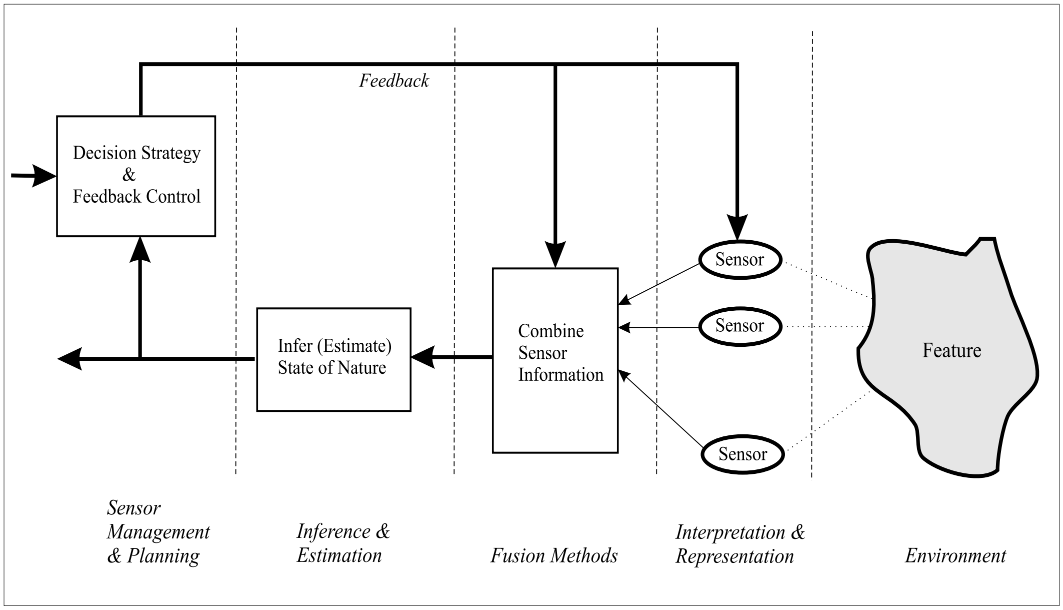
Feedback
Decision Strategy & Feedback Control
Infer (Estimate) State of Nature
Combine Sensor Information
Sensor
Sensor
Sensor
Feature
Sensor Management & Planning
Inference & Estimation
Fusion Methods
Interpretation & Representation
Environment

Figure 1: Multi-sensor data fusion.

2.3 Background and overview

The introduction to the thesis gives a flavour of the enormous variety of multi-sensor applications ranging from military systems to process plants. As was mentioned there, data fusion covers a large number of topics, from statistical estimation and signal processing to computer science and physical modelling of the sensors. Hence, not surprisingly, the main issues in the subject formulated in the previous section have mostly been addressed separately, sometimes based on well-founded theories and sometimes in an ad hoc manner and in the context of specific systems and architectures.

For instance, naturally, the problem of interpretation of measurements (sensor modelling) cannot be described in the common representation which can be made use of in a general multi-sensor system. It is only possible to say that a frequently used approach is developing probabilistic models for sensors, which were used, for example, in [21], [22], [5]. As shown in [40] and [41], the probabilistic descriptions are extremely useful in augmenting physical models thus providing a way of objectively evaluating the sensors and the information they provide using a common language.

Much work has been done in developing methods of combining information. The basic approach has been to pool the information using "weighted averaging" techniques of varying degrees of complexity [54], [9]. The Independent Opinion Pool and the Independent Likelihood Pool are described in [42], and an initial discussion of probabilistic data fusion can be found in [20].

Inferring a state of nature is, in general, a well understood problem. Such methods as Bayesian estimation [9], [46], Least Squares estimation and especially Kalman filtering [4], [34], [47], [51], [60] has been widely reported.

A major consideration which determines the form of the method employed is the multi-sensor *architecture*. They have traditionally been *centralized* but the need to relieve computational burdens at the central processor leaded to *hierarchical* systems [15], [44] which allow several levels of abstraction. However, they have some shortcomings (see [42] for the discussion) that might be overcome by the use of decentralized architectures, which are described in [23], [34], [42], [56].

The full surveys over the area are provided, for example, in [33], [42], [58] and, most recently, [28], [57].

3 A PROBABILISTIC MODEL FOR MANAGING DATA FUSION

Let us assume that the signal obtained by a sensor is carrying information relating to some physical phenomenon. In multi-sensor systems several such information sources are available and the objective is to extract this information using some suitable means, thus making the inference about the state of nature.

As a matter of fact, all these signals are corrupted by noise and, moreover, the relationship of mapping between the state and observations (the model of the signal) is never known precisely. Hence, in order to infer the true state of nature, it is necessary to find the most appropriate model to describe the obtained data, and then estimate its parameters.

The random nature of noise as well as uncertainty associated with the model can make it extremely difficult to determine what exactly is occurring. In order to develop ways to reduce the above uncertainty we turn to the methods which originate from the 18th century mathematician Reverend T. Bayes [8], [12].

3.1 Bayesian inference

3.1.1 Bayesian theorem

In *An Essay Towards Solving a Problem in the Doctrine of Chances,* Bayes creates a methodology of mathematical inference and describes how the initial information I about the hypothesis H, called a *prior* and denoted $p(H|\, I),$ and the likelihood based on observations or data D, denoted $p(D|\, H, I)$, determine the posterior probability distribution $p(H|\, D, I)$. The law that bears the author's name is, in fact, a simple relationship of conditional probabilities:

$$p(H|\, D, I) = \frac{p(D|\, H, I) p(H|\, I)}{p(D|\, I)}, \tag{1}$$

where $p(D|\, I)$ is a normalization factor, known as the "evidence".

Thus, the Bayesian posterior probability reflects our belief in the hypothesis, based on the prior information and current observations and provides a direct and easily applicable means of combining the two last-mentioned. Given this, the pervasiveness of Bayes' Theorem in data fusion problem is unsurprising.

3.1.2 Model selection and parameter estimation

The hypothesis space is different for different kinds of tasks. In general, two main problems of data analysis are model selection and parameter estimation.

In the first case, one wishes to choose from a set of candidate models $\mathcal{M}_k$ $(k = 1, \ldots, K)$ that which is best supported by the data. An auxiliary model indexing random variable k specifies which model generated the data $\mathbf{x}$, and together with the vector of model parameters $\boldsymbol{\theta}_k$ forms the "hypothesis" H that will be used in Bayes' theorem.

In the parameter estimation problem one assumes that the model is true for some unknown values of the parameters $\boldsymbol{\theta}$ and the hypothesis space is therefore the set of possible values of the parameter vector $\boldsymbol{\theta}$.

3.1.3 Assigning probabilities

Bayes' theorem tells us how to manipulate probabilities, but it does not answer the question of how to assign these probabilities, which is discussed in this section.

3.1.3.1 Likelihood function

If the assumed signal model (tested hypothesis) fits the data "exactly", the difference between the data and the inferred model is the observation noise, and hence the likelihood function should be the probability distribution of the noise. The probability distribution that has maximum entropy, subject to knowledge of the first two moments of the noise distribution is a Gaussian one, and therefore the likelihood of this form is often used unless one has further knowledge concerning the noise statistics. It also allows the integration of the nuisance parameters in many cases and has been shown to work well in practice.

3.1.3.2 Prior distribution

3.1.3.2.1 The choice of a prior

The prior distribution describes one's state of knowledge (or lack of it) about the parameter values before examining the data. The choice of it is undoubtedly the most critical and most criticized point of Bayesian analysis, since, in practice, it rarely happens that the available prior information is precise enough to lead to an exact determination of the prior distribution. The situation is especially difficult when prior information about the model is too vague or unreliable. Naturally, if a prior distribution is narrow it will dominate the posterior and can be used only to express the precise knowledge. Thus, if one has no knowledge at all about the value of a parameter prior to observing the data, the chosen prior probability function should be very broad and flat relatively to the expected likelihood function.

Non-informative priors. The most intuitively obvious non-informative prior density is a uniform density:

$$p(\boldsymbol{\theta}|\, I) = c_p, \tag{2}$$

where c_p is a constant. This prior is typically used for discrete distributions or for unbounded real valued parameters $\boldsymbol{\theta}$.

Jeffreys [36] distinguished between this case and the case of a strictly positive scale parameter and proposed the prior distribution uniformly distributed over different scales. This is the same as assuming that the logarithm of a scale parameter χ is uniformly distributed:

$$p\left(\log \chi|\, I\right) = c_p. \tag{3}$$

Using the fundamental transformation law of probabilities one obtains what is known as Jeffreys' prior:

$$p\left(\chi|\, I\right) = \frac{c_p}{\chi}. \tag{4}$$

Both uniform and Jeffreys' prior probabilities are non-normalizable and therefore improper. They can be made into a proper probability by placing bounds on the range so that the probability outside equals zero.

Conjugate priors. Another criterion for the choice of prior is its convenience. In order to simplify computation, one would prefer the prior density to be conjugate to a given likelihood function so that the posterior density takes the same form as the likelihood function. For example, in many situations the likelihood function belongs to the exponential family of probability distribution. In this case, it is convenient to use a Gaussian prior distribution for the parameters which might be positive or negative and inverse gamma distribution for scale parameters which are strictly positive.

In general, conjugate priors are not non-informative, and in order to express the ignorance of the value of the parameter, a probability density of relatively large variance can be chosen.

3.1.3.2.2 Robustness of the prior

In most cases, there is an uncertainty about the selected prior distribution used for Bayesian inference. Of course, if the precise prior information is available the prior will be better defined than in a non-informative setup, but still this information does not always lead to an exact determination of the prior distribution. It is, therefore, very important to make sure that the arbitrary part of the prior distribution does not dominate. Not surprisingly, the concern about the influence of the existent indeterminacy (*robustness* of the prior) has been reflected in a large number of works (see [10], [11], [49] and [59]), and different methods to deal with this problem have been developed.

One of the approaches used to increase robustness of the conjugate prior is Bayesian *hierarchical modelling.*

Definition 1 *A hierarchical Bayesian model is a Bayesian statistical model with the prior distribution* $p(\theta)$ *decomposed in conditional distributions* $p_1(\theta|\,\theta_1), p_2(\theta_1|\,\theta_2), \ldots, p_n(\theta_{n-1}|\,\theta_n)$ *and a marginal distribution* $p_{n+1}(\theta_n)$ *such that*

$$p\,(\theta) = \int_{\Theta_1 \times \ldots \times \Theta_n} p_1(\theta|\,\theta_1)p_2(\theta_1|\,\theta_2)\ldots p_n(\theta_{n-1}|\,\theta_n)p_{n+1}(\theta_n)d\theta_1 \ldots d\theta_n,$$

where θ *is a parameter of the Bayesian model and* θ_i *is a hyperparameter of level* i*, which belongs to a vector space* Θ_i*.*

As may be seen from the above, a hierarchical model is just a special case of a usual Bayesian model where the lack of information on the parameters of the prior distribution is expressed according to the Bayesian paradigm, i.e. through another prior distribution (*hyperprior*) on these parameters; and it seems quite intuitive that this additional level of hyperparameters in the prior modelling should robustify the prior distribution (see [49] for discussion).

3.1.3.2.3 Directed graphs

In the case of a complex system (for example, several additional levels of hyperparameters are introduced) graph theory provides a convenient way of representing the dependencies between the parameters. For instance, the following probability structure

$$p\,(u, s, x, y) = p(u)p(s|\,u)p(x|\,u, s)p(y|\,x)$$

can be visualised with a *directed acyclic graph* (DAG) (see [46]) shown in Fig. 2a. This DAG together with a set of local probability distributions associated with each variable form a *Bayesian network* (see also [35]), which is one of the examples of a *graphical model.*

Definition 2 *A graphical model is a graphical representation for probabilistic structure, along with functions that can be used to derive the joint distribution.*

Other examples of graphical models include factor graphs (see Fig. 2b), Markov random fields (see [24]) and chain graphs (see [38]).

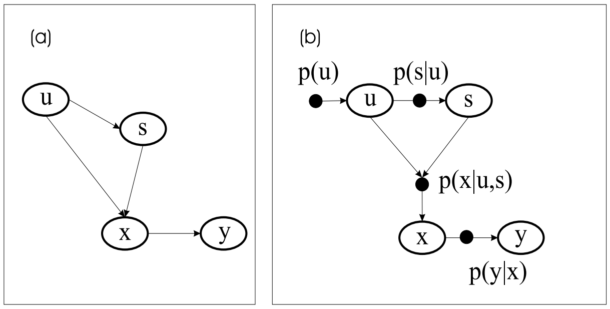

Figure 2: A directed acyclic graph (a) and a factor graph (b) for the global probability distribution $p\left(u,s,x,y\right)=p(u)p\left(\left.s\right|u\right)p\left(\left.x\right|u,s\right)p\left(\left.y\right|x\right).$

3.1.4 Bayesian inference and estimation

Once the posterior distribution is obtained it then can be used for the Bayesian estimation of the state of a system. An intuitive approach is to find the most likely values of $\mathbf{y}$ based on the information available in the form of the posterior probability distribution $p(\left.\mathbf{y}\right|\mathbf{x})$ according to some criterion. The most frequently used estimates are the following ones:

- *Maximum A Posteriori* (MAP) estimator:

$$\hat{\mathbf{y}}_{MAP}=\arg\max p(\left.\mathbf{y}\right|\mathbf{x}). \tag{5}$$

- *Minimum Mean Square Error* (MMSE) estimator

$$\hat{\mathbf{y}}_{MMSE}=\arg\min_{\hat{\mathbf{y}}}E_{p(\left.\mathbf{y}\right|\mathbf{x})}\left\{(\hat{\mathbf{y}}-\mathbf{y})(\hat{\mathbf{y}}-\mathbf{y})^{\mathrm{T}}\right\}.$$

In the same way, the evaluation of any marginal estimator is performed, though it involves extra-integration steps over the parameters that one wants to eliminate. For example, the *Marginal Maximum A Posteriori* (MMAP) estimator for the parameter y_i takes the form:

$$\hat{y}_{i\ MMAP}=\arg\max p(\left.y_i\right|\mathbf{x}). \tag{6}$$

3.1.5 The general linear model

As was mentioned before, in order to proceed with the processing of a signal it should be first described by some mathematical model, which then can be tested for a fit to the data. One of the most important signal models which may be used in a very large number of applications is the general linear model [17], [45] introduced in this section.

Let $\mathbf{x}\triangleq(x_0,x_1,\ldots,x_{T-1})^{\mathrm{T}}$ be a vector of T observations. Our prior information suggests modelling the data by a set of p model parameters or linear coefficients, arranged in the vector $\mathbf{a}=(a_1, a_2, \ldots, a_p)$. We describe the data as a linear combination of basis functions with an additive noise component. Our model thus has the form

$$x_m = \sum_{j=1}^{p} a_j g_j(t) + n_m, \qquad \text{if } 0 \leq m \leq T-1,$$

where $g_j(t)$ is a value of a basis function.

This can be written in the form of a matrix equation

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{n}, \tag{7}$$

where $\mathbf{X}$ is the $T \times p$ dimensional matrix of basis functions that determine the type of the model (for example, AR model) and $\mathbf{n}$ is a vector of noise samples. More precisely,

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T-1} \end{bmatrix} = \begin{bmatrix} g_1(0) & g_2(0) & \ldots & g_p(0) \\ g_1(1) & g_2(1) & \ldots & g_p(1) \\ \vdots & \vdots & \ddots & \vdots \\ g_1(T-1) & g_2(T-1) & \ldots & g_p(T-1) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{T-1} \end{bmatrix}. \tag{8}$$

The strength of the general linear model is its flexibility, which is explored below for several possible sets of basis functions.

3.1.5.1 Common basis functions

This section explains how to formulate the matrix $\mathbf{X}$ for several particular types of models, such as an autoregressive model (AR), autoregressive model with exogenous input model (ARX) and polynomial model.

Example 1 ***Autoregressive (AR) model.*** *An AR model is a time series where a given datum is a weighted sum of the p previous data and noise term. Equivalently, an AR model is an output of an all-pole filter excited by white noise. More precisely,*

$$x_m = \sum_{j=1}^{p} a_j x_{m-j} + n_m \qquad \textit{for } 0 \leq m < T-1, \tag{9}$$

which is in the matrix form is given by

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T-1} \end{bmatrix} = \begin{bmatrix} x_{-1} & x_{-2} & \dots & x_{-p} \\ x_0 & x_{-1} & \dots & x_{1-p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-2} & x_{T-3} & \dots & x_{T-1-p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{T-1} \end{bmatrix}. \qquad (10)$$

One difficulty with implementation exists because of the need to have initial conditions for the filter or knowledge of x_{-1} through x_{-p}. Prior information may suggest reasonable assumptions for these values. Alternatively, one can interpret the first p samples as the initial conditions and proceed with the analysis on the remaining $T - p$ data points (see [27])..

Example 2 ***Autoregressive model with exogenous input (ARX).*** *Whereas an AR model is the output of an all-pole filter excited by white noise, an ARX model is a filtered version of some input u with this filter having both pole and zeroes. Mathematically, an ARX model is*

$$x_m = \sum_{j=1}^{q} \alpha_j x_{m-j} + \sum_{j=0}^{z} \beta_j u_{m-j} + n_m \qquad \text{for } 0 \le m < T-1, \qquad (11)$$

and the matrix **X** *takes the form*

$$\mathbf{X} = \begin{bmatrix} x_{-1} & x_{-2} & \dots & x_{-q} & u_0 & u_{-1} & \dots & u_{-z} \\ x_0 & x_{-1} & \dots & x_{1-q} & u_1 & u_0 & \dots & u_{1-z} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{T-2} & x_{T-3} & \dots & x_{T-1-q} & u_{T-1} & u_{T-2} & \dots & u_{T-1-z} \end{bmatrix}, \qquad (12)$$

with a vector of parameters $\mathbf{a} = (\alpha_1, \alpha_2, \dots, \alpha_q, \beta_0, \beta_1, \dots, \beta_z)^{\mathrm{T}}$ of the length $p = q + z + 1$.

Example 3 ***Polynomial and seemingly non-linear models.*** *The flexibility of a general linear model allows us to describe polynomial and other models where the basis functions are not linear, but the models are linear in its coefficients. In the case of the polynomial model, the observation sequence is given by*

$$x_m = \sum_{j=1}^{p} a_j u_m^{j-1} + n_m \qquad \text{for } 0 \le m < T-1 \qquad (13)$$

which is in the generalized form can be rewritten as

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T-1} \end{bmatrix} = \begin{bmatrix} 1 & u_0 & u_0^2 & \dots & u_0^{p-1} \\ 1 & u_1 & u_1^2 & \dots & u_1^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{T-1} & u_{T-1}^2 & \dots & u_{T-1}^{p-1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} n_0 \\ n_1 \\ \vdots \\ n_{T-1} \end{bmatrix}$$

3.1.5.2 Marginalization of the nuisance parameters

One of the most interesting features of the Bayesian paradigm is the ability to remove nuisance parameters (i.e. parameters that are not of interest) from the analysis. This process is of both practical and theoretical interest, since it can significantly reduce the dimension of the problem being addressed.

Suppose the observed data $\mathbf{x} = (x_1, x_2, \dots, x_T)^{\mathrm{T}}$ may be described in terms of a general linear model (we repeat Eq. (7) for convenience):

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{n},$$

where $\mathbf{n}$ is a vector of i.i.d. Gaussian noise samples. Then the likelihood function is given by

$$p(\mathbf{x}|\left\{\omega\right\},\sigma,\mathbf{a}) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left[-\frac{\mathbf{n}^{\mathrm{T}}\mathbf{n}}{2\sigma^2}\right], \qquad (14)$$

where $\{\omega\}$ denotes the parameters of the basis functions $\mathbf{X}$. Substituting Eq. (7) into Eq. (14) gives

$$p(\mathbf{x}|\left\{\omega\right\},\sigma,\mathbf{a}) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left[-\frac{(\mathbf{x}-\mathbf{X}\mathbf{a})^{\mathrm{T}}(\mathbf{x}-\mathbf{X}\mathbf{a})}{2\sigma^2}\right]. \qquad (15)$$

Remark 1 *In fact, the exact likelihood expression for the case of AR and ARX modelling is of a slightly different form (see [14], [27]).*

Suppose, that a given series is generated by the pth-order stationary autoregressive model, which in an alternative form is given by:

$$\mathbf{n}_{p:T-1} = \mathbf{A}\mathbf{x}_{p:T-1},$$

where $\mathbf{A}$ *is the* $((T-p)\times(T))$ *matrix:*

$$\mathbf{A}=\begin{bmatrix} -a_p & \dots & -a_1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -a_p & \dots & -a_1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & -a_p & \dots & -a_1 & 1 \end{bmatrix}.$$

Here the first p samples are interpreted as the initial conditions and $\mathbf{n}_{p:T-1}$ *is a vector of i.i.d. Gaussian noise samples. Thus, one obtains:*

$$p(\mathbf{n}_{p:T-1})=(2\pi\sigma^2)^{-\frac{T-p}{2}}\exp(-\frac{1}{2\sigma^2}\mathbf{x}_{p:T-1}^{\mathtt{T}}\mathbf{A}^{\mathtt{T}}\mathbf{A}\mathbf{x}_{p:T-1})$$

Since the Jacobian of the transformation between $\mathbf{n}_{p:T-1}$*,* $\mathbf{x}_{p:T-1}$ *is unity and the conditional likelihood is equal to:*

$$p(\mathbf{x}_{p:T-1}|\,\mathbf{x}_{0:p-1},\mathbf{a})=(2\pi\sigma^2)^{-\frac{T-p}{2}}\exp(-\frac{1}{2\sigma^2}\mathbf{x}_{p:T-1}^{\mathtt{T}}\mathbf{A}^{\mathtt{T}}\mathbf{A}\mathbf{x}_{p:T-1}).$$

and in order to obtain the true likelihood for the whole data block, the probability chain rule can be used:

$$p(\mathbf{x}_{0:T-1}|\,\mathbf{a})=p(\{\mathbf{x}_{0:p-1},\mathbf{x}_{p:T-1}\}|\,\mathbf{a})=p(\mathbf{x}_{p:T-1}|\,\mathbf{x}_{0:p-1},\mathbf{a})p(\mathbf{x}_{0:p-1}|\,\mathbf{a}),$$

where

$$p(\mathbf{x}_{0:p-1}|\,\mathbf{a})=(2\pi\sigma^2)^{-\frac{p}{2}}\left|\mathbf{M}_{\mathbf{x}_{0:p-1}}\right|^{-\frac{1}{2}}\exp(-\frac{1}{2\sigma^2}\mathbf{x}_{0:p-1}^{\mathtt{T}}\mathbf{M}_{\mathbf{x}_{0:p-1}}^{-1}\mathbf{x}_{0:p-1})$$

and $\mathbf{M}_{\mathbf{x}_{0:p-1}}$ *is the covariance matrix for p samples of data with unit variance excitation.*

The exact likelihood expression is thus:

$$p(\mathbf{x}_{0:T-1}|\,\mathbf{a})=(2\pi\sigma^2)^{-\frac{T}{2}}\left|\mathbf{M}_{\mathbf{x}_{0:p-1}}\right|^{-\frac{1}{2}}\exp(-\frac{1}{2\sigma^2}\mathbf{x}_{0:T-1}^{\mathtt{T}}\mathbf{M}_{\mathbf{x}_{0:T-1}}^{-1}\mathbf{x}_{0:T-1}),$$

where

$$\mathbf{M}_{\mathbf{x}_{0:T-1}}^{-1}=\mathbf{A}^{\mathtt{T}}\mathbf{A}+\begin{bmatrix}\mathbf{M}_{\mathbf{x}_{0:p-1}}^{-1} & 0 \\ 0 & 0\end{bmatrix}$$

is the inverse covariance matrix for a block of T samples.

However, in many cases T will be large and the term $\mathbf{x}_{0:p-1}^{\mathtt{T}}\mathbf{M}_{\mathbf{x}_{0:p-1}}^{-1}\mathbf{x}_{0:p-1}$ *can be regarded as an insignificant "end-effect". In this case we make the approximation* $\mathbf{x}_{0:T-1}^{\mathtt{T}}\mathbf{M}_{\mathbf{x}_{0:T-1}}^{-1}\mathbf{x}_{0:T-1}\approx\mathbf{x}_{0:T-1}^{\mathtt{T}}\mathbf{A}^{\mathtt{T}}\mathbf{A}\mathbf{x}_{0:T-1}$ *and obtain the approximate likelihood of the* *form:*

$$p(\mathbf{x}|\left\{\omega\right\},\sigma,\mathbf{a}) \propto (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left[-\frac{\mathbf{n}^{\mathrm{T}}\mathbf{n}}{2\sigma^2}\right], \tag{16}$$

Similarly, the approximate likelihood for the case of ARX modelling is obtained.

We assume uniform priors over each of the elements of the vector **a** and assign a Jeffreys' prior to σ. In fact, these parameters are not of interest in our task so they can be easily integrated out. Using the following standard integral identity [45]:

$$\int_{\mathbb{R}^p} \exp\left[-\frac{\mathbf{a}^{\mathrm{T}}\mathbf{A}\mathbf{a}+\mathbf{y}^{\mathrm{T}}\mathbf{a}+c}{2\sigma^2}\right]d\mathbf{y} = (2\pi\sigma^2)^{\frac{p}{2}}\left|\mathbf{A}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(c-\frac{\mathbf{a}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{a}}{4}\right)\right], \tag{17}$$

and a gamma integral

$$\int_0^\infty \sigma^{\alpha-1}\exp\left(-Q\sigma\right)d\sigma = \Gamma(\alpha)Q^{-\alpha}, \tag{18}$$

one obtains

$$\begin{aligned} p(\left\{\omega\right\}|\,\mathbf{x}) &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^+} p(\left\{\omega\right\},\sigma,\mathbf{a}|\,\mathbf{x})d\mathbf{a}d\sigma \qquad (19)\\ &\propto \left|\mathbf{X}^{\mathrm{T}}\mathbf{X}\right|^{-\frac{1}{2}}\left[\mathbf{x}^{\mathrm{T}}\mathbf{x}-\mathbf{x}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{x}\right]^{-\frac{T-p}{2}}. \end{aligned}$$

Here the integrals have been done analytically so the dimensionality of the parameter space was reduced for each parameter integrated out. This reduction of the dimensionality is a major advantage in many applications.

3.2 Combining probabilistic information

The techniques presented thus far are, in general, well understood in terms of classical statistical theory. However, when there is a multiplicity of informational sources, a problem of combining information from them arises. In this section we consider in turn three approaches generally proposed in the literature and discuss some criticisms associated with them.

To begin with, we assume that M information sources are available and the observations from the m^{th} source are arranged in the vector $\mathbf{x}^{(m)}$ (the number of observations T is the same for all sources). What is now required is to compute the global posterior distribution $p\left(\mathbf{y}|\,\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right)$, given the information contributed by each source. In what follows, we will assume that each information source communicates either a local posterior distribution $p\left(\mathbf{y}|\,\mathbf{x}^{(m)}\right)$or a likelihood function $p\left(\mathbf{x}^{(m)}|\,\mathbf{y}\right)$.

3.2.1 Linear opinion pool

In tackling the problem of fusion, the information originating from different sources, the questions of how relevant and how reliable is the information from each source should be considered. These questions can be addressed by attaching a measure of value such as weight to the information provided by each source. Such a pool based on the probabilistic representation of the information was proposed by Stone [54]. The posteriors from each information source are combined linearly (see Fig. 3), i.e.

$$p\left(\mathbf{y}|\,\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right)=\sum_{m=1}^{M}\omega_m p\left(\mathbf{y}|\,\mathbf{x}^{(m)}\right), \tag{20}$$

where ω_m is a weight such that, $0 \leq \omega_m \leq 1$ and $\sum_{m=1}^{M}\omega_m = 1$. The weight ω_m reflects the significance attached to the m^{th} information source. It can be used to model the reliability or trustworthiness of an information source and to "weight out" faulty sensors.

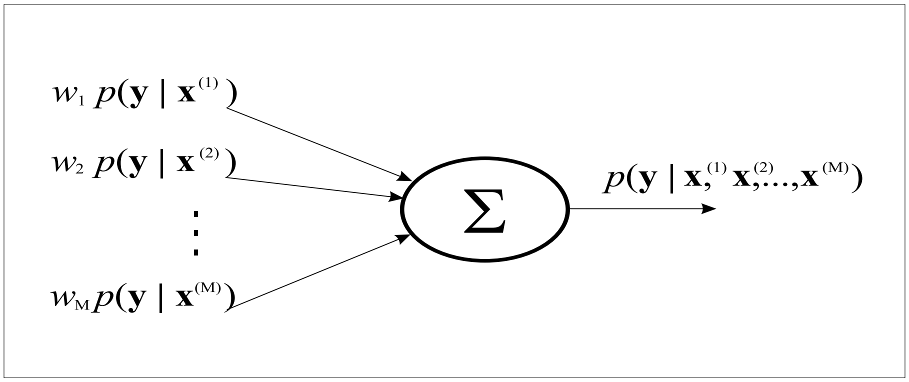
$w_1\,p(\mathbf{y}\mid\mathbf{x}^{(1)})$
$w_2\,p(\mathbf{y}\mid\mathbf{x}^{(2)})$
$w_M p(\mathbf{y}\mid\mathbf{x}^{(M)})$
Σ
$p(\mathbf{y}\mid\mathbf{x},^{(1)}\mathbf{x},^{(2)}\ldots,\mathbf{x}^{(M)})$

Figure 3: Linear Opinion Pool.

However, in the case of equal weights, the Linear Opinion Pool can give an erroneous result if one sensor is dissenting even if M is relatively large. This is because the Linear Opinion Pool gives undue credence to the opinion of the m^{th} source. The need to redress this leads to the second approach.

3.2.2 Independent opinion pool

In the Independent Opinion Pool [42] it is assumed that the information obtained conditioned on the observation set is independent. More precisely, the Independent Opinion Pool is defined by the product

$$p\left(\mathbf{y}|\,\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right)\propto\prod_{m=1}^{M}p\left(\mathbf{y}|\,\mathbf{x}^{(m)}\right), \tag{21}$$

which is illustrated in Fig. 4.

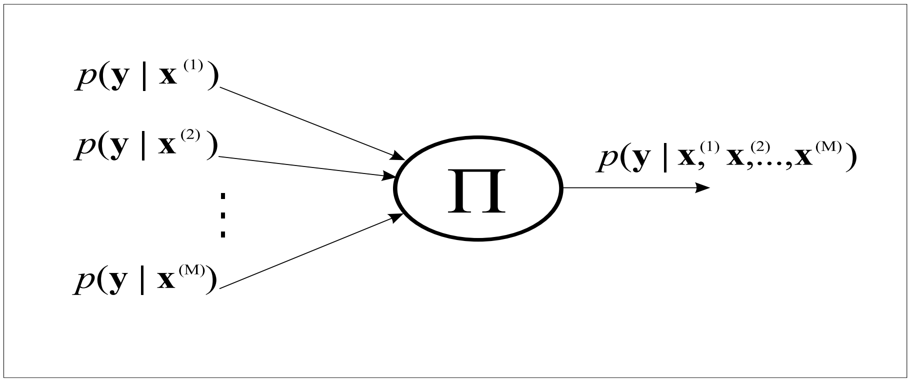

Figure 4: Independent Opinion Pool.

In general, this is a difficult condition to satisfy, though in the realm of measurement the conditional independence can often be justified experimentally.

A more serious problem is that the Independent Opinion Pool is extreme in its reinforcement of opinion when the prior information at each node is common, i.e. obtained from the same source. Indeed, the global posterior can be rewritten as

$$p\left(\mathbf{y}\middle|\,\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right) \propto \frac{p\left(\mathbf{x}^{(1)}\middle|\,\mathbf{y}\right)p_1\left(\mathbf{y}\right)}{p\left(\mathbf{x}^{(1)}\right)} \times \frac{p\left(\mathbf{x}^{(2)}\middle|\,\mathbf{y}\right)p_2\left(\mathbf{y}\right)}{p\left(\mathbf{x}^{(2)}\right)} \times \ldots \times \frac{p\left(\mathbf{x}^{(M)}\middle|\,\mathbf{y}\right)p_M\left(\mathbf{y}\right)}{p\left(\mathbf{x}^{(M)}\right)}, \tag{22}$$

and if the prior information is obtained from the same source, then

$$p_1\left(\mathbf{y}\right) = p_2\left(\mathbf{y}\right) = \ldots = p_M\left(\mathbf{y}\right), \tag{23}$$

which results in unwarranted reinforcement of the posterior through the product of the priors $\prod_{m=1}^{M} p_m\left(\mathbf{y}\right)$. Thus the Independent Opinion Pool is only appropriate when the priors are obtained independently on the basis of subjective prior information at each information source.

3.2.3 Independent likelihood pool

When each information source has common prior information, i.e. information obtained from the same origin, the situation is better described by the Independent Likelihood Pool [42], which is derived as follows. According to Bayes' theorem for the global posterior one obtains

$$p\left(\mathbf{y}|\,\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right)\propto\frac{p\left(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\middle|\,\mathbf{y}\right)p\left(\mathbf{y}\right)}{p\left(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right)}. \tag{24}$$

For a sensor system is reasonable to assume that the likelihoods from each informational source $p\left(\mathbf{x}^{(m)}\middle|\,\mathbf{y}\right)$, $m=1,\ldots,M$, are independent since the only parameter they have in common is the state.

$$p\left(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\middle|\,\mathbf{y}\right)=p\left(\mathbf{x}^{(1)}\middle|\,\mathbf{y}\right)p\left(\mathbf{x}^{(2)}\middle|\,\mathbf{y}\right)\ldots p\left(\mathbf{x}^{(M)}\middle|\,\mathbf{y}\right). \tag{25}$$

Thus, the Independent Likelihood Pool is defined by the following equation

$$p\left(\mathbf{y}|\,\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(M)}\right)\propto p\left(\mathbf{y}\right)\prod_{m=1}^{M}p\left(\mathbf{x}^{(m)}\middle|\,\mathbf{y}\right), \tag{26}$$

and is illustrated in Fig. 5.

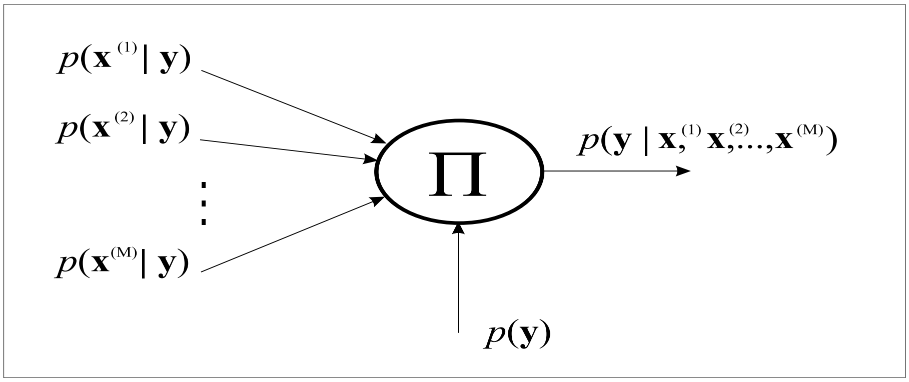
$p(\mathbf{x}^{(1)}|\,\mathbf{y})$
$p(\mathbf{x}^{(2)}|\,\mathbf{y})$
$p(\mathbf{x}^{(M)}|\,\mathbf{y})$
$\prod$
$p(\mathbf{y}\,|\,\mathbf{x}^{(1)}_{,}\mathbf{x}^{(2)}_{,}...,\mathbf{x}^{(M)})$
$p(\mathbf{y})$

Figure 5: Independent Likelihood Pool.

3.2.4 Remarks

As may be seen from the above both the Independent Opinion Pool and the Independent Likelihood Pool more accurately describe the situation in multi-sensor systems where the conditional distribution of the observation can be shown to be independent. However, in most cases in sensing the Independent Likelihood Pool is the most appropriate way of combining information since the prior information tends to be from the same origin. If there are dependencies between information sources the Linear Opinion Pool should be used.

4 MCMC METHODS

As shown in the previous chapter, the Bayesian approach typically requires the evaluation of high-dimensional integrals involving posterior (or marginal posterior) distributions that do not admit any closed form analytical expression. In order to perform Bayesian inference it is necessary to numerically approximate these integrals. However, classical numerical integration methods are difficult to use when the dimension of the integrand is large and impose a huge computational burden. An attractive approach to solving this problem consists of using Markov chain Monte Carlo (MCMC) methods - powerful stochastic algorithms that have revolutionized applied statistics; see [13], [50], [55] for some reviews.

4.1 Markov chains

The basic idea of MCMC methods is to simulate an ergodic Markov chain whose samples are asymptotically distributed according to some *target* probability distribution known up to a normalising constant $\pi(d\mathbf{x}) = \pi(\mathbf{x})d\mathbf{x}$.

Definition 3 *A Markov chain [2], [55] is a sequence of random variables* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ *defined in the same space* $(E, \mathcal{E})$ *such that the influence of random variables* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i$ *on the value of the* $\mathbf{x}_{i+1}$ *is mediated by the value of* $\mathbf{x}_i$ *alone, i.e. for any* $A \in \mathcal{E}$

$$\Pr(\mathbf{x}_{i+1} \in A | \, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i) = \Pr(\mathbf{x}_{i+1} \in A | \, \mathbf{x}_i).$$

One can define for any $(\mathbf{x}, A) \in E \times \mathcal{E}$:

$$P(\mathbf{x}, A) \triangleq \Pr(\mathbf{x}_{i+1} \in A | \, \mathbf{x}_i = \mathbf{x}), \tag{27}$$

where $P(\mathbf{x}, A)$ is the *transition kernel* of the Markov chain and

$$P(\mathbf{x}, A) = \int_A P(\mathbf{x}, d\mathbf{x}'), \tag{28}$$

where $P(\mathbf{x}, d\mathbf{x}')$ is the probability of going to a "small" set $d\mathbf{x}' \in \mathcal{E}$, starting from $\mathbf{x}$.

There are two properties required of the Markov chain for it to be of any use in sampling a prescribed density: there must exist a unique invariant distribution and the Markov chain must be ergodic.

Definition 4 *A probability distribution $\pi(d\mathbf{x})$ is an invariant or stationary distribution for the transition kernel P if for any*

$$\pi(A) = \int_E \pi(d\mathbf{x}) P(\mathbf{x}, A) = \int_E \pi(d\mathbf{x}) \int_A P(\mathbf{x}, d\mathbf{x}').$$

This implies that if a state of the Markov chain $\mathbf{x}_i$ is distributed according to $\pi(d\mathbf{x})$ then $\mathbf{x}_{i+1}$ and all the following states are distributed marginally according to $\pi(d\mathbf{x})$; and therefore it is important to ensure that π is the invariant distribution of the Markov chain.

Definition 5 *A transition kernel P is π-reversible [2], [55] if it satisfies for any $(A, B) \in \mathcal{E} \times \mathcal{E}$:*

$$\int_A \pi(d\mathbf{x}) P(\mathbf{x}, B) = \int_B \pi(d\mathbf{x}) P(\mathbf{x}, A).$$

Stated in words, the probability of a transition from A to B is equal to the probability of a transition in the reverse direction. It is easy to show that this condition of *detailed balance* implies invariance and, therefore, is very often used in the framework of the MCMC algorithms.

We also require that the Markov chain be *ergodic.*

Definition 6 *A Markov chain is said to be ergodic [43] if, regardless of the initial distribution, the probabilities at time N converge to the invariant distribution as $N \to \infty$.*

Of course, the rate of convergence of a Markov chain or, indeed, whether it converges at all is of crucial interest. This question is well developed and presented by many authors such as Meyn and Tweedie [39], Neal [43] and Tierney [55].

4.2 MCMC algorithms

In the following subsections some classical methods for constructing a Markov chain that admits as invariant distribution $\pi(d\mathbf{x}) = \pi(\mathbf{x}) d\mathbf{x}$ are presented (see also [2], [50]).

4.2.1 Gibbs sampler

The Gibbs sampler was first introduced in image processing by Geman and Geman [25]. The algorithm proceeds as follows

Gibbs sampling

1. Set randomly $\mathbf{x}^{(0)} = \mathbf{x}_0$.
2. Iteration i, $i \geq 1$.
   - Sample $x_1^{(i)} \sim \pi(x_1 | \mathbf{x}_{-1}^{(i)})$.
   - Sample $x_2^{(i)} \sim \pi(x_2 | \mathbf{x}_{-2}^{(i)})$.

     $\vdots$
3. Goto 2.

■

where $\mathbf{x}_{-k}^{(i)} \triangleq (x_1^{(i)}, x_2^{(i)}, \ldots, x_{k-1}^{(i)}, x_{k+1}^{(i-1)}, \ldots)$ and $\pi(x_k | \mathbf{x}_{-k}^{(i)})$ is the full conditional density with all components but one x_k held constant.

4.2.2 Metropolis-Hastings algorithm

Another very popular MCMC algorithm is the Metropolis-Hastings (MH) algorithm, which uses a candidate *proposal* distribution $q(\mathbf{x} | \mathbf{x}^{(i)})$.

Metropolis-Hastings algorithm

1. Set randomly $\mathbf{x}^{(0)} = \mathbf{x}_0$.
2. Iteration i, $i \geq 1$.
   - Sample a candidate $\mathbf{x} \sim q_1(\mathbf{x} | \mathbf{x}^{(i-1)})$.
   - Evaluate the acceptance probability

$$\alpha(\mathbf{x}^{(i-1)}, \mathbf{x}) = \min\left\{1, \frac{\pi\left(\mathbf{x}\right)}{\pi\left(\mathbf{x}^{(i-1)}\right)} \frac{q\left(\mathbf{x}^{(i-1)} \middle| \mathbf{x}\right)}{q\left(\mathbf{x} \middle| \mathbf{x}^{(i-1)}\right)}\right\}.$$

   - Sample $u \sim \mathcal{U}_{(0,1)}$. If $u \leq \alpha(\mathbf{x}^{(i-1)}, \mathbf{x})$ then $\mathbf{x}^{(i)} = \mathbf{x}$ otherwise $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)}$.
3. Goto 2.

One may want to select the candidate independently of the current state according to a distribution $q(\mathbf{x}|\,\mathbf{x}^{(i)}) = \varphi(\mathbf{x})$ in which case the acceptance probability is given by

$$\alpha(\mathbf{x}^{(i-1)}, \mathbf{x}) = \min\left\{1, \frac{\pi\left(\mathbf{x}\right)}{\pi\left(\mathbf{x}^{(i-1)}\right)} \frac{\varphi\left(\mathbf{x}^{(i-1)}\right)}{\varphi\left(\mathbf{x}\right)}\right\}. \tag{29}$$

It is worth noticing that the algorithm does not require knowledge of the normalising constant of $\pi\left(d\mathbf{x}\right)$ as only the ratio $\dfrac{\pi\left(\mathbf{x}\right)}{\pi\left(\mathbf{x}^{(i-1)}\right)}$ appears in the acceptance probability.

Metropolis-Hastings one-at-a-time. In the case where $\mathbf{x}$ is high-dimensional it is very difficult to select a good proposal distribution so that the level of rejections will be low. To solve this problem one can modify the method and update only one parameter at a time similar to the Gibbs sampling algorithm. More precisely,

Metropolis-Hastings one-at-a-time

1. Set randomly $\mathbf{x}^{(0)} = \mathbf{x}_0$.
2. Iteration i, $i \geq 1$.
    - Sample a candidate $x_1^{(i)}$ according to MH step with proposal distribution $q_1(x_1|\,\mathbf{x}_{-1}^{(i-1)})$ and invariant distribution $\pi(x_1|\,\mathbf{x}_{-1}^{(i-1)})$.
    - Sample a candidate $x_2^{(i)}$ according to MH step with proposal distribution $q_2(x_2|\,\mathbf{x}_{-2}^{(i-1)})$ and invariant distribution $\pi(x_2|\,\mathbf{x}_{-2}^{(i-1)})$.

    $\vdots$

    - Sample a candidate $x_k^{(i)}$ according to MH step with proposal distribution $q_k(x_k|\,\mathbf{x}_{-k}^{(i-1)})$ and invariant distribution $\pi(x_k|\,\mathbf{x}_{-k}^{(i-1)})$.

    $\vdots$

3. Goto 2.

where $\mathbf{x}_{-k}^{(i)} \triangleq (x_1^{(i)}, x_2^{(i)}, \ldots, x_{k-1}^{(i)}, x_{k+1}^{(i-1)}, \ldots)$. As might be seen from the above this algorithm includes the Gibbs sampler as a special case when the proposal distributions of the MH steps are equal to the full conditional distributions, so that the acceptance probability is equal to 1 and no candidate is rejected.

4.2.3 Reversible jump MCMC

Such an important area of signal processing as model uncertainty problem can be treated very elegantly through the use of MCMC methods, reversible jump MCMC [29] in particular. In fact, this method might be viewed as a direct generalisation of the Metropolis-Hastings method. In the case of model selection the problem is that the posterior distribution to be evaluated is defined on a finite disconnected union of subspaces of various dimensions, corresponding to different models. The reversible jump sampler achieves such model space moves by Metropolis-Hastings proposals with an acceptance probability which is designed to preserve detailed balance (reversibility) within each move type. If a move from model k with parameters $\boldsymbol{\theta}_k$ to the model k' with parameters $\boldsymbol{\theta}_{k'}$ is proposed then such an acceptance probability is given by

$$\alpha = \min \left\{ 1, \frac{\pi\left(k', \boldsymbol{\theta}_{k'}\right)}{\pi\left(k, \boldsymbol{\theta}_k\right)} \frac{q\left(k, \boldsymbol{\theta}_k \middle| k', \boldsymbol{\theta}_{k'}\right)}{q\left(k', \boldsymbol{\theta}_{k'} \middle| k, \boldsymbol{\theta}_k\right)} \right\}. \tag{30}$$

In the above equation it is assumed that the proposal is made directly in the new parameter space rather than via “dimensional” matching random variables (see [29]) and the Jacobian term is therefore equal to 1.

5 APPLICATION TO CHANGEPOINT DETECTION

5.1 Introduction

The theory of changepoint detection has its origins in segmentation - a problem which is fundamental to many areas of data and image analysis. The process involves dividing a large sequence of data into small homogeneous segments, the boundaries of which may be interpreted as changes in the physical system. This approach has proved extremely useful for different practical problems arising in recognition-oriented signal processing, such as continuous speech processing, biomedical and seismic signal processing, monitoring of industrial processes, etc. Not surprisingly, the task is of great practical and theoretical interest, which is reflected in a large number of surveys. For example, the problem of automatic analysis of continuous speech signals is addressed in [1]; segmentation algorithms for recognition-oriented geophysical signals are described in [3]; and an application of the changepoint detection method to an electroencephalogram (EEG) is presented in [37].

Of course, different authors propose various approaches to the problem of detection of abrupt changes and, in particular, segmentation. This issue is thoroughly surveyed in [7], where different methods are proposed and an exhaustive list of references is given. Since then, several contributions have been made to the field of changepoint theory. For example, the General Piecewise Linear Model and its extension to study multiple changepoints in non-Gaussian impulsive noise environments is introduced in [45], segmentation in a linear regression framework is investigated in [30] and [32], and a general segmentation method suitable for both parametric and nonparametric models is described in [37]. The main goal of these last approaches and, indeed, [18] is the use of the maximum a posteriori (MAP), or maximum-likelihood (ML), estimate. According to [31], this technique eliminates some shortcomings of the Generalised Likelihood Ratio (GLR) test (see [31], [32] for discussion), introduced in [61] and widely used in segmentation in the 1980s (see [1], [6], [7]). Some approaches to solve the problem of multiple changepoint detection in a Bayesian framework, using Markov Chain Monte Carlo (MCMC) [50], are also presented in [3] and [53].

In [1], [7], [37] it is also shown that the algorithms designed for signals modelled as piecewise constant autoregressive (AR) processes excited by white Gaussian noise, have proved useful for processing real signals, such as speech, seismic and EEG data. In all these cases the order of AR model was the same for different segments and was chosen by the user. However, in practice, there are numerous applications (speech processing, for example) where different model orders should be considered for different segments. Thus, not only the number of segments, but the correct model orders for each of them should be estimated. To the best of our knowledge, this joint detection/estimation problem has never been addressed before and in this paper a new methodology to solve it is proposed.

In this chapter the problem of retrospective changepoint detection is considered; thus all the data are assumed to be available at a same time. The chapter begins by examining the observations from a single source, and the segmentation of piecewise constant AR processes in particular. Following a Bayesian approach, the unknown parameters, including the number of AR processes needed to represent the data, the model orders, the values of the parameters and noise variances for each segment are regarded as random quantities with known prior distributions. Moreover, some of the hyperparameters are considered random as well and drawn from the appropriate hyperprior distribution, whereas they are usually tuned heuristically by the user (see [32], [37]). The main problem of this approach is that the resulting posterior distribution appears highly non-linear in its parameters, thus precluding analytical calculations. The case treated here is even more complex. Indeed, since the number of changepoints and the orders of the models are assumed random, the posterior distribution is defined on a finite disconnected union of subspaces of various dimensions. Each subspace corresponds to a model with a fixed number of changepoints and fixed model order for each segment. To evaluate this joint posterior distribution, an efficient stochastic algorithm based on reversible jump Markov chain Monte Carlo (MCMC) methods [50], [29] is proposed. Once the posterior distribution, and more specifically some of its features such as marginal distributions, are estimated, model selection can be performed using the marginal maximum a-posteriori (MMAP) criterion. The proposed algorithm is applied to synthetic and real data (a speech signal examined in the literature before, see [1], [6], [7], and [32]) and the results confirm the good performance of both the model and the algorithm when put into practice.

Then in Subsection 5.2.2 the framework for identification of multiple changepoints in linearly modelled data, where the noise corrupting the signal is i.i.d. Gaussian, is presented. This approach is just a generalization of the method proposed before for the segmentation of piecewise constant AR processes, and the strength of it is its flexibility with one algorithm for multiple simple steps, ramps, autoregressive changepoints, polynomial coefficient changepoints and changepoints in other piecewise linear models.

Finally, the problem of the centralized fusion of the information originating from a number of different sources, as a way of reducing uncertainty and obtaining more complete knowledge of changes in the state of nature, is addressed. Practical applications of this technique abound in diverse areas, and one of the examples is monitoring changes in a reservoir in oil production, described in Section 5.3 in more detail. In the method introduced here, all available signals are assumed to be in the form of the general linear piecewise model, and the probabilistic information is combined according to the Independent Likelihood Pool. The developed algorithm is applied to the synthetic data obtained from three different sources (multiple simple steps, ramps and a piecewise constant AR process) and, in addition, the case of the failure of one sensor is simulated.

5.2 Single information source

In this section the segmentation of the signals obtained from a single information source is considered. First, the case of piecewise constant AR models is presented (see Subsection 5.2.1) and then the application of the proposed method to any signal which might be represented in the form of the general linear model is discussed (see Subsection 5.2.2).

5.2.1 Segmentation of piecewise constant AR processes

This section specifically develops the method for segmentation of piecewise constant AR processes excited by white Gaussian noise and is organised as follows: the model of the signal is given in Subsection 5.2.1.1; in Subsection 5.2.1.2, we propose a hierarchical Bayesian model and state the estimation objectives. As mentioned above, this model implies that the posterior distribution and the associated Bayesian estimators do not admit any closed-form expression. Therefore, in order to perform estimation, an algorithm based on a reversible jump MCMC algorithm (see [29]), is developed in Subsection 5.2.1.3. The results for both synthetic and real data (speech signal examined in the literature before, see [1], [6], [7], and [32]) are presented in Subsection 5.2.1.4 and confirm the good performance of both the model and the algorithm when put into practice.

5.2.1.1 Problem Statement

Let $\mathbf{x}_{0:T-1} \triangleq (x_0, x_1, \ldots, x_{T-1})^{\mathrm{T}}$ be a vector of T observations. The elements of $\mathbf{x}_{0:T-1}$ may be represented by one of the models $\mathcal{M}_{k,\mathbf{p}_k}$, corresponding to the case when the signal is modelled as an AR process with piecewise constant parameters and k ($k = 0, \ldots, k_{\max}$) changepoints. More precisely:

$$\mathcal{M}_{k,\mathbf{p}_k} : \quad x_t = \mathbf{a}_{i,k}^{(p_{i,k})\mathrm{T}} \mathbf{x}_{t-1:t-p_{i,k}} + n_t \quad \text{for } \tau_{i,k} \leq t < \tau_{i+1,k}, \quad i = 0, \ldots, k, \tag{31}$$

where a set of $p_{i,k}$ model parameters ($p_{i,k} = 0, \ldots, p_{\max}$, $\mathbf{p}_k \triangleq \mathbf{p}_{1:k,k}$) for the i^{th} segment under the assumption of k changepoints in the signal is arranged in the vector $\mathbf{a}_{i,k}^{(p_{i,k})} = \left(a_{i,k,1}^{(p_{i,k})}, \ldots, a_{i,k,p_{i,k}}^{(p_{i,k})} \right)^{\mathtt{T}}$ and n_t is i.i.d. Gaussian noise of variance $\sigma_{i,k}^2$ ($\boldsymbol{\sigma}_k^2 \triangleq \boldsymbol{\sigma}_{1:k,k}^2$) associated with this AR model. The changepoints of the model $\mathcal{M}_{k,\mathbf{p}_k}$ are denoted $\boldsymbol{\tau}_k \triangleq \boldsymbol{\tau}_{1:k,k}$ and we adopt the convention $\tau_{0,k} = 0$ and $\tau_{k+1,k} = T - 1$ for notational convenience.

The models can be rewritten in the following matrix form:

$$\mathcal{M}_{k,\mathbf{p}_k} : \quad \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1} = \mathbf{X}_{i,k}^{(p_{i,k})} \mathbf{a}_{i,k}^{(p_{i,k})} + \mathbf{n}_{\tau_{i,k}:\tau_{i+1,k}-1}, \quad i = 0, \ldots, k, \tag{32}$$

where $\mathbf{X}_{i,k}^{(p_{i,k})}$ for the i^{th} segment ($i = 0, ..., k$) is given by:

$$\mathbf{X}_{i,k}^{(p_{i,k})} = \begin{bmatrix} x_{\tau_{i,k}-1} & x_{\tau_{i,k}-2} & \cdots & x_{\tau_{i,k}-p_{i,k}} \\ x_{\tau_{i,k}} & x_{\tau_{i,k}-1} & \cdots & x_{\tau_{i,k}+1-p_{i,k}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\tau_{i+1,k}-2} & x_{\tau_{i+1,k}-3} & \cdots & x_{\tau_{i+1,k}-1-p_{i,k}} \end{bmatrix}. \tag{33}$$

We interpret the first $p_{\max}$ samples as the initial conditions and proceed with analysis on the remaining $T - p_{\max}$ data points.

We assume that the number of changepoints k and the associated parameters $\boldsymbol{\Psi}_k \triangleq \left(\boldsymbol{\tau}_k, \mathbf{p}_k, \{\mathbf{a}_{i,k}^{(p_{i,k})}\}_{i=0,\ldots,k}, \boldsymbol{\sigma}_k^2 \right)$ are unknown. Given $\mathbf{x}_{0:T-1}$, our aim is to estimate k and $\boldsymbol{\Psi}_k$.

5.2.1.2 Bayesian model and estimation objectives

We follow a Bayesian approach where the unknown parameters k, $\boldsymbol{\tau}_k$, $\mathbf{p}_k$, $\{\mathbf{a}_{i,k}^{(p_{i,k})}\}_{i=0,\ldots,k}$, $\boldsymbol{\sigma}_k^2$ are regarded as random with a known prior that reflects our degree of belief in the different values of these quantities. In order to increase robustness of the prior, an additional level of hyperprior distributions [49] is introduced. Thus, an extended hierarchical Bayesian model is proposed, which allows us to define a posterior distribution on the space of all possible structures of the signal. Subsequently, the detection/estimation aims are specified and, finally, we derive the posterior distribution marginalised with respect to the unknown nuisance parameters.

5.2.1.2.1 Prior distribution

In our case it is natural to introduce a binomial distribution as a prior distribution for the number of changepoints and their positions (see [19], [32], [37] for a similar choice).

This implies that:

$$p\left(k,\boldsymbol{\tau}_k\middle|\,\lambda\right)=\lambda^k\left(1-\lambda\right)^{T-2-k}\mathbb{I}_{\boldsymbol{\Upsilon}_k}(\boldsymbol{\tau}_k),\quad 0<\lambda<1, \tag{34}$$

where $\boldsymbol{\Upsilon}_k\triangleq\{\boldsymbol{\tau}_{1:k,k}\in\{1,\dots,T-2\}^k$ such that $\tau_{1,k}\neq\tau_{2,k}\neq\dots\neq\tau_{k,k}\}$. For the model order prior we adopt a truncated Poisson distribution:

$$p\left(p_{i,k}\middle|\,\theta\right)\propto\tfrac{\theta^{p_{i,k}}}{p_{i,k}!}\mathbb{I}_{\{0,\dots,p_{\max}\}}(p_{i,k}), \tag{35}$$

where the mean θ is interpreted as the expected number of poles for the AR model and the normalizing constant is:

$$C_{p_{\max}}=\frac{1}{\sum_{p_{i,k}=0}^{p_{\max}}\frac{\theta^{p_{i,k}}}{p_{i,k}!}}. \tag{36}$$

Furthermore, we assign a normal distribution to the parameters of the AR models

$$\mathbf{a}_{i,k}^{(p_{i.k})}\Big|\,\sigma_{i,k}^2,\delta_{i,k}^2\sim\mathcal{N}\left(0,\sigma_{i,k}^2\delta_{i,k}^2\mathbf{I}_{p_{i,k}}\right),\quad i=0,\dots,k. \tag{37}$$

and a conjugate Inverse-Gamma distribution to the noise variances

$$\sigma_{i,k}^2\Big|\,\tfrac{\nu_0}{2},\tfrac{\gamma_0}{2}\sim\mathcal{IG}(\tfrac{\nu_0}{2},\tfrac{\gamma_0}{2}),\quad\nu_0>0,\,\gamma_0>0,\quad i=0,\dots,k. \tag{38}$$

This choice of prior, given the Gaussian noise model, allows the marginalization of the parameters $\left(\{\mathbf{a}_{i,k}^{(p_{i,k})}\}_{i=0,\dots,k},\boldsymbol{\sigma}_k^2\right)$ in this case.

The algorithm requires the specification of $\lambda,\theta,\delta_{i,k}^2,\nu_0$ and γ_0. It is clear that these parameters play a fundamental role in the segmentation of signals, and in order to robustify the prior, we propose to estimate $\lambda,\theta,\delta_{i,k}^2,\gamma_0$ from the data (see [48], [49] for a similar approach), i.e. we consider $\lambda,\theta,\delta_{i,k}^2,\gamma_0$ to be random. We assign a vague conjugate Inverse-Gamma distribution to the scale hyperparameter $\delta_{i,k}^2$:

$$\delta_{i,k}^2\big|\,\alpha_\delta,\beta_\delta\sim\mathcal{IG}(\alpha_\delta,\beta_\delta),\quad i=0,\dots,k. \tag{39}$$

Moreover, since in our particular case the acceptance ratio for the birth/death of a change-point depends on the hyper-hyperparameter β_δ (see Eq. (60)), we assume that it is also randomly distributed according to a conjugate prior Gamma distribution:

$$\beta_\delta|\,\zeta_\beta,\varkappa_\beta\sim\mathcal{IG}\left(\zeta_\beta,\varkappa_\beta\right). \tag{40}$$

Similarly, we assign a conjugate prior Gamma density to θ :

$$\theta |\, \zeta, \varkappa \sim \mathcal{IG}\left(\zeta, \varkappa\right). \tag{41}$$

We set $\upsilon_0 = 2, \alpha_\delta = 1, \zeta = 1$ and $\zeta_\beta = 1$ to ensure an infinite variance to express ignorance of the value of the parameters and $\varkappa = \epsilon, \varkappa_\beta = \epsilon_\beta$, $(\epsilon, \epsilon_\beta \ll 1)$; and we choose a uniform prior distribution $\lambda \sim \mathcal{U}_{(0,1)}$ and a non-informative improper Jeffreys' prior for γ_0.

Thus, the following hierarchical structure is assumed for the prior of the parameters:

$$\begin{aligned} & p\left(k, \mathbf{\Psi}_k, \lambda, \theta, \boldsymbol{\delta}_k^2, \gamma_0, \beta_\delta\right) = \\ & \quad \prod_{i=0}^{k} \left[p\left(p_{i,k} \middle|\, \theta\right) p\left(\mathbf{a}_{i,k}^{(p_{i,k})} \middle|\, \sigma_{i,k}^2, \delta_{i,k}^2\right) p\left(\sigma_{i,k}^2 \middle|\, \gamma_0\right) p\left(\delta_{i,k}^2 \middle|\, \beta_\delta\right)\right] \\ & \quad \times p\left(k, \boldsymbol{\tau}_k \middle|\, \lambda\right) p\left(\lambda\right) p\left(\theta\right) p\left(\beta_\delta\right) p\left(\gamma_0\right), \end{aligned} \tag{42}$$

which can be visualised with a directed acyclic graph (DAG) as shown in Fig. 6 (for convenience we do not show $\upsilon_0, \alpha_\delta, \zeta, \varkappa, \zeta_\beta, \varkappa_\beta$).

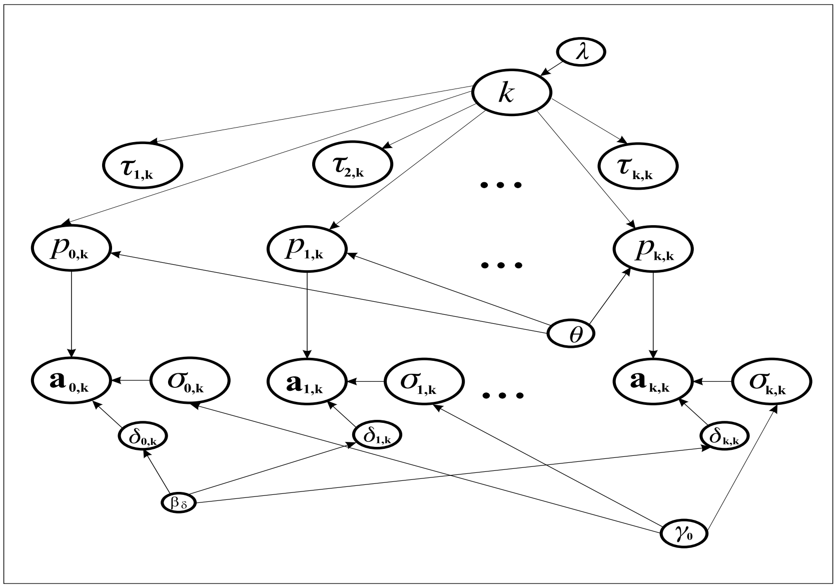

Figure 6: Directed acyclic graph for the prior distribution.

5.2.1.2.2 *Bayesian hierarchical model*

For our problem, the overall parameter space can be written as a finite union of subspaces $\mathbf{\Theta} \triangleq \cup_{k=0}^{k_{\max}} \{k\} \times \mathbf{\Upsilon}_k \times \prod_{i=0}^{k} \mathbf{\Phi}_{p_{i,k}} \times \mathbf{\Xi}_k$, where $\mathbf{\Phi}_{p_{i,k}}$ denotes the space of the parameters $p_{i,k}, \mathbf{a}_{i,k}^{(p_{i,k})}, \sigma_{i,k}^2$ for the i^{th} segment, i.e. $\mathbf{\Phi}_0 \triangleq \mathbb{R}^+$, $\mathbf{\Phi}_{p_{i,k}} \triangleq \cup_{p_{i,k}=0}^{p_{\max}} \{p_{i,k}\} \times \left(\mathbb{R}^{p_{i,k}} \times \mathbb{R}^+\right)$, $\mathbf{\Xi}_k$ denotes the hyperparameter $\left(\lambda, \theta, \boldsymbol{\delta}_k^2, \gamma_0\right)$ and hyper-hyperparameter $\left(\beta_\delta\right)$ space, which is given by $\boldsymbol{\Xi}_k \triangleq (0,1) \times \mathbb{R}^+ \times (\mathbb{R}^+)^k \times \mathbb{R}^+ \times \mathbb{R}^+$, and $k_{\max} = T - 2$.

There is a natural hierarchical structure for this set-up, which we can formalise by modelling the joint distribution of all variables as follows:

$$p\left(k, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_k, \mathbf{x}_{0:T-1}\right) = p\left(k, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_k\right) p\left(\mathbf{x}_{0:T-1} \middle| k, \boldsymbol{\Psi}_k\right), \tag{43}$$

where $\boldsymbol{\xi}_k = \{\lambda, \theta, \boldsymbol{\delta}_k^2, \gamma_0, \beta_\delta\}$. As the excitation is assumed to be i.i.d Gaussian (see Section 5.2.1.1), the likelihood takes the form:

$$p\left(\mathbf{x}_{0:T-1} \middle| k, \boldsymbol{\Psi}_k\right) = \\ \prod_{i=0}^{k} \left(2\pi\sigma_{i,k}^2\right)^{-\frac{\tau_{i+1,k}-\tau_{i,k}}{2}} \exp\left(-\frac{\left(\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1} - \mathbf{X}_{i,k}^{(p_{i,k})}\mathbf{a}_{i,k}^{(p_{i,k})}\right)^{\mathrm{T}}\left(\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1} - \mathbf{X}_{i,k}^{(p_{i,k})}\mathbf{a}_{i,k}^{(p_{i,k})}\right)}{2\sigma_{i,k}^2}\right). \tag{44}$$

This assumption is commonly used one, since the probability distribution that has maximum entropy, subject to knowledge of the first two moments of the noise distribution, is Gaussian, and, therefore, the likelihood of this form is often used unless one has further knowledge concerning the noise statistics. It has been shown to work well in practice and allows the marginalization of the nuisance parameters in our case.

5.2.1.2.3 Bayesian detection and estimation

Any Bayesian inference on k and $\boldsymbol{\Psi}_k, \boldsymbol{\xi}_k$ is based on the following posterior obtained using Bayes' theorem:

$$p\left(k, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right) \propto p\left(\mathbf{x}_{0:T-1} \middle| k, \boldsymbol{\Psi}_k\right) p\left(k, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_k\right). \tag{45}$$

Our aim is to estimate this posterior distribution, and more specifically some of its features such as the marginal distributions. In our case, however, it is not possible to obtain these quantities analytically, as it requires the evaluation of high-dimensional integrals of nonlinear functions in the parameters (see Section 5.2.1.2.4). Therefore, we apply MCMC methods and a reversible jump MCMC method in particular (see Section 5.3.3 for details). The key idea is to build an ergodic Markov chain $\left(k^{(j)}, \boldsymbol{\Psi}_k^{(j)}, \boldsymbol{\xi}_k^{(j)}\right)_{j\in\mathbb{N}}$ whose equilibrium distribution is the desired posterior distribution. Under weak additional assumptions, the $P \gg 1$ samples generated by the Markov chain are asymptotically distributed according to the posterior distribution and thus allow easy evaluation of all posterior features of interest. For example,

$$\widehat{p}\left(k = l \middle| \mathbf{x}_{0:T-1}\right) = \frac{1}{P}\sum_{j=1}^{P} \mathbb{I}_{\{l\}}\left(k^{(j)}\right). \tag{46}$$

In practice, we take the most straightforward approach to obtain marginal densities: the samples $k^{(j)}$ from the joint posterior density $p\left(k, \mathbf{\Psi}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right)$ are collected into frequency bins, ignoring other parameters, and the histogram is plotted directly. Once the estimate of $p\left(k \middle| \mathbf{x}_{0:T-1}\right)$ is obtained, the model selection is performed using the MMAP criterion, from which the number of changepoints is estimated as

$$\widehat{k} = \underset{k \in \{0,\ldots,k_{\max}\}}{\arg\max}\ \widehat{p}\left(k \middle| \mathbf{x}_{0:T-1}\right). \tag{47}$$

Having fixed $k = \widehat{k}$, we proceed with the estimation of $p\left(\tau_{i,\hat{k}} \middle| \widehat{k}, \mathbf{x}_{0:T-1}\right)$, $i = 1, \ldots, \hat{k}$, and $p\left(p_{i,\hat{k}} \middle| \widehat{k}, \mathbf{x}_{0:T-1}\right)$, $i = 0, \ldots, \hat{k}$, by plotting the corresponding histograms, from which the estimates of the positions of changepoints and the model orders for each segment are obtained in exactly the same way (using MMAP):

$$\widehat{\tau}_{i,\hat{k}} = \underset{\tau_{i,\hat{k}} \in \{1,\ldots,T-1\}}{\arg\max}\ \widehat{p}\left(\tau_{i,\hat{k}} \middle| \widehat{k}, \mathbf{x}_{0:T-1}\right),\ i = 1, \ldots, \hat{k}, \tag{48}$$

$$\widehat{p}_{i,\hat{k}} = \underset{p_{i,\hat{k}} \in \{0,\ldots,p_{\max}\}}{\arg\max}\ \widehat{p}\left(p_{i,\hat{k}} \middle| \widehat{k}, \mathbf{x}_{0:T-1}\right),\ i = 0, \ldots, \hat{k}. \tag{49}$$

In fact, as shown in the next section, the parameters $\left(\{\mathbf{a}_{i,k}^{(p_{i,k})}\}_{i=0,\ldots,k}, \boldsymbol{\sigma}_k^2\right)$ can be integrated out analytically due to the Gaussian noise assumption and the choice of prior distribution, and, if necessary, can then be straightforwardly estimated.

5.2.1.2.4 Integration of the nuisance parameters

The proposed Bayesian model allows for the integration of the nuisance parameters $\left(\{\mathbf{a}_{i,k}^{(p_{i,k})}\}_{i=0,\ldots,k}, \boldsymbol{\sigma}_k^2\right)$ and subsequently gives us the expression for $p\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right)$ up to a normalizing constant:

$$\begin{aligned} &p\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right) \propto \\ &\qquad \int_{\mathbb{R}^+} \int_{\mathbb{R}^{p_{0,k}}} \ldots \int_{\mathbb{R}^+} \int_{\mathbb{R}^{p_{k,k}}} p\left(k, \mathbf{\Psi}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right) d\mathbf{a}_{0,k} d\sigma_{0,k}^2 \ldots d\mathbf{a}_{k,k} d\sigma_{k,k}^2. \end{aligned} \tag{50}$$

Thus, from Eq. (45):

$$
\begin{aligned}
&p\left(k,\boldsymbol{\tau}_k,\mathbf{p}_k,\boldsymbol{\xi}_k\middle|\,\mathbf{x}_{0:T-1}\right)\propto\\
&\quad\prod_{i=0}^{k}\left[\left(2\pi\sigma_{i,k}^2\right)^{-\frac{p_{i,k}}{2}}\exp\left(-\frac{\left(\mathbf{a}_{i,k}^{(p_{i,k})}-\mathbf{m}_{i,k}^{(p_{i,k})}\right)^{\mathtt{T}}\left[\mathbf{M}_{i,k}^{(p_{i,k})}\right]^{-1}\left(\mathbf{a}_{i,k}^{(p_{i,k})}-\mathbf{m}_{i,k}^{(p_{i,k})}\right)}{2\sigma_{i,k}^2}\right)\right]\\
&\quad\times\prod_{i=0}^{k}\left[\left(\sigma_{i,k}^2\right)^{-\frac{\nu_0}{2}-\frac{\tau_{i+1,k}-\tau_{i,k}}{2}-1}\exp\left(-\frac{\left(\gamma_0+\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}^{\mathtt{T}}\mathbf{P}_{i,k}^{(p_{i,k})}\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}\right)}{2\sigma_{i,k}^2}\right)\right]\\
&\quad\times\prod_{i=0}^{k}\left[(2\pi)^{-\frac{\tau_{i+1,k}-\tau_{i,k}}{2}}\,\frac{(\gamma_0)^{\frac{\nu_0}{2}}}{\Gamma(\frac{v_0}{2})}\frac{\beta_\delta^{\alpha_\delta}}{\Gamma(\alpha_\delta)}\left(\delta_{i,k}^2\right)^{-\frac{p_{i,k}}{2}}\left(\delta_{i,k}^2\right)^{-\alpha_\delta-1}\exp\left(-\frac{\beta_\delta}{\delta_{i,k}^2}\right)\right]\\
&\quad\times\prod_{i=0}^{k}\left[\frac{c_{p_{\max}}\theta^{p_{i.k}}}{p_{i,k}!}\right]\lambda^k\left(1-\lambda\right)^{T-k-2}\gamma_0^{-1}\beta_\delta^{\zeta-1}\exp\left(-\varkappa\beta_\delta\right)\mathbb{I}_{\boldsymbol{\Upsilon}_k}(\boldsymbol{\tau}_k)\mathbb{I}_{\mathcal{F}_k}(\mathbf{p}_k),
\end{aligned}
\tag{51}
$$

where $\mathcal{F}_k\triangleq\{0,\ldots,p_{\max}\}^k$ and

$$
\begin{aligned}
&\mathbf{M}_{i,k}^{(p_{i,k})}=\left[\mathbf{X}_{i,k}^{(p_{i,k})\mathtt{T}}\mathbf{X}_{i,k}^{(p_{i,k})}+\frac{1}{\delta_{i,k}^2}\mathbf{I}_{p_{i,k}}\right]^{-1},\quad\mathbf{m}_{i,k}^{(p_{i,k})}=\mathbf{M}_{i,k}^{(p_{i,k})}\mathbf{X}_{i,k}^{(p_{i,k})\mathtt{T}}\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1},\\
&\mathbf{P}_{i,k}^{(p_{i,k})}=\mathbf{I}_{\tau_{i,k}:\tau_{i+1,k}-1}-\mathbf{X}_{i,k}^{(p_{i,k})}\mathbf{M}_{i,k}^{(p_{i,k})}\mathbf{X}_{i,k}^{(p_{i,k})\mathtt{T}}.
\end{aligned}
\tag{52}
$$

The marginalised expression becomes:

$$
\begin{aligned}
&p\left(k,\boldsymbol{\tau}_k,\mathbf{p}_k,\boldsymbol{\xi}_k\middle|\,\mathbf{x}_{0:T-1}\right)\propto\\
&\quad\prod_{i=0}^{k}\left[\Gamma(\tfrac{v_0+\tau_{i+1,k}-\tau_{i,k}}{2})\left(\gamma_0+\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}^{\mathtt{T}}\mathbf{P}_{i,k}^{(p_{i,k})}\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}\right)^{-\frac{v_0+\tau_{i+1,k}-\tau_{i,k}}{2}}\right]\\
&\quad\times\prod_{i=0}^{k}\left[\left|\mathbf{M}_{i,k}^{(p_{i,k})}\right|^{\frac{1}{2}}\pi^{-\frac{\tau_{i+1,k}-\tau_{i,k}}{2}}\frac{(\gamma_0)^{\frac{\nu_0}{2}}}{\Gamma(\frac{v_0}{2})}\frac{\beta_\delta^{\alpha_\delta}}{\Gamma(\alpha_\delta)}\left(\delta_{i,k}^2\right)^{-\alpha_\delta-\frac{p_{i,k}}{2}-1}\exp\left(-\frac{\beta_\delta}{\delta_{i,k}^2}\right)\right]\\
&\quad\times\prod_{i=0}^{k}\left[\frac{c_{p_{\max}}\theta^{p_{i.k}}}{p_{i,k}!}\right]\times\lambda^k\left(1-\lambda\right)^{T-k-2}\gamma_0^{-1}\beta_\delta^{\zeta-1}\exp\left(-\varkappa\beta_\delta\right)\mathbb{I}_{\boldsymbol{\Upsilon}_k}(\boldsymbol{\tau}_k)\mathbb{I}_{\mathcal{F}_k}(\mathbf{p}_k),
\end{aligned}
\tag{53}
$$

As it was already pointed out in Section 5.2.1.2.3, this posterior distribution is complex in the parameters $(k,\boldsymbol{\tau}_k,\mathbf{p}_k,\boldsymbol{\xi}_k)$ and the posterior model probability $p\left(k,\mathbf{p}_k\middle|\,\mathbf{x}_{0:T-1}\right)$ cannot be determined analytically. In the next section we develop a method to estimate $p\left(k,\boldsymbol{\tau}_k,\mathbf{p}_k,\boldsymbol{\xi}_k\middle|\,\mathbf{x}_{0:T-1}\right)$ or, if needed, $p\left(k,\boldsymbol{\tau}_k,\mathbf{p}_k,\{\mathbf{a}_{i,k}^{(p_{i,k})}\}_{i=0,\ldots,k},\boldsymbol{\sigma}_k^2,\boldsymbol{\xi}_k\middle|\,\mathbf{x}_{0:T-1}\right)$..

5.2.1.3 MCMC algorithm

The problem addressed here is, in fact, a model uncertainty problem of variable dimensionality in terms of both the number of changepoints and the model order for each segment. It can be treated very efficiently through the use of MCMC methods, and reversible jump MCMC [29] is particularly suitable for this case. As it was described before (see Chapter 4) this method extends the traditional Metropolis-Hastings algorithm to the case where moves from one dimension to another are proposed with a certain acceptance probability. This probability should be designed in a special way in order to preserve reversibility and thus ensure that $p\left(k, \mathbf{\Psi}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right)$ is the invariant distribution of the Markov chain (MC). In general, if we propose a move from model $(k, \mathbf{p}_k)$ with parameters $(\boldsymbol{\tau}_k, \boldsymbol{\xi}_k)$ to model $(k', \mathbf{p}_{k'})$ with parameters $(\boldsymbol{\tau}_{k'}, \boldsymbol{\xi}_{k'})$ using a proposal distribution $q\left(k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k'} \middle| k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k\right)$, the acceptance probability is given by:

$$\alpha = \min\left\{1, \frac{p\left(k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k'} \middle| \mathbf{x}_{0:T-1}\right)}{p\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k \middle| \mathbf{x}_{0:T-1}\right)} \frac{q\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k \middle| k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k'}\right)}{q\left(k', \boldsymbol{\tau}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\xi}_{k'} \middle| k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\xi}_k\right)}\right\}. \tag{54}$$

Here the proposal is made directly in the new parameter space rather than via "dimensional" matching random variables (see [29]) and the Jacobian term is therefore equal to 1 ([16], [26], [52]).

In fact, the only condition to be fulfilled in selecting different types of moves is to be able to maintain the correct invariant distribution. A particular choice will only affect the convergence rate of the algorithm. To ensure a low level of rejections, we want the proposed "jumps" to be small; therefore the following moves have been selected:

- birth of a changepoint (proposing a new changepoint at random),
- death of a changepoint (removing a changepoint chosen randomly),
- update of the positions of changepoints (proposing a new position for each of the existing changepoints).

At each iteration one of the moves described above is randomly chosen with probabilities b_k, d_k and u_k such that $b_k + d_k + u_k = 1$ for all $0 \leq k \leq k_{\max}$. For $k = 0$ the death of a changepoint is impossible and for $k = k_{\max}$, the birth is impossible, thus $d_0 \triangleq 0$, $b_{k_{\max}} \triangleq 0$. Otherwise we choose $b_k = d_k = u_k$. After each move we perform the update of the number of poles for each AR model. We now describe the main steps of the algorithm:

Reversible Jump MCMC algorithm (main procedure).

1. Initialize $\left(k^{(0)}, \boldsymbol{\tau}_k^{(0)}, \mathbf{p}_k^{(0)}, \lambda^{(0)}, \theta^{(0)}, \boldsymbol{\delta}_k^{2(0)}, \gamma_0^{(0)}, \beta_\delta^{(0)}\right) \in \mathbf{\Theta}$. Set $j = 1$.
2. Iteration j.

   If $\left(u \sim \mathcal{U}_{(0,1)}\right) \leq b_{k^{(j)}}$ then birth of a new changepoint (see Section 5.2.1.3.1).

   else if $u \leq b_{k^{(j)}} + d_{k^{(j)}}$ then death of a changepoint (see Section 5.2.1.3.1).

else update the changepoints positions (see Section 5.2.1.3.2).

3. Update of the number of poles (see Section 5.2.1.3.3).

4. $j \leftarrow j+1$ and goto 2.

We now detail these different steps of the algorithm. To simplify the notation, we drop the superscript (j) from all variables at iteration j.

5.2.1.3.1 Death/birth of the changepoints

First, let the current state of the MC be $\left(k+1, \boldsymbol{\tau}_{k+1}, \mathbf{p}_{k+1}, \boldsymbol{\delta}^2_{k+1}, \lambda, \theta, \gamma_0, \beta_\delta\right)$ and consider the death move, which implies a modification of the dimension of the model respectively from $k+1$ to k. Our proposal begins by choosing a changepoint to be removed among $k+1$ existing ones. If the move is accepted then two segments $(l-1)^{th}$ and l^{th} will be merged, thus reducing $k+1$ by 1, and a new AR model will be created. We choose the order of the new proposed AR model to be $p_{ol} = p_{1l} + p_{2l}$, where p_{1l}, p_{2l} are the orders of the existing $(l-1)^{th}$ and l^{th} AR models[1] (see Fig. 7). The choice of proposal distribution for the hyperparameter δ^2_{ol} will be described later. The algorithm proceeds as follows:

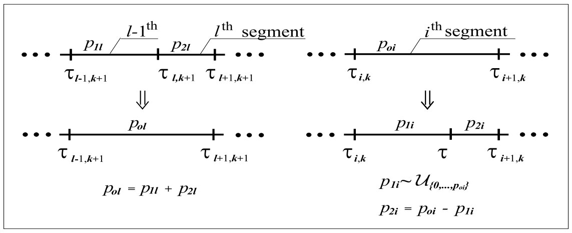

Figure 7: Death (left) and birth (right) moves.

Algorithm for the death move

- Choose a changepoint among the $k+1$ existing ones $l \sim \mathcal{U}_{\{1,\dots,k+1\}}$.
- The proposed model order is $p_{ol} = p_{1l} + p_{2l}$, where $p_{1l} = p_{l-1,k+1}$, $p_{2l} = p_{l,k+1}$; sample $\delta^2_{ol} \big| \left(\tau_{l-1,k+1}, \tau_{l+1,k+1}, p_{ol}, \mathbf{x}_{\tau_{l-1,k+1}:\tau_{l+1,k+1}-1}\right)$, see Eq. (58)

[1] We keep this notation to obtain a general equation for acceptance probabilities.

- Evaluate α_{death}, see Eq. (60).
- If $\left(u_d \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{death}$ then the new state of the MC becomes
  $(k, \{\boldsymbol{\tau}_{1:l-1,k+1}, \boldsymbol{\tau}_{l+1:k+1,k+1}\}, \{\mathbf{p}_{1:l-2,k+1}, p_{ol}, \mathbf{p}_{l+1:k+1,k+1}\},$
  $\{\boldsymbol{\delta}^2_{1:l-2,k+1}, \delta^2_{ol}, \boldsymbol{\delta}^2_{l+1:k+1,k+1}\}, \lambda, \theta, \gamma_0, \beta_\delta),$
  otherwise it stays $\left(k+1, \boldsymbol{\tau}_{k+1}, \mathbf{p}_{k+1}, \boldsymbol{\delta}^2_{k+1}, \lambda, \theta, \gamma_0, \beta_\delta\right)$.

For the birth move ($k \to k+1$), again, first the position of a new changepoint τ is proposed. For $\tau_{i,k} < \tau < \tau_{i+1,k}$ the i^{th} segment should be split into two and the new AR model orders should be $p_{1i} \sim \mathcal{U}_{\{0,\ldots,p_{oi}\}}$ and $p_{2i} = p_{oi} - p_{1i}$, where p_{oi} is the order of the i^{th} model (see Fig. 7). This choice for the number of poles ensures that birth/death moves are reversible ($p_{oi} = p_{1i} + p_{2i}$). Thus, assuming that the current state of the MC is $\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}^2_k, \lambda, \theta, \gamma_0, \beta_\delta\right)$, we have:

Algorithm for the birth move

- Propose a new changepoint τ in $\{1, \ldots, T-2\}$: $\tau \sim \mathcal{U}_{\{1,\ldots,T-2\}\setminus\{\boldsymbol{\tau}_k\}}$.
- The proposed model orders are: $p_{1i} = \mathcal{U}_{\{1,\ldots,p_{oi}\}}$, $p_{2i} = p_{oi} - p_{1i}$, where $p_{oi} = p_{i,k}$
  for $\tau_{i,k} \leq \tau < \tau_{i+1,k}$;
  sample $\delta^2_{1i} | \left(\tau_{i,k}, \tau, p_{1i}, \mathbf{x}_{\tau_{i,k}:\tau-1}\right)$, $\delta^2_{2i} | \left(\tau, \tau_{i+1,k}, p_{2i}, \mathbf{x}_{\tau:\tau_{i+1,k}-1}\right)$ see Eq. (58);
- Evaluate α_{birth}, see Eq. (60).
- If $\left(u_b \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{birth}$ then the new state of the MC becomes
  $(k+1, \{\boldsymbol{\tau}_{1:i,k}, \tau, \boldsymbol{\tau}_{i+1:k,k}\}, \{\mathbf{p}_{1:i-1,k}, p_{1i}, p_{2i}, \mathbf{p}_{i+1:k,k}\},$
  $\{\boldsymbol{\delta}^2_{1:i-1,k}, \delta^2_{1i}, \delta^2_{2i}, \boldsymbol{\delta}^2_{i+1:k,k}\}, \lambda, \theta, \gamma_0, \beta_\delta),$
  otherwise it stays $\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}^2_k, \lambda, \theta, \gamma_0, \beta_\delta\right)$.

To perform these moves in practice, it is necessary to choose a proposal distribution for the elements of $\boldsymbol{\delta}^2_k$ in such a way that we avoid rejecting too many candidates. If the number of changepoints were fixed and we wanted just to update the values of $\boldsymbol{\delta}^2_k$, we would sample the elements of the vector according to a standard Gibbs move (see [50]), i.e. from Eq. (51)

$$\mathcal{IG}\left(\delta^2_{i,k};\, \alpha_\delta + \frac{p_{i,k}}{2},\, \beta_\delta + \frac{\mathbf{a}_{i,k}^{(p_{i,k})\mathrm{T}} \mathbf{a}_{i,k}^{(p_{i,k})}}{2\sigma^2_{i,k}}\right), \tag{55}$$

where $\mathbf{a}_{i,k}^{(p_{i,k})}, \sigma_{i,k}^2$ are sampled from the following distributions (see Eq. (51)):

$$\mathcal{IG}\left(\sigma_{i,k}^2;\ \frac{\nu_0+\tau_{i+1,k}-\tau_{i,k}}{2},\ \frac{\gamma_0+\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}^{\mathrm{T}}\mathbf{P}_{i,k}^{(p_{i,k})}\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}}{2}\right), \tag{56}$$

$$\mathcal{N}\left(\mathbf{a}_{i,k}^{(p_{i,k})};\ \mathbf{m}_{i,k}^{(p_{i,k})},\ \sigma_{i,k}^2\mathbf{M}_{i,k}^{(p_{i,k})}\right), \tag{57}$$

with matrices $\mathbf{M}_{i,k}^{(p_{i,k})}, \mathbf{P}_{i,k}^{(p_{i,k})}$ and vector $\mathbf{m}_{i,k}^{(p_{i,k})}$ depending on the value of $\delta_{i,k}^2$ before updating. In the case of a birth/death move, we do not have any previous value but instead we can use the mean of the distribution $\mathcal{IG}\left(\alpha_\delta + \frac{p_{i,k}}{2},\ \beta_\delta\right)$: $\delta_{i,k}^{2*}=\frac{\beta_\delta}{\alpha_\delta+\frac{p_{i,k}}{2}-1}$and sample $\delta_{i,k}^2$ using Metropolis-Hastings steps (see [50]); the corresponding matrices are denoted as $\mathbf{M}_{i,k}^{(p_{i,k})*}, \mathbf{P}_{i,k}^{(p_{i,k})*}$and $\mathbf{m}_{i,k}^{(p_{i,k})*}$. Taking this into account, we construct our proposal distribution in the following way:

$$\mathcal{IG}\big(\delta_{i,k}^2;\ \tilde{\alpha}_\delta,\ \tilde{\beta}_\delta\big)=\mathcal{IG}\left(\delta_{i,k}^2;\ \alpha_\delta+\frac{p_{i,k}}{2},\ \beta_\delta+\frac{\overline{\mathbf{a}}_{i,k}^{(p_{i,k})\mathrm{T}}\overline{\mathbf{a}}_{i,k}^{(p_{i,k})}}{2\overline{\sigma_{i,k}^2}}\right), \tag{58}$$

where $\overline{\mathbf{a}}_{i,k}^{(p_{i,k})}$, $\overline{\sigma_{i,k}^2}$ are the means of the distributions corresponding to Eq. (56), (57) but with $\mathbf{M}_{i,k}^{(p_{i,k})*}, \mathbf{P}_{i,k}^{(p_{i,k})*}$and $\mathbf{m}_{i,k}^{(p_{i,k})*}$:

$$\overline{\mathbf{a}}_{i,k}^{(p_{i,k})} = \mathbf{m}_{i,k}^{(p_{i,k})*},\ \overline{\sigma_{i,k}^2} = \frac{\gamma_0+\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}^{\mathrm{T}}\mathbf{P}_{i,k}^{(p_{i,k})*}\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}}{\nu_0+\tau_{i+1,k}-\tau_{i,k}-2}. \tag{59}$$

The acceptance ratio of the birth and death (of changepoint) moves is evaluated according to the general expression (54). We obtain the acceptance probabilities:

$$\alpha_{birth} = \min\{1, r_{birth}\} \text{ and } \alpha_{death} = \min\left\{1, r_{birth}^{-1}\right\} \tag{60}$$

where

$$\begin{aligned} r_{birth} = \quad & \frac{p\big(k+1,\boldsymbol{\tau}_{k+1},\mathbf{p}_{k+1},\lambda,\theta,\boldsymbol{\delta}_{k+1}^2,\gamma_0,\beta_\delta\big|\mathbf{x}_{0:T-1}\big)}{p\big(k,\boldsymbol{\tau}_k,\mathbf{p}_k,\lambda,\theta,\boldsymbol{\delta}_k^2,\gamma_0,\beta_\delta\big|\mathbf{x}_{0:T-1}\big)} \times \frac{q(k,\boldsymbol{\tau}_k,\mathbf{p}_k|k+1,\boldsymbol{\tau}_{k+1},\mathbf{p}_{k+1})}{q(k+1,\boldsymbol{\tau}_{k+1},\mathbf{p}_{k+1}|k,\boldsymbol{\tau}_k,\mathbf{p}_k)} \\ & \times \frac{q\big(\boldsymbol{\delta}_k^2\big|k+1,\boldsymbol{\tau}_{k+1},\mathbf{p}_{k+1},\boldsymbol{\delta}_{k+1}^2,\gamma_0,\beta_\delta,\mathbf{x}_{0:T-1}\big)}{q\big(\boldsymbol{\delta}_{k+1}^2\big|k,\boldsymbol{\tau}_k,\mathbf{p}_k,\boldsymbol{\delta}_k^2,\gamma_0,\beta_\delta,\mathbf{x}_{0:T-1}\big)} \end{aligned} \tag{61}$$

and

$$\begin{aligned} \frac{q(k,\boldsymbol{\tau}_k,\mathbf{p}_k|k+1,\boldsymbol{\tau}_{k+1},\mathbf{p}_{k+1})}{q(k+1,\boldsymbol{\tau}_{k+1},\mathbf{p}_{k+1}|k,\boldsymbol{\tau}_k,\mathbf{p}_k)} &= \quad \frac{q(k|k+1)}{q(k+1|k)} \times \frac{q(\boldsymbol{\tau}_k|k+1,\boldsymbol{\tau}_{k+1})}{q(\boldsymbol{\tau}_{k+1}|k,\boldsymbol{\tau}_k)} \times \frac{q(\mathbf{p}_k|k+1,\mathbf{p}_{k+1})}{q(\mathbf{p}_{k+1}|k,\mathbf{p}_k)} \\ &= \quad \frac{d_{k+1}}{b_k} \times \frac{(T-2-k)}{(k+1)} \times \frac{(p_{oi}+1)}{1}. \end{aligned} \tag{62}$$

Finally, from Eq. (53) for the birth of the changepoint τ, $(\tau_{i,k} \leq \tau < \tau_{i+1,k})$ we obtain:

$$r^{i}_{birth} = \frac{\lambda}{(1-\lambda)} \frac{f(\tau_{i,k},\tau,p_{1i},\delta^2_{1i}) f(\tau,\tau_{i+1,k},p_{2i},\delta^2_{2i})}{f(\tau_{i+1,k},\tau_{i,k},p_{i,k},\delta^2_{i,k})} \frac{d_{k+1}}{b_k} \frac{(T-2-k)(p_{oi}+1)}{(k+1)}, \tag{63}$$

where, for convenience, we denote for the segment between the changepoints $\tau_{i,k}, \tau_{i+1,k}$:

$$f(\tau_{i,k}, \tau_{i+1,k}, p_{i,k}, \delta^2_{i,k}) = \frac{(\gamma_0)^{\frac{\nu_0}{2}}}{\Gamma(\frac{v_0}{2})} \frac{c_{p_{\max}} \theta^{p_{i.k}}}{p_{i,k}!} \frac{\beta_\delta^{\alpha_\delta}}{\Gamma(\alpha_\delta)} \left[\frac{\tilde{\beta}_\delta^{\tilde{\alpha}_\delta}}{\Gamma(\tilde{\alpha}_\delta)} \right]^{-1} \exp(-\frac{\beta_\delta - \tilde{\beta}_\delta}{\delta^2_{i,k}})$$
$$\times \Gamma\left(\frac{v_0 + \tau_{i+1,k} - \tau_{i,k}}{2}\right) \left| \mathbf{M}^{(p_{i,k})}_{i,k} \right|^{\frac{1}{2}} \left(\gamma_0 + \mathbf{x}^{\mathrm{T}}_{\tau_{i,k}:\tau_{i+1,k}-1} \mathbf{P}^{(p_{i,k})}_{i,k} \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1} \right)^{-\frac{v_0 + \tau_{i+1,k} - \tau_{i,k}}{2}} . \tag{64}$$

5.2.1.3.2 Update of the changepoint positions

Although the update of the changepoint positions does not involve the change in dimension k, it is somewhat more complicated than the birth/death moves. In fact, updating the position of changepoint $\tau_{l,k}$ means removing the l^{th} changepoint and proposing instead a new one τ. We determine i such that $\tau_{i,k} < \tau < \tau_{i+1,k}$ and it is worth noticing that if $i \neq l$ the update move may actually be described as a combination of the birth of the changepoint τ and the death of the changepoint $\tau_{l,k}$ (see Fig. 8). Otherwise, we leave the model orders the same and just sample the values of the hyperparameters δ^2_{1l}, δ^2_{2l}. This process is repeated for all existing changepoints, $l = 1, \ldots, k$, and is described below in more detail.

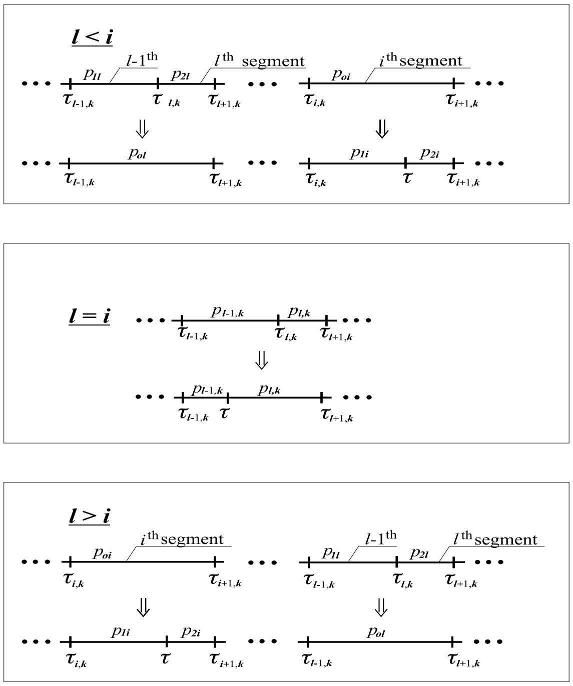

Figure 8: Update of the changepoint positions

Algorithm for the update of the changepoint positions

```
For $l = 1, \dots, k$

• Propose a new position for the $l^{th}$ changepoint $\tau \sim \mathcal{U}_{\{1,\dots,T-2\} \setminus \{\boldsymbol{\tau}_k\}}$
  and determine $i$ such that $\tau_{i,k} < \tau < \tau_{i+1,k}$.

• If $l \neq i$ then
  $p_{1i} = \mathcal{U}_{\{1,\dots,p_{oi}\}}$, $p_{2i} = p_{oi} - p_{1i}$, where $p_{oi} = p_{i,k}$,
  $p_{ol} = p_{1l} + p_{2l}$, where $p_{1l} = p_{l-1,k}$, $p_{2l} = p_{l,k}$;
  sample $\delta^2_{1i} \big| \left(\tau_{i,k}, \tau, p_{1i,} \mathbf{x}_{\tau_{i,k}:\tau-1}\right)$, $\delta^2_{2i} \big| \left(\tau, \tau_{i+1,k}, p_{2i,} \mathbf{x}_{\tau:\tau_{i+1,k}-1}\right)$
  and $\delta^2_{ol} \big| \left(\tau_{l-1,k+1}, \tau_{l+1,k+1}, p_{o,l}, \mathbf{x}_{\tau_{l-1,k+1}:\tau_{l+1,k+1}-1}\right)$, see Eq. (58);
  else sample $\delta^2_{1l} \big| \left(\tau_{l-1,k}, \tau, p_{l-1,k,} \mathbf{x}_{\tau_{l-1,k}:\tau-1}\right)$, $\delta^2_{2l} \big| \left(\tau, \tau_{l+1,k}, p_{l,k,} \mathbf{x}_{\tau:\tau_{l+1,k}-1}\right)$, see Eq. (58);

• Evaluate $\alpha_{update}$, if $l \neq i$ then see Eq. (65) else see Eq. (66)

• If $\left(u_u \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{update}$ then the new state of the MC becomes

    – if $l < i$ then
      $(k, \{\boldsymbol{\tau}_{1:l-1,k}, \boldsymbol{\tau}_{l+1:i,k}, \boldsymbol{\tau}, \boldsymbol{\tau}_{i+1:k,k}\}, \{\mathbf{p}_{1:l-2,k}, p_{ol}, \mathbf{p}_{l+1:i-1,k}, p_{1i}, p_{2i}, \mathbf{p}_{i+1:k,k}\},$
      $\{\boldsymbol{\delta}^2_{1:l-2,k}, \delta^2_{ol}, \boldsymbol{\delta}^2_{l+1:i-1,k} \delta^2_{1i}, \delta^2_{2i}, \boldsymbol{\delta}^2_{i+1:k,k}\}, \lambda, \theta, \gamma_0, \beta_\delta);$
    – else if $l > i$ then
      $(k, \{\boldsymbol{\tau}_{1:i,k}, \boldsymbol{\tau}, \boldsymbol{\tau}_{i+1:l-1,k}, \boldsymbol{\tau}_{l+1:k,k}\}, \{\mathbf{p}_{1:i-1,k}, p_{1i}, p_{2i}, \mathbf{p}_{i+1:l-2,k}, p_{ol}, \mathbf{p}_{l+1:k,k}\},$
      $\{\boldsymbol{\delta}^2_{1:i-1,k}, \delta^2_{1i}, \delta^2_{2i}, \boldsymbol{\delta}^2_{i+1:l-2,k} \delta^2_{ol}, \boldsymbol{\delta}^2_{l+1:k,k}\}, \lambda, \theta, \gamma_0, \beta_\delta)$
    – else $(k, \{\boldsymbol{\tau}_{1:l-1,k}, \tau, \boldsymbol{\tau}_{l+1:k,k}\}, \mathbf{p}_k, \{\boldsymbol{\delta}^2_{1:l-2,k}, \delta^2_{1l}, \delta^2_{2l}, \boldsymbol{\delta}^2_{l+1:k,k}\}, \lambda, \theta, \gamma_0, \beta_\delta)$.

  otherwise it stays $\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}^2_k, \lambda, \theta, \gamma_0, \beta_\delta\right)$.
```

■

Since for $l \neq i$ the update of the positions of changepoints combines the birth of the i^{th} changepoint and death of the l^{th} changepoint at the same time, the acceptance ratio for the proposed move is given by:

$$\alpha_{update} = \min\left\{1, r^{i}_{birth} r^{l}_{death}\right\}. \tag{65}$$

If $l = i$, it becomes:

$$r_{update} = \frac{f(\tau_{l-1,k},\tau,p_{l-1,k},\delta^2_{1l})f(\tau,\tau_{l+1,k},p_{l,k},\delta^2_{2l})}{f(\tau_{l-1,k},\tau_{l,k},p_{l-1,k},\delta^2_{l-1,k})f(\tau_{l,k},\tau_{l+1,k},p_{l,k},\delta^2_{l,k})}. \tag{66}$$

where $f(\cdot)$ is defined in (64).

5.2.1.3.3 Update of the number of poles

The update of the number of poles for each segment does not involve changing the number of changepoints and their positions. However, we still have to perform "jumps" between the subspaces of different dimensions $p_{i,k}$ and will therefore continue using the reversible jump MCMC method, though it is formulated now in a less complicated form. Similarly, the moves are chosen to be: (1) birth of the pole ($p_{i,k} \to p_{i,k}+1$), (2) death of the pole ($p_{i,k} \to p_{i,k}-1$) and (3) just the update of the hyperparameter $\delta^2_{i,k}$. The probabilities for choosing these moves are defined in exactly the same way: $b_{p_{i,k}} + d_{p_{i,k}} + u_{p_{i,k}} = 1$; $d_0 \triangleq 0$, $b_{p_{\max}} \triangleq 0$, otherwise $b_{p_{i,k}} = d_{p_{i,k}} = u_{p_{i,k}}$ for $i = 0, ..., k$. The procedure is performed for each segment and the main steps are described as follows.

Algorithm for the update of the number of poles.

```
1. For $i = 1, \ldots, k$
    (a) If $\left(u_p \sim \mathcal{U}_{(0,1)}\right) \leq b_{p_{i,k}}$ then propose $p'_{i,k} = p_{i,k} + 1$;
        else if $u_p \leq b_{p_{i,k}} + d_{p_{i,k}}$ then propose $p'_{i,k} = p_{i,k} - 1$;
        else goto $(d)$.
    (b) If $\left(u_{p_d} \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{(p_{i,k} \to p'_{i,k})}$ (see Eq. (67)) then the new state of the MC becomes
        $\left(k, \boldsymbol{\tau}_k, \{\mathbf{p}_{1:i-1,k}, p'_{i,k}, \mathbf{p}_{i+1:k,k}\}, \boldsymbol{\delta}^2_k, \lambda, \theta, \gamma_0, \beta_\delta\right)$
        otherwise it stays $\left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}^2_k, \lambda, \theta, \gamma_0, \beta_\delta\right)$
    (c) Sample $\sigma^2_{i,k}\Big| \left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}^2_k, \mathbf{x}_{0:T-1}\right)$ see Eq. (56);
        sample $\mathbf{a}_{i,k}\Big| \left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \boldsymbol{\delta}^2_k, \mathbf{x}_{0:T-1}, \sigma^2_{i,k}\right)$ see Eq. (57);
        sample $\delta^2_{i,k}\Big| \left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \mathbf{x}_{0:T-1}, \mathbf{a}_{i,k}, \sigma^2_{i,k}\right)$ see Eq. (71).
2. Propose $\theta'\big| (k, \mathbf{p}_k)$ (see Eq. (69))
    if $\left(u_\theta \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_\theta$ (see Eq. (70)) then $\theta = \theta'$.
3. Sample $\lambda\big| (k, \boldsymbol{\tau}_k, \mathbf{x}_{0:T-1})$ see Eq. (72).
```

4. Sample $\gamma_0 | \left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \mathbf{x}_{0:T-1}, \boldsymbol{\sigma}_k^2\right)$ see Eq. (72).

5. Sample $\beta_\delta | \left(k, \boldsymbol{\tau}_k, \mathbf{p}_k, \mathbf{x}_{0:T-1}, \{\mathbf{a}_{i,k}^{p_{i,k}}\}_{i=0,\ldots,k}, \boldsymbol{\sigma}_k^2, \boldsymbol{\delta}_k^2\right)$ see Eq. (72).

■

The acceptance probability for the different types of moves (in terms of the number of poles) is given by:

$$\alpha_{(p_{i,k}\to p'_{i,k})} = \min\left\{1, r_{(p_{i,k}\to p'_{i,k})}\right\}, \tag{67}$$

where from Eq. (54)

$$r_{(p_{i,k}\to p'_{i,k})} = \frac{\left|\mathbf{M}_{i,k}^{(p'_{i,k})}\right|^{\frac{1}{2}} \frac{\theta^{p'_{i.k}}}{p'_{i,k}!} \delta_{i,k}^{-\frac{p'_{i,k}}{2}} \left(\gamma_0 + \mathbf{x}^{\mathrm{T}}_{\tau_{i,k}:\tau_{i+1,k}-1} \mathbf{P}_{i,k}^{(p'_{i,k})} \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}\right)^{-\frac{v_0+\tau_{i+1,k}-\tau_{i,k}}{2}}}{\left|\mathbf{M}_{i,k}^{(p_{i,k})}\right|^{\frac{1}{2}} \frac{\theta^{p_{i.k}}}{p_{i,k}!} \delta_{i,k}^{-\frac{p_{i,k}}{2}} \left(\gamma_0 + \mathbf{x}^{\mathrm{T}}_{\tau_{i,k}:\tau_{i+1,k}-1} \mathbf{P}_{i,k}^{(p_{i,k})} \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}\right)^{-\frac{v_0+\tau_{i+1,k}-\tau_{i,k}}{2}}} \tag{68}$$

Thus, for the birth move ($p_{i,k} \to p_{i,k}+1$) the acceptance ratio is $\alpha^p_{birth} = \min\{1, r_{birth}\}$, where $r_{birth} = r_{(p_{i,k}\to p_{i,k}+1)}$. Assuming that the current number of poles is $(p_{i,k}+1)$, one obtains the acceptance ratio for the death move $(p_{i,k}+1 \to p_{i,k})$ as $\alpha^p_{death} = \min\left\{1, r_{birth}^{-1}\right\}$. Thus, the birth/death moves are, indeed, reversible.

Taking into account that the series representation of the exponential function is $\exp(\theta) = \sum_{p=0}^{\infty} \frac{\theta^p}{p!}$, we adopt the following proposal distribution for the parameter θ :

$$\mathcal{G}\big(\theta;\, \zeta + \textstyle\sum_{i=0}^{k} p_{i,k},\, \varkappa + (k+1)\big) \tag{69}$$

and sample θ according to a Metropolis-Hastings step with the acceptance probability equal to:

$$\alpha_\theta = \left[\frac{\sum_{p=0}^{p_{\max}} \theta^p}{\sum_{p=0}^{p_{\max}} (\theta')^p} \frac{\exp(-\theta)}{\exp(-\theta')}\right]^{(k+1)}. \tag{70}$$

The hyperparameters $\delta^2_{i,k}$ are sampled using a standard Gibbs move in exactly the same way as described in Section 5.2.1.3.1:

$$\mathcal{IG}\left(\delta^2_{i,k};\, \alpha_\delta + \frac{p_{i,k}}{2},\, \beta_\delta + \frac{\mathbf{a}_{i,k}^{(p_{i,k})\mathrm{T}} \mathbf{a}_{i,k}^{(p_{i,k})}}{2\sigma^2_{i,k}}\right) \tag{71}$$

Similarly, we sample $\beta_\delta, \lambda, \gamma_0$ according to:

$$\begin{gathered}
\mathcal{G}a\left(\beta_\delta;\ \alpha_\delta(k{+}1),\ \textstyle\sum_{i=1}^{k} \frac{1}{\delta_{i,k}^2}\right), \\
\mathcal{B}e(\lambda;\ k{+}1,\ T{-}k{-}1), \\
\mathcal{G}a\left(\gamma_0;\ \frac{\nu_0(k{+}1)}{2},\ \frac{1}{2}\textstyle\sum_{i=1}^{k} \frac{1}{\sigma_{i,k}^2}\right).
\end{gathered} \tag{72}$$

5.2.1.4 Simulations

We assess the performance of the segmentation method proposed above by applying it to the synthetic data with $T = 500$ and $k = 5$. The parameters of the AR models $\{\mathbf{a}_{i,5}^{(p_{i,5})}\}_{i=0,\ldots,5}$ and noise variances $\boldsymbol{\sigma}_5^2$, drawn at random, are given in Table 1.

i^{th} *segment*	$\sigma_{i,5}$	$\mathbf{a}_{i,5}^{(p_{i,5})}$			
0	1.6	−2.3000	−2.6675	−1.8437	−0.5936
1	0.8	1.3000	−0.9200	0.2600	
2	1.7	0.8000	−0.5200		
3	0.5	2.0000	−1.6350	0.5075	
4	0.6	−1.7000	−0.7450		
5	1.8	−0.5000	0.6100	0.5850	

Table 1: The parameters of the AR model and noise variance for each segment.

The number of iterations of the algorithm was 10000, which seemed to be sufficient since the histograms of the posterior distribution were stabilized. As was described in Section 5.2.1.2, we adopt the MMAP of $\hat{p}\left(k|\,\mathbf{x}_{0:T-1}\right)$ as a detection criterion and, indeed, find $\hat{k} = 5$ changepoints. Then, for fixed $k = \hat{k}$, the model order for each segment $p_{i,\widehat{k}}$ and the positions of changepoints $\tau_{i,\widehat{k}}$, $i = 1, \ldots, \hat{k}$ are estimated by MMAP. The results are presented in Table 2. In Fig. 10 and 9 the segmented signal and the estimation of the marginal posterior distributions of the number of changepoints $\hat{p}\left(k|\,\mathbf{x}_{0:T-1}\right)$ and their positions $\hat{p}\left(\tau_{i,\widehat{k}}\middle|\,\widehat{k}, \mathbf{x}_{0:T-1}\right)$ are given. Fig. 11 shows the estimates of the marginal posterior distribution of the model order for each signal $\hat{p}\left(p_{i,\widehat{k}}\middle|\,\widehat{k}, \mathbf{x}_{0:T-1}\right)$.

i^{th}segment	0	1	2	3	4	5
$\tau_{i,5}$ (true value)	-	90	160	250	365	430
$\widehat{\tau_{i,\widehat{k}}} = \max \hat{p}\left(\tau_{i,\widehat{k}}\middle|\,\widehat{k}, \mathbf{x}_{0:T-1}\right)$	-	91	162	249	366	434
$p_{i,5}$ (true value)	4	3	2	3	2	3
$\widehat{p_{i,\widehat{k}}} = \max \hat{p}\left(p_{i,\widehat{k}}\middle|\,\widehat{k}, \mathbf{x}_{0:T-1}\right)$	4	3	2	3	2	3

Table 2: Real and estimated values for changepoint and model order.

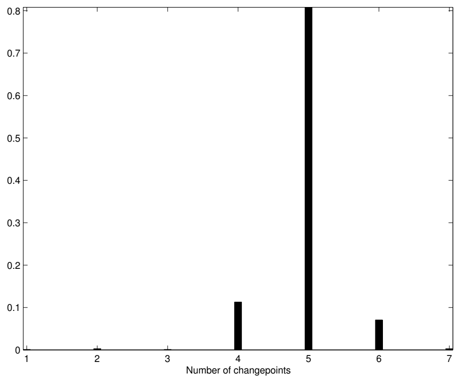

Figure 9: Estimation of the marginal posterior distribution of the number of changepoints $\hat{p}\left(k\middle|\,\mathbf{x}_{0:T-1}\right)$.

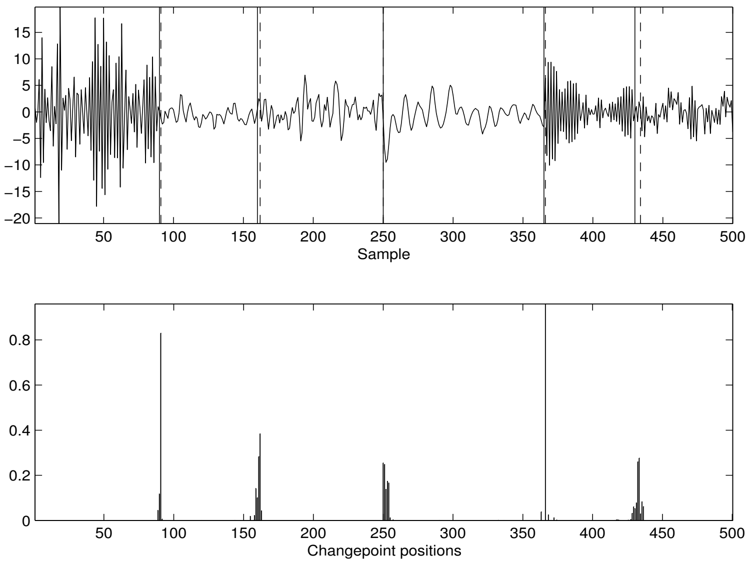

Figure 10: Top: segmented signal (the original changepoints are shown as a solid line, and the estimated changepoints are shown as a dotted line). Bottom: estimation of the marginal posterior distribution of the changepoint positions $\hat{p}\left(\tau_{i,\widehat{k}}\middle|\,\widehat{k},\mathbf{x}_{0:T-1}\right)$, $i = 1, \ldots, \hat{k}$.

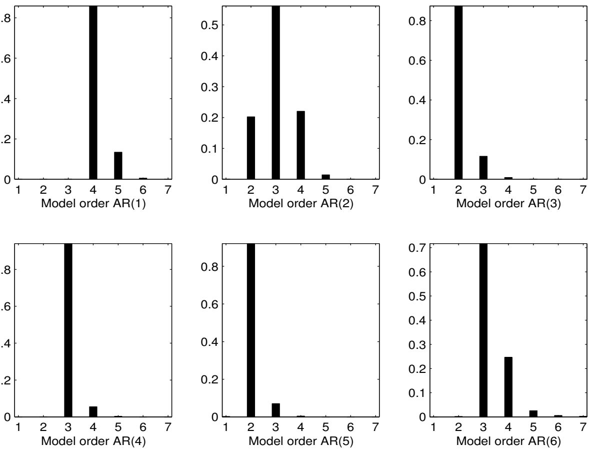

Figure 11: Estimates of the marginal posterior distributions of the number of poles for each segment $\hat{p}\left(p_{i,\widehat{k}}\middle|\,\widehat{k},\mathbf{x}_{0:T-1}\right)$, $i=0,\ldots,\hat{k}$.

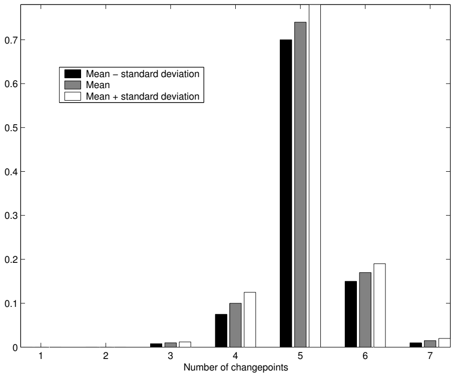

Figure 12: Mean and standard deviation for 50 realizations of the posterior distribution $p\left(k\middle|\,\mathbf{x}_{0:T-1}^{(i)}\right)$.

Then we estimated the mean and the associated standard deviation of the marginal posterior distributions $\left(p\left(k|\,\mathbf{x}_{0:T-1}^{(i)}\right)\right)_{i=1,\ldots,50}$ for 50 realisations of the experiment with fixed model parameters and changepoint positions. The results are presented in Fig. 12 and it is worth noticing that they are very stable regarding the fluctuations of the realization of the excitation noise.

5.2.1.5 Speech Segmentation

In this section we implemented the proposed algorithm for processing a real speech signal which was examined in the literature before (see [1], [7] and [32]). It was recorded inside a car by the French National Agency for Telecommunications for testing and evaluating speech recognition algorithms as described in [7]. According to [32], the sampling frequency was 12 kHz, and a high-pass filtered version of the signal with cut-off frequency 150Hz and the resolution equal to 16 bits is presented in Fig. 13.

Different segmentation methods (see [1], [6], [7], and [32]) were applied to the signal and the summary of the results can be found in [32]. We show these results in Table 3 in order to compare them to the ones obtained using our proposed method (see also Fig. 13 and 14). The estimated orders of the AR models are presented in Table 4 and as one can see they are quite different from segment to segment. This resulted in the different positions of the changepoints, which is especially crucial in the case of the third changepoint. Its position changed significantly due to the estimated model orders for the second ($\hat{p}_{2,5} = 19$) and third segments ($\hat{p}_{3,5} = 27$). As it is illustrated in Fig. 14, the changepoints obtained by the proposed method visually seem to be more accurate.

Method	AR order	Estimated changepoints							
Divergence	16	445	645	1550	1800	2151	2797	-	3626
Brand's GLR	16	445	645	1550	1800	2151	2797	-	3626
Brand's GLR	2	445	645	1550	1750	2151	2797	3400	3626
Approx. ML ([32])	2	445	626	1609	-	2151	2797	-	3627
Proposed method	estimated	448	624	1377	-	2075	2807	-	3626

Table 3: Changepoint positions for different methods.

Segment	0	1	2	3	4	5	6
Model order	6	5	19	27	16	9	11

Table 4: Estimated model orders.

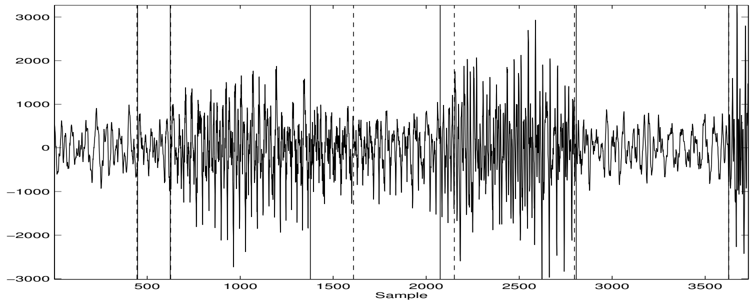

Figure 13: Segmented speech signal (the changepoints estimated by Gustafsson are shown as a dotted line and ones estimated using our proposed method are shown as a solid line).

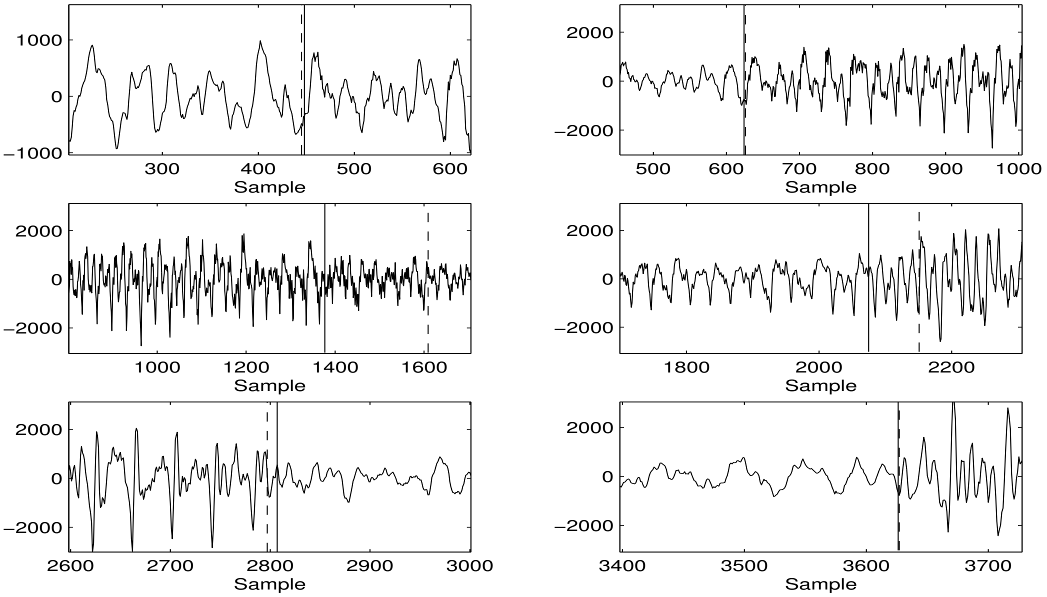

Figure 14: The changepoint positions (the changepoints estimated by Gustafsson are shown as a dotted line and the ones estimated using our proposed method are shown as a solid line).

5.2.1.6 Conclusion

In this section the problem of segmentation of piecewise constant AR processes was addressed. An original algorithm based on a reversible jump MCMC method was proposed, which allows the estimation of the number of changepoints, as well as the estimation of model orders, parameters and noise variances for each of the segments. The results obtained for synthetic and real data confirm the good performance of the algorithm in practice.

In exactly the same way the segmentation of any data which might be described in terms of a linear combination of basis functions with an additive Gaussian noise component (general piecewise linear model, [17], [45]) can be considered. This generalisation of the proposed method is presented in the next section.

5.2.2 General linear changepoint detector

The framework proposed in the previous section is in most cases suitable for the segmentation of any signal in the form of the general linear model with piecewise constant parameters. In this case the possible models $\mathcal{M}_{k,\mathbf{p}_k}$, which might now represent the signal, have exactly the same form as Eq. (31):

$$\mathcal{M}_{k,\mathbf{p}_k}:\quad \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1} = \mathbf{X}_{i,k}^{(p_{i,k})}\mathbf{a}_{i,k}^{(p_{i,k})} + \mathbf{n}_{\tau_{i,k}:\tau_{i+1,k}-1},\quad i=0,\dots,k, \tag{73}$$

where a set of $p_{i,k}$ model parameters ($p_{i,k}=0,\dots,p_{\max}$, $\mathbf{p}_k\triangleq \mathbf{p}_{1:k,k}$) for the i^{th} segment is still arranged in the vector $\mathbf{a}_{i,k}^{(p_{i,k})} = \left(a_{i,k,1}^{(p_{i,k})},\dots,a_{i,k,p_{i,k}}^{(p_{i,k})}\right)^{\mathrm{T}}$ and n_t is i.i.d. Gaussian noise of variance $\sigma^2_{i,k}$ ($\boldsymbol{\sigma}^2_k\triangleq \boldsymbol{\sigma}^2_{1:k,k}$) associated with the i^{th} model. The only difference is the form of the matrix $\mathbf{X}_{i,k}^{(p_{i,k})}$, the example of which for the polynomial model is presented below.

Example 4 ***Polynomial and seemingly non-linear models.*** *The flexibility of the general linear model allows us to detect changepoints in polynomials and other models where the basis functions are not linear, but the model is linear in its coefficients. In this case the observation sequence may be represented as*

$$\mathcal{M}_{k,\mathbf{p}_k}:\quad x_t = \textstyle\sum_{j=1}^{p_{i,k}} a_{i,k,j}^{(p_{i,k})} u_t^{j-1} + n_t \quad \text{for } \tau_{i,k}\le t<\tau_{i+1,k},\quad i=0,\dots,k, \tag{74}$$

which in the generalised form for the i^{th} segment can be rewritten as

$$\begin{bmatrix} x_{\tau_{i,k}} \\ x_{\tau_{i,k}+1} \\ \vdots \\ x_{\tau_{i+1,k}-1} \end{bmatrix} = \begin{bmatrix} 1 & u_{\tau_{i,k}} & u^2_{\tau_{i,k}} & \cdots & u^{p_{i,k}-1}_{\tau_{i,k}} \\ 1 & u_{\tau_{i,k}+1} & u^2_{\tau_{i,k}+1} & \cdots & u^{p_{i,k}-1}_{\tau_{i,k}+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{\tau_{i+1,k}-1} & u^2_{\tau_{i+1,k}-1} & \cdots & u^{p_{i,k}-1}_{\tau_{i+1,k}-1} \end{bmatrix} \begin{bmatrix} a_{i,k,1}^{(p_{i,k})} \\ a_{i,k,2}^{(p_{i,k})} \\ \vdots \\ a_{i,k,p_{i,k}}^{(p_{i,k})} \end{bmatrix} + \begin{bmatrix} n_{\tau_{i,k}} \\ n_{\tau_{i,k}+1} \\ \vdots \\ n_{\tau_{i+1,k}-1} \end{bmatrix}.$$

Remark 2 *In fact, a special case of polynomial models is* ***a piecewise constant source*** *with added Gaussian noise. In this case $p_{i,k} = 1$ and for the i^{th} segment Eq. (73) takes the form:*

$$\begin{bmatrix} x_{\tau_{i,k}} \\ x_{\tau_{i,k}+1} \\ \vdots \\ x_{\tau_{i+1,k}-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [a_{i,k}] + \begin{bmatrix} n_{\tau_{i,k}} \\ n_{\tau_{i,k}+1} \\ \vdots \\ n_{\tau_{i+1,k}-1} \end{bmatrix}. \tag{75}$$

Remark 3 *Since the method allows the estimation of the model order for each segment, a so called* ***sloping-step detector*** *(see Fig. 15 for the type of a signal), which is also a special case of the polynomial model***,** *can be introduced. The model order $p_{i,k}$ may be either 1 or 2 in this case and the matrix $\mathbf{X}_{i,k}^{(p_{i,k})}$ for the i^{th} segment is given by*

$$\mathbf{X}_{i,k}^{(p_{i,k})} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & \tau_{i+1,k} - \tau_{i,k} \end{bmatrix}.$$

Similarly, the equations for other common models, presented in Subsection 3.1.5 are derived.

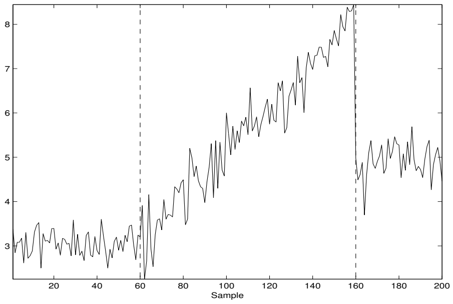

Figure 15: Illustration to a sloping-step detector.

It is worth noticing that the case of a piecewise constant source is more simple as the model order is known. The changes in the algorithm are obvious in this case. There is no need to propose a new model order for different segments, performing the birth/death and update of the changepoint positions, and, taking also into account that $p_{i,k}$ is not a random variable any more, Eq. (61) takes the form:

$$\begin{aligned} r^i_{birth} &= \frac{p\left(k+1,\boldsymbol{\tau}_{k+1},\lambda,\theta,\boldsymbol{\delta}^2_{k+1},\gamma_0,\beta_\delta \middle| \mathbf{x}_{0:T-1}\right)}{p\left(k,\boldsymbol{\tau}_k,\lambda,\theta,\boldsymbol{\delta}^2_k,\gamma_0,\beta_\delta \middle| \mathbf{x}_{0:T-1}\right)} \frac{q\left(k,\boldsymbol{\tau}_k,\boldsymbol{\delta}^2_k \middle| k+1,\boldsymbol{\tau}_{k+1},\boldsymbol{\delta}^2_{k+1},\gamma_0,\beta_\delta,\mathbf{x}_{0:T-1}\right)}{q\left(k+1,\boldsymbol{\tau}_{k+1},\boldsymbol{\delta}^2_{k+1} \middle| k,\boldsymbol{\tau}_k,\boldsymbol{\delta}^2_k,\gamma_0,\beta_\delta,\mathbf{x}_{0:T-1}\right)} \\ &= \frac{\lambda}{(1-\lambda)} \frac{f(\tau_{i,k},\tau,\delta^2_{1i}) f(\tau,\tau_{i+1,k},\delta^2_{2i})}{f(\tau_{i+1,k},\tau_{i,k},\delta^2_{i,k})} \frac{d_{k+1}}{b_k} \frac{(T-2-k)}{(k+1)}, \end{aligned} \tag{76}$$

where

$$f(\tau_{i,k},\tau_{i+1,k},p_{i,k},\delta^2_{i,k}) = \frac{(\gamma_0)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\upsilon_0}{2})}\frac{\beta_\delta^{\alpha_\delta}}{\Gamma(\alpha_\delta)}\left[\frac{\tilde{\beta}_\delta^{\tilde{\alpha}_\delta}}{\Gamma(\tilde{\alpha}_\delta)}\right]^{-1}\exp(-\frac{\beta_\delta - \tilde{\beta}_\delta}{\delta^2_{i,k}})$$
$$\times\Gamma\left(\frac{\upsilon_0+\tau_{i+1,k}-\tau_{i,k}}{2}\right)\left|\mathbf{M}_{i,k}^{(p_{i,k})}\right|^{\frac{1}{2}}\left(\gamma_0+\mathbf{x}^{\mathtt{T}}_{\tau_{i,k}:\tau_{i+1,k}-1}\mathbf{P}_{i,k}^{(p_{i,k})}\mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}\right)^{-\frac{\upsilon_0+\tau_{i+1,k}-\tau_{i,k}}{2}}. \tag{77}$$

Then the parameters of the model are updated according to the algorithm for the update of the number of poles but with the steps 1a, 1b, 1c removed.

In exactly the same way any model with fixed order may be considered.

On the other hand, the case of the ARX model is more complicated since the number of zeroes z as well as the number of poles q is unknown. Thus, the number of possible models $\mathcal{M}_{k,\mathbf{q}_k,\mathbf{z}_k}$ goes up to $k_{\max} \times q_{\max} \times (z_{\max}+1)$, and each of them can be described as

$$\mathcal{M}_{k,\mathbf{q}_k,\mathbf{z}_k}: \quad \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1} = \mathbf{X}_{i,k}^{(q_{i,k},z_{i,k})}\mathbf{a}_{i,k}^{(q_{i,k},z_{i,k})} + \mathbf{n}_{\tau_{i,k}:\tau_{i+1,k}-1}, \quad i = 0,\dots,k,$$

where $\mathbf{a} = \left(\alpha_{i,k,1},\alpha_{i,k,2},\dots,\alpha_{i,k,q_{i,k}},\beta_{i,k,0},\beta_{i,k,1},\dots,\beta_{i,k,z_{i,k}}\right)^{\mathtt{T}}$ and

$$\mathbf{X}_{i,k}^{(q_{i,k},z_{i,k})} =$$
$$\begin{bmatrix} x_{\tau_{i,k}-1} & x_{\tau_{i,k}-2} & \cdots & x_{\tau_{i,k}-q_{i,k}} & u_{\tau_{i,k}} & u_{\tau_{i,k}-1} & \cdots & u_{\tau_{i,k}-z_{i,k}} \\ x_{\tau_{i,k}} & x_{\tau_{i,k}-1} & \cdots & x_{\tau_{i,k}+1-q_{i,k}} & u_{\tau_{i,k}+1} & u_{\tau_{i,k}} & \cdots & u_{\tau_{i,k}+1-z_{i,k}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{\tau_{i+1,k}-2} & x_{\tau_{i+1,k}-3} & \cdots & x_{\tau_{i+1,k}-1-q_{i,k}} & u_{\tau_{i+1,k}-1} & u_{\tau_{i+1,k}-2} & \cdots & u_{\tau_{i+1,k}-1-z_{i,k}} \end{bmatrix}.$$

However, the same technique as was used for the estimation of the number of poles in the AR model could be applied for the estimation of the number of both poles and zeros in ARX model.

5.3 Data fusion for changepoint detection problem

As was described before (see Chapter 2), one of the ways of reducing uncertainty and obtaining more complete knowledge of the state of nature is the fusion of information originating from several sources. Typically, the measurements of different quantities, parameters or variables associated with various features of the state of nature are obtained, and the changepoints (discontinuities) in the data indicate the changes in the physical system (or the state of nature) we are interested in. One of the examples of application of this technique is monitoring changes in a reservoir in oil production. As hydrocarbons (oil or gas) are withdrawn from the reservoir, its behavior is altered and a new production approach may be needed to continue efficient drainage. Repeating measurements with sensors installed permanently in the well or lowered into the well on a cable and comparing them with previous measurements can reveal changes that point the way to a revised production strategy. The key idea here is that the data provided by sensors can be modelled with the aid of parametric models (in which the parameters are subject to changes) and thus the problem can be addressed in the proposed framework.

In this section the problem of detecting and estimating the location of changepoints in such (centralized) multi-sensor systems is considered. It is assumed that all available signals can be represented in the form of the general linear piecewise model (see Subsection 5.3.1) and the way of combining probabilistic information from different sources is described in Subsection 5.3.2, where also the Bayesian model and estimation objectives are formalized. In Subsection 5.3.3 the generalisation of the algorithm for the case of multiple observations is presented, and the performance of this algorithm is illustrated on synthetic data in Subsection 5.3.4, where in addition the failure of one of the sensors is simulated.

5.3.1 Problem Statement

Let $\mathbf{x}_{0:T-1}^{(m)}$ be information source m's set of observations (the number of observations T is the same for all sources[2]). The elements of $\mathbf{x}_{0:T-1}^{(m)}$ may be represented by one of the models $\mathcal{M}_{k,\mathbf{p}_k^{(m)}}^{(m)}$, corresponding to the case when the signal is represented in the form of the general linear model with piecewise constant parameters and k changepoints. The models can be written in the following matrix form:

$$\mathcal{M}_{k,\mathbf{p}_k^{(m)}}^{(m)}: \quad \mathbf{x}_{\tau_{i,k}:\tau_{i+1,k}-1}^{(m)} = \mathbf{X}_{i,k}^{(m)}\mathbf{a}_{i,k}^{(m)} + \mathbf{n}_{\tau_{i,k}:\tau_{i+1,k}-1}^{(m)}, \quad i = 0, \ldots, k, \tag{78}$$

where $\mathbf{a}_{i,k}^{(m)}$ is a vector of $p_{i,k}^{(m)}$ model parameters ($p_{i,k}^{(m)} = 0, \ldots, p_{\max}^{(m)}$, $\mathbf{p}_k^{(m)} \triangleq \mathbf{p}_{1:k,k}^{(m)}$) for the i^{th} segment under the assumption of k changepoints and $\mathbf{n}_{\tau_{i,k}:\tau_{i+1,k}-1}^{(m)}$ is a vector of samples of i.i.d. Gaussian noise. (The associated noise variances are arranged in the vector $\boldsymbol{\sigma}_k^{2(m)} \triangleq \boldsymbol{\sigma}_{1:k,k}^{2(m)}$.)

The number of the available sources is M and all of them contain the information about k ($k = 0, \ldots, k_{\max}$) changepoints denoted $\boldsymbol{\tau}_k \triangleq \boldsymbol{\tau}_{1:k,k}$ ($\tau_{0,k} = 0$ and $\tau_{k+1,k} = T - 1$).

We assume that the number of changepoints k, their positions $\boldsymbol{\tau}_k$ and the associated signal model parameters $\boldsymbol{\Psi}_k^{(m)} \triangleq \left(\mathbf{p}_k^{(m)}, \{\mathbf{a}_{i,k}^{(m)}\}_{i=0,\ldots,k}, \boldsymbol{\sigma}_k^{2(m)}\right)$, $m = 1, \ldots, M$ are unknown. Given $\mathbf{x}_{0:T-1}$, our aim is to estimate $k, \boldsymbol{\tau}_k$ and $\boldsymbol{\Psi}_k^{(m)}$ for $m = 1, \ldots, M$.

[2] In general, the same technique can be used for the case when the number of observations is different for different sources.

5.3.2 Bayesian model and estimation objectives

In our case it is reasonable to assume that the likelihoods from each information source m, that is, $p\left(\mathbf{x}_{0:T-1}^{(m)}\middle| k, \boldsymbol{\tau}_k, \boldsymbol{\Psi}_k^{(m)}\right)$ are independent since the only parameters that the observations have in common are the positions of the changepoints $\boldsymbol{\tau}_k$ and their number k. This approach is known as the Independent Likelihood Pool (see Section 3.2) and Fig. 16 illustrates it for our particular problem. Thus, according to the Bayes' theorem, the posterior is given by:

$$\begin{array}{l} p\left(k, \boldsymbol{\tau}_k, \lambda, \boldsymbol{\Psi}_k^{(1)}, \ldots, \boldsymbol{\Psi}_k^{(M)}, \boldsymbol{\xi}_k^{(1)}, \ldots, \boldsymbol{\xi}_k^{(M)}\middle| \mathbf{x}_{0:T-1}^{(1)}, \ldots, \mathbf{x}_{0:T-1}^{(M)}\right) \propto \\ \quad p\left(k, \boldsymbol{\tau}_k, \lambda, \boldsymbol{\Psi}_k^{(1)}, \ldots, \boldsymbol{\Psi}_k^{(M)}, \boldsymbol{\xi}_k^{(1)}, \ldots, \boldsymbol{\xi}_k^{(M)}\right) \left[\prod_{m=1}^{M} p\left(\mathbf{x}_{0:T-1}^{(m)}\middle| k, \boldsymbol{\tau}_k, \boldsymbol{\Psi}_k^{(m)}\right)\right], \end{array} \tag{79}$$

where the prior distribution has the following hierarchical structure:

$$p\left(k, \boldsymbol{\tau}_k, \lambda, \boldsymbol{\Psi}_k^{(1)}, \ldots, \boldsymbol{\Psi}_k^{(M)}, \boldsymbol{\xi}_k^{(1)}, \ldots, \boldsymbol{\xi}_k^{(M)}\right) = p\left(k, \boldsymbol{\tau}_k\right| \lambda) \, p\left(\lambda\right) \prod_{m=1}^{M} p\left(\boldsymbol{\Psi}_k^{(m)}, \boldsymbol{\xi}_k^{(m)}\right), \tag{80}$$

with $\boldsymbol{\xi}_k^{(m)} = \{\theta^{(m)}, \boldsymbol{\delta}_k^{2(m)}, \gamma_0^{(m)}, \beta_\delta^{(m)}\}$ and

$$\begin{aligned} p\left(\boldsymbol{\Psi}_k^{(m)}, \boldsymbol{\xi}_k^{(m)}\right) \quad &= \prod_{i=0}^{k} \left[p\left(\mathbf{a}_{i,k}^{(m)}\middle| \sigma_{i,k}^{2(m)}, \delta_{i,k}^{2(m)}\right) p\left(\sigma_{i,k}^{2(m)}\middle| \gamma_0^{(m)}\right) p\left(\delta_{i,k}^{2(m)}\middle| \beta_\delta^{(m)}\right)\right] \\ &\times \prod_{i=0}^{k} \left[p\left(p_{i,k}^{(m)}\middle| \theta^{(m)}\right)\right] p\left(\theta^{(m)}\right) p\left(\gamma_0^{m)}\right) p\left(\beta_\delta^{(m)}\right). \end{aligned} \tag{81}$$

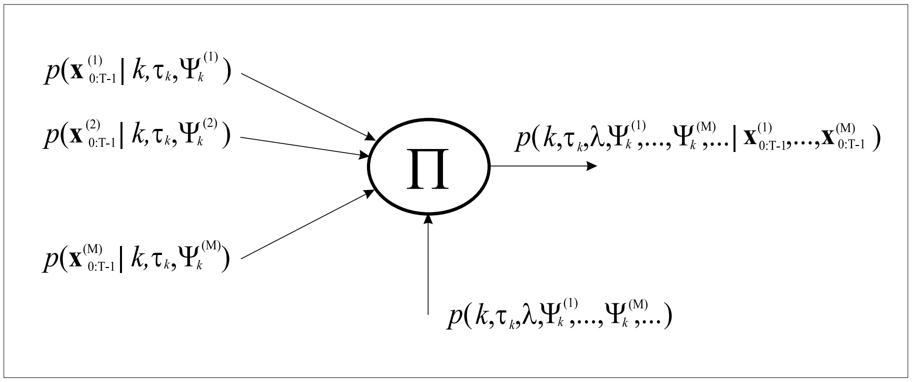
p(x^(1)_0:T-1| k,τ_k,Ψ^(1)_k)
p(x^(2)_0:T-1| k,τ_k,Ψ^(2)_k)
p(x^(M)_0:T-1| k,τ_k,Ψ^(M)_k)
Π
p(k,τ_k,λ,Ψ^(1)_k,...,Ψ^(M)_k,...| x^(1)_0:T-1,...,x^(M)_0:T-1)
p(k,τ_k,λ,Ψ^(1)_k,...,Ψ^(M)_k,...)

Figure 16: Independent Likelihood Pool for changepoint detection problem.

The prior distributions for the model parameters and hyperparameters are assigned in exactly the same way as described in Subsection 5.2.1.2.1 and the likelihoods from each source have the form corresponding to the form of Eq. (44).

The marginal posterior distribution $p\left(k\middle| \mathbf{x}_{0:T-1}^{(1)}, \ldots, \mathbf{x}_{0:T-1}^{(M)}\right)$ should be now estimated using the same method as was used for a single information source (see Section 5.2) in order to perform the model selection (MMAP criterion):

$$\widehat{k} = \underset{k\in\{0,\ldots,k_{\max}\}}{\arg\max}\ \widehat{p}\left(k|\,\mathbf{x}_{0:T-1}^{(1)},\ldots,\mathbf{x}_{0:T-1}^{(M)}\right). \tag{82}$$

Then, having estimated $p\left(\tau_{i,k}|\,\hat{k},\mathbf{x}_{0:T-1}^{(1)},\ldots,\mathbf{x}_{0:T-1}^{(M)}\right), i = 1,\ldots,k$ the estimates of the changepoint positions can be obtained according to the same criterion:

$$\widehat{\tau}_{i,k} = \underset{\tau_{i,k}\in\{1,\ldots,T-1\}}{\arg\max}\ \widehat{p}\left(\tau_{i,k}|\,\widehat{k},\mathbf{x}_{0:T-1}^{(1)},\ldots,\mathbf{x}_{0:T-1}^{(M)}\right),\ i = 1,\ldots,k. \tag{83}$$

Obviously, the parameters $\left(\{\mathbf{a}_{i,k}^{(m)}\}_{i=0,\ldots,k},\boldsymbol{\sigma}_k^{2(m)}\right)$, $m = 1,\ldots,M$ can be integrated out analytically due to the Gaussian noise assumption and the choice of the prior distribution (see Subsection 5.2.1.2.4) and, if necessary, can then be straightforwardly estimated.

5.3.3 MCMC algorithm

The algorithm based on the reversible jump MCMC method, which is presented in Subsection 5.2.1.3 can be easily generalised to the case of multiple sources. Thus the main steps of the algorithms are described as follows.

Reversible Jump MCMC algorithm (main procedure).

1. Initialize $\left(k_{(0)},\boldsymbol{\tau}_{k(0)},\mathbf{p}_{k(0)}^{(1)},\ldots,\mathbf{p}_{k(0)}^{(M)},\lambda_{(0)}^{(1)},\ldots,\lambda_{(0)}^{(M)},\boldsymbol{\xi}_{k(0)}^{(1)},\ldots,\boldsymbol{\xi}_{k(0)}^{(M)}\right)$. Set $j = 1$.
2. Iteration j.

   If $\left(u\sim\mathcal{U}_{(0,1)}\right)\leq b_{k^{(j)}}$ then birth of a new changepoint;

   else if $u\leq b_{k^{(j)}}+d_{k^{(j)}}$ then death of a changepoint;

   else update the changepoints positions.
3. Update of the number of poles.
4. $j\leftarrow j+1$ and goto 2.

■

Algorithm for the birth move $(k\rightarrow k+1)$

- Propose a new changepoint τ in $\{1,\ldots,T-2\}$: $\tau\sim\mathcal{U}_{\{1,\ldots,T-2\}\backslash\{\boldsymbol{\tau}_k\}}$.

- For $m = 1, \dots, M$,

  the proposed model orders are: $p_{1i}^{(m)} = \mathcal{U}_{\left\{1,\dots,p_{oi}^{(m)}\right\}}$, $p_{2i}^{(m)} = p_{oi}^{(m)} - p_{1i}^{(m)}$,

  where $p_{oi}^{(m)} = p_{i,k}^{(m)}$ for $\tau_{i,k} \leq \tau < \tau_{i+1,k}$;

  sample $\delta_{1i}^{2(m)}\Big| \left(\tau_{i,k}, \tau, p_{1i}^{(m)}, \mathbf{x}_{\tau_{i,k}:\tau-1}^{(m)}\right)$, $\delta_{2i}^{2}\Big| \left(\tau, \tau_{i+1,k}, p_{2i}^{(m)}, \mathbf{x}_{\tau:\tau_{i+1,k}-1}^{(m)}\right)$ see Eq. (58).

- Evaluate α_{birth}, see Eq. (84).

- If $\left(u_b \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{birth}$ then the new state of the MC is accepted.

■

Algorithm for the death move $(k+1 \rightarrow k)$

- Choose a changepoint among the $k+1$ existing ones $l \sim \mathcal{U}_{\{1,\dots,k+1\}}$.

- For $m = 1, \dots, M$,

  the proposed model order is $p_{ol}^{(m)} = p_{1l}^{(m)} + p_{2l}^{(m)}$, where $p_{1l}^{(m)} = p_{l-1,k+1}^{(m)}$, $p_{2l}^{(m)} = p_{l,k+1}^{(m)}$;

  sample $\delta_{ol}^{2(m)}\Big| \left(\tau_{l-1,k+1}, \tau_{l+1,k+1}, p_{ol}^{(m)}, \mathbf{x}_{\tau_{l-1,k+1}:\tau_{l+1,k+1}-1}^{(m)}\right)$, see Eq. (58).

- Evaluate α_{death}, see Eq. (84).

- If $\left(u_d \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{death}$ then the new state of the MC is accepted.

■

Algorithm for the update of the changepoint positions

For $l = 1, \dots, k$,

- Propose a new position of the l^{th} changepoint $\tau \sim \mathcal{U}_{\{1,\dots,T-2\}\backslash\{\boldsymbol{\tau}_k\}}$

  determine i such that $\tau_{i,k} < \tau < \tau_{i+1,k}$.

- For $m = 1, \dots, M$,

  If $l \neq i$ then

  $p_{1i} = \mathcal{U}_{\{1,\dots,p_{oi}\}}$, $p_{2i} = p_{oi} - p_{1i}$, where $p_{oi} = p_{i,k}$,

  $p_{ol} = p_{1l} + p_{2l}$, where $p_{1l} = p_{l-1,k}$, $p_{2l} = p_{l,k}$;

  sample $\delta_{1i}^{2}\big| \left(\tau_{i,k}, \tau, p_{1i}, \mathbf{x}_{\tau_{i,k}:\tau-1}\right)$, $\delta_{2i}^{2}\big| \left(\tau, \tau_{i+1,k}, p_{2i}, \mathbf{x}_{\tau:\tau_{i+1,k}-1}\right)$,

  and $\delta_{ol}^{2}\big| \left(\tau_{l-1,k+1}, \tau_{l+1,k+1}, p_{o,l}, \mathbf{x}_{\tau_{l-1,k+1}:\tau_{l+1,k+1}-1}\right)$, see Eq. (58);

  else sample $\delta_{1l}^{2}\big| \left(\tau_{l-1,k}, \tau, p_{l-1,k}, \mathbf{x}_{\tau_{l-1,k}:\tau-1}\right)$, $\delta_{2l}^{2}\big| \left(\tau, \tau_{l+1,k}, p_{l,k}, \mathbf{x}_{\tau:\tau_{l+1,k}-1}\right)$, see Eq. (58).

- Evaluate α_{update}, see Eq. (85).
- If $\left(u_u \sim \mathcal{U}_{(0,1)}\right) \leq \alpha_{update}$ then the new state of the MC is accepted.

■

The algorithm for the update of the number of poles for each model is the same as presented in Subsection 5.2.1.3.3.

The acceptance probabilities for the birth/death and update of the changepoint positions moves are still given by Eq. (60), (65):

$$\alpha_{birth} = \min\left\{1, r_{birth}\right\} \text{ and } \alpha_{death} = \min\left\{1, r_{birth}^{-1}\right\}, \tag{84}$$

$$\alpha_{update} = \min\left\{1, r_{update}\right\}, \tag{85}$$

where however

$$r_{birth}^{i} = \frac{\lambda}{(1-\lambda)}\frac{d_{k+1}}{b_k}\frac{(T-2-k)}{(k+1)}\prod_{m=1}^{M}\left[\frac{f(\tau_{i,k},\tau,p_{1i}^{(m)},\delta_{1i}^{2(m)})f(\tau,\tau_{i+1,k},p_{2i}^{(m)},\delta_{2i}^{2(m)})(p_{oi}^{(m)}+1)}{f(\tau_{i+1,k},\tau_{i,k},p_{i,k}^{(m)},\delta_{i,k}^{2(m)})}\right], \tag{86}$$

$$\begin{aligned} r_{update} &= \prod_{m=1}^{M}\frac{f(\tau_{l-1,k},\tau,p_{l-1,k}^{(m)},\delta_{1l}^{2(m)})f(\tau,\tau_{l+1,k},p_{l,k}^{(m)},\delta_{2l}^{2(m)})}{f(\tau_{l-1,k},\tau_{l,k},p_{l-1,k}^{(m)},\delta_{l-1,k}^{2(m)})f(\tau_{l,k},\tau_{l+1,k},p_{l,k}^{(m)},\delta_{l,k}^{2(m)})}, & \text{if } l = i, \\ r_{update} &= r_{birth}^{i} r_{death}^{l}, & \text{if } l \neq i. \end{aligned} \tag{87}$$

The sampling of the model parameters and hyperparameters is described in detail in Subsection 5.2.1.3.

5.3.4 Simulations

The proposed algorithm was applied to the synthetic data with $T = 500$, $k = 5$ and the number of the available sources was equal to $M = 3$. The signals are presented in Fig (17) and the parameters of the first two models drawn at random are presented in Table 5**.** The third (AR) model has the same parameters as the one tested in Subsection 5.2.1.4 (see Table 1).

The number of iterations was 10000, and using MMAP as a detection criterion one finds $\hat{k} = 5$. The results are presented in Fig. 17 and 18 and in Table 6.

Then the following experiment was carried out in order to assess the performance of the algorithm in the case of a sensor failure. It was assumed that the first sensor failed to observe the last (in fact, existing) changepoint for some reason (see Fig. 19), whereas the other signals remained absolutely the same (see Table 1, 5 for the parameters of the second

Model 1			Model 2		
piecewise constant			*sloping-step*		
i^{th}*segment*	$\sigma_{i,5}$	$\mathbf{a}_{i,5}^{(p_{i,5})}$	$\sigma_{i,5}$	$\mathbf{a}_{i,5}^{(p_{i,5})}$	
0	0.5	3.0	0.5	6.0	–
1	1.3	2.0	1.3	6.0	–
2	0.9	6.0	0.9	8.0	0.05
3	0.5	8.0	0.6	8.0	−0.03
4	0.6	6.0	1.0	6.0	−0.02
5	1.8	3.5	0.4	5.0	–

Table 5: The parameters of the first and second models and noise variances for each segment.

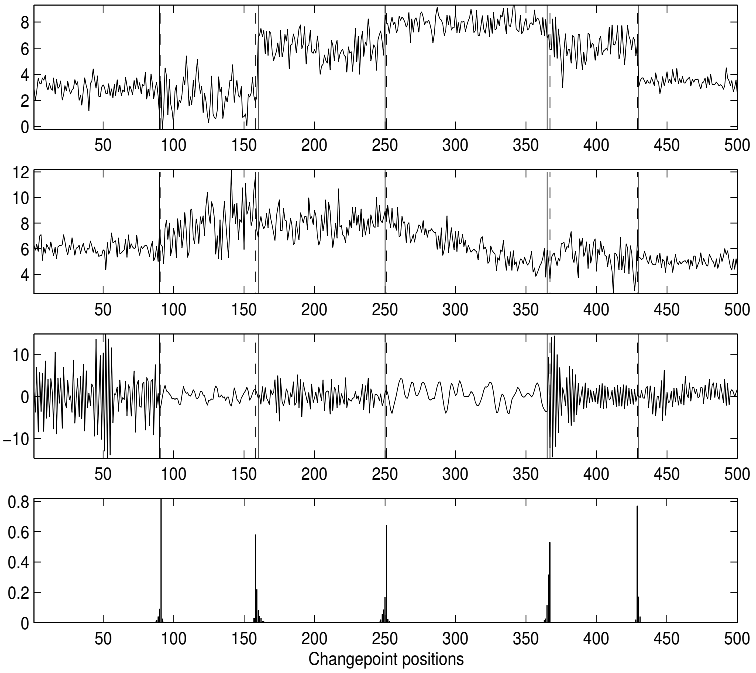

Figure 17: Signal from each source (the original changepoints are shown as a solid line, and the estimated changepoints are shown as a dotted line) and estimation of the marginal posterior distribution of the changepoint positions $p\left(\tau_{i,k} \middle| \widehat{k}, \mathbf{x}_{0:T-1}^{(1)}, \ldots, \mathbf{x}_{0:T-1}^{(M)}\right)$, $i = 1, \ldots, \hat{k}$.

i^{th}segment	0	1	2	3	4	5
$\tau_{i,5}$ (true value)	-	90	160	250	365	430
$\widehat{\tau_{i,\widehat{k}}} = \max \hat{p}\left(\tau_{i,\widehat{k}} \middle| \hat{k}, \mathbf{x}_{0:T-1}\right)$	-	91	158	251	367	429

Table 6: Real and estimated positions of changepoints.

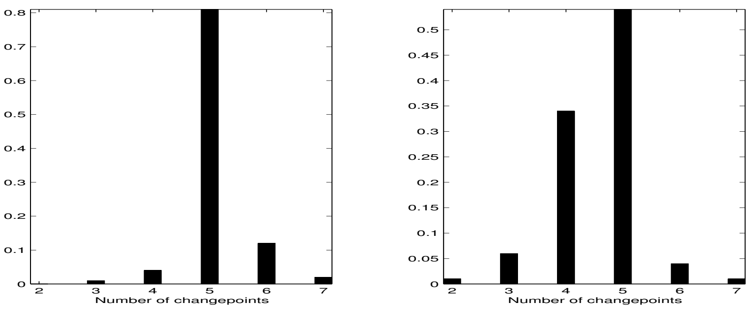

Figure 18: Estimation of the marginal posterior distribution of the number of changepoints $p\left(k|\,\mathbf{x}_{0:T-1}^{(1)},\ldots,\mathbf{x}_{0:T-1}^{(M)}\right)$ for the first (left) and second (right) experiments.

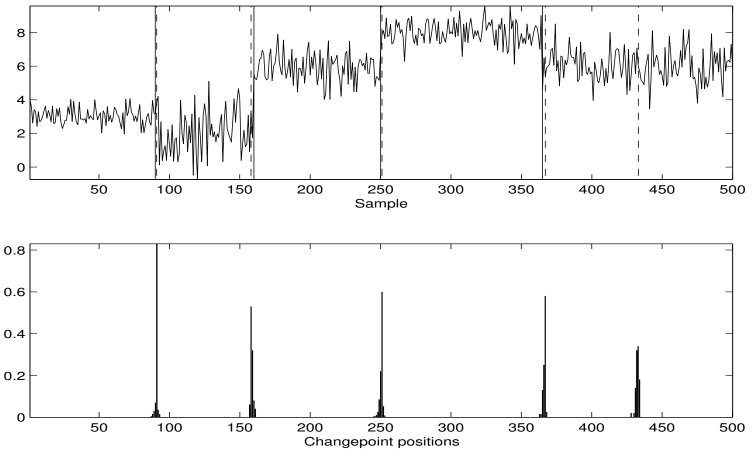

Figure 19: Signal from the first sourse in the case of a sensor failure (the original changepoints are shown as a solid line, and the estimated changepoints are shown as a dotted line) and estimation of the marginal posterior distribution of the changepoint positions $p\left(\tau_{i,k}|\,\widehat{k},\mathbf{x}_{0:T-1}^{(1)},\ldots,\mathbf{x}_{0:T-1}^{(M)}\right)$, $i=1,\ldots,\hat{k}$.

i^{th}segment	0	1	2	3	4	5
$\tau_{i,5}$ (true value)	-	90	160	250	365	430
$\widehat{\tau_{i,\widehat{k}}} = \max \hat{p}\left(\tau_{i,\widehat{k}}\middle|\,\widehat{k},\mathbf{x}_{0:T-1}\right)$	-	91	158	251	366	433

Table 7: Real and estimated positions of changepoints for the case of a sensor failure.

and third signals and Fig. 17 for their form). As one can see from Fig. 19 and Table 7 the results for the estimated number of changepoints and their positions are very similar to the ones obtained in the previous experiment.

5.3.5 Conclusion

In this section the proposed algorithm was applied to address the problem of multi-sensor retrospective changepoint detection. The results obtained for the synthetic data demonstrate the efficiency of this method and the case of a sensor failure was simulated in order to demonstrate the effectiveness of this approach.

6 CONCLUSIONS AND FURTHER RESEARCH

6.1 Conclusions

This dissertation has explored the application of Bayesian techniques and Markov chain Monte Carlo methods to the task of fusion information originating from several sources in the example of a retrospective changepoint detection problem.

Firstly, the use of observations from a single source was considered and some contributions to MCMC model selection were made along the way. In particular, the problem of optimal segmentation of signals modelled as piecewise constant autoregressive (AR) processes excited by white Gaussian noise was addressed. An original Bayesian model was proposed in order to perform so called "double model selection," where the number of segments as well as the model orders, parameters and noise variances for each of them were regarded as unknown parameters. Then an efficient reversible jump MCMC algorithm was developed to overcome the intractability of analytic Bayesian inference. In addition, in order to increase robustness of the prior, the estimation of the hyperparameters was performed, whereas they were usually tuned heuristically by the user in other methods [32], [37]. The method was applied to the speech signal examined in the literature before and the results for both synthetic and real data demonstrate the efficiency of this method and confirm the good performance of both the model and the algorithm in practice.

The approach was then extended such that segmentation of any data which might be described in terms of a linear combination of basis functions with an additive Gaussian noise component (general piecewise linear model) can be considered. The strength of this algorithm is its flexibility with one algorithm for multiple simple steps, ramps, autoregressive changepoints, polynomial coefficient changepoints and changepoints in other piecewise linear models.

Finally, the proposed method was applied to address the problem of multi-sensor retrospective changepoint detection and the effectiveness of this approach was illustrated on the synthetic data.

6.2 Further research

There are several possible extensions to this work, which are discussed in this section.

6.2.1 Application to different signal models

6.2.1.1 Non-linear time series models

We have so far assumed that the observed signals can be described as a linear combination of basis functions with an additive noise component. However, in practice, in a variety of applications one is concerned with data which in fact cannot be represented by linear models. A number of possible model structures, such as non-linear autoregressive, Volterra input-output and radial basis function models, are capable of reflecting this non-linear relationship, and can be expressed in the form of a Linear in The Parameters (LITP) Model. Thus, the technique proposed for detecting and estimating the locations of changepoints using the general linear models can be easily transferred to the case of these non-linear systems.

6.2.1.2 Mixed models

It may well turn out that changepoints divide the sequence into segments with signals of completely different structures (models). To some extent this problem can already be solved in the proposed framework. For example, an AR process may be replaced by an ARX process, multiple steps can become a polynomial sequence and the changes between segments with any signal (of whatever model), and segments with only noise can be detected. However, it will be ideal to develop a general method suitable for addressing the challenging task of finding the changepoints from one model type of any kind into another one of an absolutely different structure.

6.2.1.3 Time delays

It might also happen that the changes in the state of nature are not reflected in the signals from some (or all available) sources at the same time as they occur. Thus, another possible enhancement to the proposed method would be to take such observation time delays into account.

6.2.2 Non-Gaussian noise assumption

As it was described in Subsection 3.1.3.1, statistical inferences frequently make a Gaussian assumption about the underlying noise statistics. However, there are cases where the overall noise distribution is determined by a dominant non-Gaussian noise, and an assumption which does not agree with reality can hardly be desirable.

The difficulty, traditionally associated with the non-Gaussian noise model is analytically intractable integrals. Therefore, if one wants to perform Bayesian inference in this case, it is necessary to numerically approximate these integrals. This problem can be certainly solved by using stochastic algorithms based on MCMC methods, and the algorithm proposed above can also be adapted to address the problem of detecting and estimating the locations of changepoints in non-Gaussian noise environments.

6.2.3 On-line changepoint detection

In a large number of applications it is necessary to recognise the changes in a certain state of nature sequentially while the measurements are taken. For example, in the problem of quality control the changepoints are associated with the situation when the process leaves the *in control* condition and enters the *out of control* state. In such conditions, the quickest detection of the *disorder* with as few *false alarms* as possible might be a question of quality of the production or even safety of a technological process. The similar problem of *on-line* changepoint detection arises in monitoring of industrial processes and in seismic data processing (when the seismic waves should be identified and detected on-line). In all these cases, the observations from several sources are available and the information provided by each source should be combined. It is certainly of great interest to develop a method capable of solving this problem and, in the author's opinion, this topic is an important subject for future research.

References

[1] R. Andre-Obrecht, “A new statistical approach for automatic segmentation of continuous speech signals,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, pp. 29-40, 1988.

[2] C. Andrieu, A. Doucet, S.J. Godsill, and W.J. Fitzgerald, “An introduction to the theory and applications of simulation-based computational methods in Bayesian signal processing,” technical Report , Univ. Cambridge, Dept. Engineering, CUED/F-INFENG/TR 324, 1998.

[3] D. Barry and J. Hartigan, “A Bayesian analysis for changepoint problems,” *J. Am. Stat. Assoc.*, vol. 88, pp. 309-319, 1993.

[4] Y. Bar-Shalom and T.E. Fortmann, *Tracking and Data Association,* Academic Press, 1988.

[5] B. Barshan and H.F. Durrant-Whyte,“Inertial navigation system for a mobile robot,” In *1st IFAC International Workshop on Intelligent Autonomous Vehicles, Southampton, U.K.*, 1993.

[6] M. Basseville and A. Benveniste, “Design and comparative study of some sequential jump detection algorithms for digital signals,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 31, pp. 521-535, 1983.

[7] M. Basseville and I.V. Nikiforov, *Detection of Abrupt Changes: Theory and Application,* Prentice Hall, Information and system science series, 1993

[8] T. Bayes, “An essay towards solving a problem in the doctrine of chances,” *Biometrika*, vol. 45, pp. 293-315, 1958. Reproduced with permission of the Council of the Royal Society from *Philosophical Transactions of the Royal Society of London,* vol. 53, pp. 370-418, 1763.

[9] J.O. Berger, *Statistical Decisions (second edition)*, Springer-Verlag, Berlin, 1985.

[10] J.O. Berger, “The robust Bayesian viewpoint (with discussion),” In *Robustness of Bayesian analysis*, J. Kadane (Ed.), North-Holland, Amsterdam, 1984.

[11] J.O. Berger, “Robust Bayesian analysis: sensitivity to the priors,” *J. Statist. Plann. Inference*, vol. 25, pp. 303-328, 1990.

[12] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, Wiley, 1994.

[13] J. Besag, P.J. Green, D. Hidgon and K. Mengersen, “Bayesian computation and stochastic systems,” *Statistical Science,* vol. 10, pp. 3-66, 1995.

[14] G.E.P. Box, G.M. Jenkins and G.C. Reinsel, *Time Series Analysis, Forecasting and Control*, Prentice Hall, 1994.

[15] C. Chong, “Hierarchical estimation,” In *2nd MIT/ONR CCC Workshop,* Monterey CA, 1979.

[16] P. Dellaportas, J.J. Forster and I. Ntzoufras, “On Bayesian model and variable selection using MCMC,” paper based upon a talk presented at the HSSS Workshop on Variable Dimension MCMC, New Forest, September 1997.

[17] D. Denison, B. Mallick and A.F.M. Smith, “Automatic Bayesian Curve Fitting,” *J. Roy. Stat. Soc. B*, vol. 60, pp. 333-350, 1998.

[18] P. Djurić, “Segmentation of nonstationary signals,” in *Proc. Conf. IEEE ICASSP’92*, pp. 161–164, 1992.

[19] P. Djurić, “A MAP solution to off-line segmentation of signals,” in *Proc. Conf. IEEE ICASSP’94*, pp. 505-508, 1994.

[20] H.F. Durrant-Whyte, “Consistent integration and propagation of disparate sensor information,” *Int. J. Robotics and Research*, vol. 6(3), pp. 3-24, 1987.

[21] H.F. Durrant-Whyte, “Sensor models and multi-sensor integration,” *Int. J. Robotics and Research*, vol. 7(6), pp. 97-113, 1988.

[22] H.F. Durrant-Whyte, “Uncertain geometry in robotics,” *IEEE J. Robotics and Automation*, vol. 4(1), pp. 23-31, 1988.

[23] H.F. Durrant-Whyte, B.Y. Rao and H. Hu“Toward a fully decentralized architecture for multi-sensor data fusion,” In *Proc. IEEE Int. Conf. Robotics and Automation*, pp. 1331-1336, 1990.

[24] B.J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.

[25] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images,” *IEEE Trans. Patt. Ana. Mac. Int.,* vol. 6, pp. 721-741, 1984.

[26] S.J. Godsill, “On the relationship between MCMC model uncertainty methods,” technical report, Univ. Cambridge, Dept. Engineering, CUED-F-INFENG/TR.305, 1997.

[27] S.J. Godsill, "The restoration of degraded audio signals", PhD thesis, Univ. Cambridge, Dept. Engineering, 1993.

[28] I.R. Goodman, R.P.S. Mahler, and H.T. Nguyen, *Mathematics of Data Fusion,* Kluwer Academic Publishers, London, 1997.

[29] P.J. Green, "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711-732, 1995.

[30] F. Gustafsson, "Optimal segmentation in a linear regression framework," in *1991 Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1677-1680, Toronto, Canada,1991.

[31] F. Gustafsson, "The marginalized likelihood ratio test for detecting abrupt changes," *IEEE Trans. Automatic control*, vol. 41, pp. 66-77, 1996.

[32] F. Gustafsson, "Segmentation of signals using piecewise constant linear regression models," to appear *IEEE Trans. Signal Processing*, 1999.

[33] D.L. Hall, *Mathematical Techniques in Multi-sensor Data Fusion,* Artech House, Boston, 1992.

[34] H.R. Hashemipour, S. Roy and A.J. Laub, "Decentralized structures for parallel Kalman filtering," *IEEE Trans. Automatic Control,* vol. 33(1), pp.88-93, 1988.

[35] D. Heckerman, "A tutorial on learning with Bayesian networks," technical report, Microsoft Research, Advanced Technology Devision, MSR-TR-95-06, 1995.

[36] H. Jeffreys, *Theory of Probability,* Oxford University Press, 3rd edn.,1961.

[37] M. Lavielle, "Optimal segmentation of random processes", *IEEE Trans. Signal Processing*, vol. 46, pp. 1365-1373, 1998.

[38] S.L. Lauritzen and N. Wermuth, "Graphical models for associations between variables, some of which are qualitative and some quantitative," *Ann. Stat,* vol. 17, pp. 31-57, 1989.

[39] S.P. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[40] J. Manyika, I.M. Treherne and H. Durrant-Whyte, "A modular architecture for decentralized sensor data fusion. A sonar-based sensing node," In *IARP 2nd Workshop on Sensor Fusion and Environmental Modeling*, 1991.

[41] J. Manyika and H. Durrant-Whyte, "A tracking sonar for vehicle guidance," In *Proc. IEEE Robotics and Automation*, 1993.

[42] J. Manyika and H. Durrant-Whyte, *Data Fusion and Sensor Management: A Decentralized Information-Theoretic Approach,* Ellis Horwood, New-York, London, 1994.

[43] R. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," technical report, Univ. Toronto, Dept. Computer Science, CRG-TR-93-1, 1993.

[44] N.E. Orlando, "An intelligent robotics control scheme," In *American Control Conference,* pp. 204, 1984.

[45] J.J.K. O'Ruanaidh and W.J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing,* Springer-Verlag, 1996.

[46] J. Pearl, *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*, Morgan Kaufmann, San Mateo CA, 1988.

[47] B. Rao, H. Durrant-Whyte and A. Sheen, "A fully decentralized multi-sensor system for tracking and surveillance," *Int. J. Robotics Research,* vol. 12(1), pp. 20-45, 1991.

[48] S. Richardson and P.J. Green, "On Bayesian analysis of mixtures with unknown number of components," *J. Roy. Stat. Soc. B*, vol. 59, no. 4, pp. 731-792, 1997.

[49] C.P. Robert, *The Bayesian Choice. A Decision- Theoretic Motivation*, Springer Texts in Statistics, Springer-Verlag, 1994.

[50] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag Series in Statistics, 1999.

[51] R.C. Smith and P. Cheesman, "On the representation of spatial uncertainty," *Int. J. Robotics Research,* vol. 5(4), pp. 56-68, 1987.

[52] J.A. Stark, W.J. Fitzgerald, S.B. Hladky, "Multiple-order Markov chain Monte Carlo sampling methods with application to a changepoint model," technical report, Univ. Cambridge, Dept. Engineering, CUED-F-INFENG/TR.302, 1997.

[53] D.A. Stephens, "Bayesian retrospective multiple-changepoint identification," *Applied Statistics*, vol. 43, pp. 159-178, 1994.

[54] M. Stone, "The opinion pool," *Ann. Stat.,* vol. 32, pp. 1339-1342, 1961.

[55] L. Tierney, "Markov chains for exploring posterior distributions," (with discussion) *Ann. Stat.*, pp. 1701-1762, 1994.

[56] J.N. Tsitsikis and M. Athans, "On the complexity of decentralised decision-making and detection problem," *IEEE Trans. Automatic Control,* vol. 30(5), pp. 440-446, 1985.

[57] P.K. Varshney, *Distributed detection and data fusion,* Springer, New-York, 1997.

[58] E.L. Waltz and J. Llinas, *Multi-Sensor Data Fusion,* Artech House, 1991.

[59] L. Wasserman, "Recent methodological advances in robust Bayesian inference," In *Bayesian Statistics* 4, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (Eds.), pp.483-490, Oxford University Press, London, 1992.

[60] A.S. Willsky, M.G. Bello and D.A. Caston, "Combining and updating of local estimates and regional maps along sets of one-dimensional tracks," *IEEE Trans. Automatic Control,* vol. 27(4), pp. 799-812, 1982.

[61] A.S. Willsky, H.L. Jones, "A generalised likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Trans. Automatic Control*, pp. 108-112, 1976.