University of Potsdam
Institute for Mathematics
Faculty of Science
Data Assimilation and Statistics

Master Thesis for the Degree M.Sc. Mathematics

Towards Robust Inference for Bayesian Filtering of Linear Gaussian Dynamical Systems Subject to Additive Change

by

Hans Reimann
Matrikel-Nr.: 792612

Supervisor:

Prof. Dr. Ing. Sebastian Reich
Prof. Dr. Alexandra Carpentier

Potsdam, 01.06.2024

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-64946
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-649469

Noise Mis-specification in Bayesian Filtering: A Perspective via Robust Inference for Bayesian Filtering of Dynamical Systems Subject to Change

Author:

Hans Reimann

Acknowledgments

It is an astonishing observation that there is very little in my studies of mathematics that has not been required during work on this thesis in one way or another. Hereby, especially my classes in mathematical statistics and numerics need pointing out. I was lucky to have highly knowledgeable lecturers in these fields and even more so that two of them agreed to become my supervisors for this master's thesis. I want to thank Prof Dr Sebastian Reich and Prof Dr Alexandra Carpentier for their support and valuable advice in the course of my work. Both in their teaching and supervision they helped me feel like mathematics is a, relevant, accessible and interesting topic with challenging, yet also fun and interesting tasks. Especially, after other lecturers discouraged me and advised me not to pursue my degree doubting my proficiency after early pitfalls during my master studies.

Similarly, I want to thank my lecturers and internship advisors at Macquarie University in Sydney - Benoit, Sarat and Georgy. All of you helped me shape the topic of this thesis with your teaching and provided me with essential tools. More important however, you also made me feel like I could achieve interesting insights and contributions in statistics with my work instead of being not good enough.

My studies were accompanied by several other students who helped me through valleys in motivation and frustrating experiences as well as to celebrate and enjoy success. Thank you Alexandra, Matthias, Lea and Sönke. Further, I want to thank the Uni Potsdam research group for Complexity Science, the students in the CRC (SFB) 1294, my Bayesian inference study group and the SIAM student chapter Potsdam for providing me a home during my master studies. Even at the very end of my studies in Potsdam I am still at awe what a fun, social experience my studies in mathematics were and maybe even needed to be to get here.

Abstract

State space models enjoy wide popularity in mathematical and statistical modelling across disciplines and research fields. Frequent solutions to problems of estimation and forecasting of a latent signal such as the celebrated Kalman filter hereby rely on a set of strong assumptions such as linearity of system dynamics and Gaussianity of noise terms.

We investigate fallacy in mis-specification of the noise terms, that is signal noise and observation noise, regarding heavy tailedness in that the true dynamic frequently produces observation outliers or abrupt jumps of the signal state due to realizations of these heavy tails not considered by the model. We propose a formalisation of observation noise mis-specification in terms of Huber's ε-contamination as well as a computationally cheap solution via generalised Bayesian posteriors with a diffusion Stein divergence loss resulting in the diffusion score matching Kalman filter - a modified algorithm akin in complexity to the regular Kalman filter. For this new filter interpretations of novel terms, stability and an ensemble variant are discussed. Regarding signal noise mis-specification, we propose a formalisation in the frame work of change point detection and join ideas from the popular CUSUM algorithm with ideas from Bayesian online change point detection to combine frequent reliability constraints and online inference resulting in a Gaussian mixture model variant of multiple Kalman filters. We hereby exploit open-end sequential probability ratio tests on the evidence of Kalman filters on observation sub-sequences for aggregated inference under notions of plausibility.

Both proposed methods are combined to investigate the double mis-specification problem and discussed regarding their capabilities in reliable and well-tuned uncertainty quantification. Each section provides an introduction to required terminology and tools as well as simulation experiments on the popular target tracking task and the non-linear, chaotic Lorenz-63 system to showcase practical performance of theoretical considerations.

Key Words:
Statistical Model Mis-Specification, Robust Filtering, State Space Change Point Detection, Bayesian Filtering, Bayesian Modelling

Category:
Data Assimilation, Machine Learning, Statistical Methodology, Signal Processing

Contents

Abbreviations and Frequent Notation

The following is an non-exhaustive list providing abbreviations and frequent notation.

iid Independent and identical distributed; frequent assumption in statistical modelling

ARL Average run length; expectation of a stopping time under the null hypothesis

BOCPD Bayesian online change point detection; see (Adams and MacKay, 2007) and (Fearnhead and Liu, 2007)

CADD Conditional average detection delay; see (Pollak, 1985)

CR-BOCPD CUSUM restarted BOCPD; proposed in this work via connecting ideas from R-BOCPD and CUSUM strategies

CUSUM Cumulative sum algorithm; popular approach for sequential change point detection via open-end SPRTS

EnKF Ensemble Kalman filter

FAR False alarm rate or rate of false alarm; reciprocal of ARL; substitute for PFA

FFS Fixed sample size; usually referring to classical hypothesis test settings

GLR Generalized likelohood ratio; usually in combination with the CUSUM rule

GMM Gaussian mixture model; a weighted sum of densities functions of Gaussian random variables

IMQ Inverse multi-quadratic; usually referring to the shape of the kernel in the diffusion weight matrix

innovation The difference between an observation and the forecast observation mean; usually via $\gamma_n = y_n - H_n m_n^f$

KF Kalman filter

LLR Log-likelihood ratio; the logarithm of the fraction of two density functions

PFA Probability of false alarm; probability of supposedly detecting a change when there is non

R-BOCPD Restarted BOCPD; see (Alami et al., 2020)

RV Random vector or random variable depending on dimension

SPRT Sequential probability ratio test; open-end or closed depending on null hypothesis

UBCP Uniformly best constant power; usually referring to a notion of optimality for statistical test with "rich" hypotheses

UMP Uniformly most powerful; usually referring to a notion of optimality for statistical tests with simple hypotheses

WADD Worst average detection delay; see (Lorden, 1971)

α_0 FAR of a detection rule; in $[0, 1]$

α_0^* Type-I error or PFA of a detection rule; in $[0, 1]$

α_1^* Type-II error; in $[0, 1]$

$\hat{w}(y)$ Scalar diffusion weight; see $w(y)$

$\mathbf{1}_{p \times p}$ Identity or unit matrix of dimension $p \times p$

$\mathcal{X}$ Signal state space; usually $\mathcal{X} = \mathbb{R}^d$

$\mathcal{Y}$ Signal state space; usually $\mathcal{Y} = \mathbb{R}^p$

Σ_n Time-varying innovation and marginal observation covariance matrix via $\Sigma_n = H_n P_n^f H_n^T + R_n$

$\Sigma_{i\bullet}$ Row i of the matrix Σ

$\tilde{K}_n(y_n)$ Adjusted Kalman gain matrix balancing forecast and adjusted observation uncertainty

$\{x_n^{(l)}\}_{l \in \{1,2,\ldots,M\}}$ An ensemble with ensemble members $x_n^{(l)}$ at time n and total number of ensemble members M

A_n Time-varying linear signal process operator of the state space model

H_n Time-varying forward or observation map of the state space model

K_n Kalman gain matrix balancing forecast and observation uncertainty

$L_{k,n}$ Cumulative loss of the scenario initialized at time k up to a time n; $L_{k,n} = \sum_{s=k}^{n} l_{k,s}$

$n(x; m, P)$ Multivariate Gaussian random variable with mean vector m and covariance matrix P in covariance form; $X \sim \mathcal{N}(m, P)$

$n^{-1}(x; \theta, J)$ Multivariate Gaussian random variable with potential vector $\theta = Jm$ and precision matrix $J = P^{-1}$ in information or precision form; $X \sim \mathcal{N}(J^{-1}\theta, J^{-1})$

Q_n Time-varying signal noise covariance matrix via $Q_n = C_n C_n^T$

R_n Time-varying observation noise covariance matrix via $R_n = \Gamma_n \Gamma_n^T$

V_n Observation noise, usually standard Gaussian

$w(y)$ Diffusion matrix in diffusion score matching; a point wise invertible matrix valued function $\mathcal{Y}$; $w(y) = \hat{w}(y) R^{\frac{1}{2}}$

W_n Signal noise, usually standard Gaussian

X_n Random vector of the time-varying latent signal state of the dynamical system;

Y_n Random vector of the observation at time n

$\pi_n(\cdot)$ True data generating process at time n and a measure on $\mathcal{Y}$

$\overset{+C}{=}$ Equality up to a constant additional term independent of the variable of interest

$\propto$ Proportionality up to a constant factor independent of the variable of interest

$\mathcal{D}(\pi(\cdot), p(\cdot, x))$ Discrepancy of two measures; usually on $\mathcal{Y}$; with empirical estimator $\hat{\mathcal{D}}$

$\det(P)$ Determinant of the matrix P

$\mathbf{1}\{a = b\}$ Indicator function taking value 1 for true statements and value 0 for false statements

$\mathrm{Tr}(P)$ Trace of the matrix P; the Sum of the diagonal entries

∇f Jacobian matrix of the function f; usually reduces to the gradient, so the vector of the partial derivatives

$\nabla \cdot f$ Divergence operator of the function f; the dot product of the partial derivative operator and the columns of f

$s_p(y)$ Score function via the gradient of the log density function; $s_p(y) = \nabla \log(p(y)$

1. Introduction

Volatile values,
Low-key data masks their grace,
Detection unveils.

The fundamental ideas of this work are rooted in the practice of mathematical and statistical modelling. Both are essential to modern scientific convention and are thus subject to understanding modern science as a social construct (Ritchie, 2020). This is not as in the methods of science and scientific results are socially constructed, but in that scientific progress is a social process. Research enables us to move towards a truth, yet we may never fully get there with knowledge evolving and changing with new realizations. This evolution takes place and is driven in social context - it takes convincing a scientific community in communal review and eventually mutual agreement to progress (Ritchie, 2020).

Scientific models are central to precisely this social aspect of science. Further, the practice of modelling itself is subject to this discourse, even more so for intersecting fields in between theory and application with applied mathematics and statistics at the frontier. Exclamations such as Box's «All models are wrong, but some models are useful» and Wiener's «The best model of a cat is a cat» are both popular and widely debated for their implications and (mis-)interpretations, but regardless provide an insightful glimpse of the diversity of philosophical takes on the matter. These takes directly impact theory and practice in that they utilize the philosophical understanding for interpretation and completeness (see (Gelman and Shalizi, 2013) for details).

The presented work is motivated by one such philosophical take on statistical modelling and its implications, more precisely on frequent interpretations of Bayesian modelling and its theory. In (Gelman and Shalizi, 2013) the authors argue that associating Bayesian statistics with inductive approaches of learning in «the rise and fall of posteriors» as contrary to the hypothetico-deductive interpretation of frequentist methods is wrong - not regarding its inference or theory, but this specific perceived view. Further, they argue that Bayesian approaches are not inherently more inductive than any other approach and they should instead also be understood in a hypothetico-deductive framework. The central criticism of Gelman and Shalizi in (Gelman and Shalizi, 2013) is in the frequently neglected central assumption of Bayesian models in being well-specified - assuming that the model covers a ground truth and that it can be recovered given sufficient information. As important as this assumptions is, it can also only be satisfied or confirmed to very limited extent.

This master thesis is motivated in developing theory and algorithms addressing this criticism in the context of Bayesian filtering of (linear) dynamical systems. Picking up on (Gelman and Shalizi, 2013), we want to address two aspects in which Bayesian filtering may fail under mis-specification and tackle particular errors challenging reliable inference and forecasting. To specify on Box's models being inherently wrong, we want to work towards Gelman's and Shalizi's «[All] models have errors of approximations. Statistical models, however, typically assert that their errors of approximation will be unsystematic and patternless». The work presented and its results propose adapting the popular Kalman filter to maintain patternless in error even under reduced assumptions on the noise terms or substantial threat of mis-specification.

1.1. Problem Statement

Pattern and system in error may plausibly result from mis-specification in statistical models. In Bayesian filtering there are two different sources of stochasticity and accordingly two potential instances of mis-specification - the system noise and the observation noise. Both are frequently assumed to be Gaussian for simplicity as well as lacking better knowledge, resulting in the popular Kalman filter for linear system and observation dynamics (Kalman, 1960). However, for mis-specification in tailedness in that the true data generating process has noise terms with much heavier tails, this quickly results in high potential for volatility and mistakes in inference and especially in forecasting.

Assume a model for a dynamical system and a corresponding model producing observations from that system. In many applications the system dynamics and forward map or observation map can be derived from first principles to some extent, however, the noise terms can then only be estimated experimentally or from prior knowledge as given in (Morzfeld and Reich, 2018). This leads to a somewhat semi-parametric intuition of the setting. The dynamics can be assumed to be fairly accurate or at least good enough in that errors may be assumed patternelss, ideally only carrying negligible inaccuracies of the system dynamics. On the flip-side, the noise terms, i.e. the signal noise and observation noise, therefore carry high potential volatility under mis-specification, i.e. when they still encompass relevant system behavior. It is a fairly debatable practice to force Gaussian noise just to suit the framework of the Kalman filter given linear system dynamics yet it is often most practicable if not the only practical solution from a feasibility perspective.

To go more in depth on this volatility introduced via mis-specified noise terms, it is of major concern when modelling Gaussian noise yet with the true noise generating process being heavy-tailed or prone to outliers. The resulting unexpected behavior impacts inference in filtering and forecasting mainly in two ways.

At the heart of the Bayesian filtering problem is the name-giving Bayes' theorem. While Bayesian inference obeys certain desirable concepts with the likelihood principle and Zellner's information optimality (Zellner, 1988) by utilizing Kullback-Leibler divergence to incorporate newly obtained information, this intuitive choice of loss function is also a major source of volatility regarding observation outliers as it tends to substantially overweight observations subject to mis-specification of the likelihood. The idea proposed is to re-build the Kalman filter with breaking open the involved Bayesian inverse inference problem and utilizing a more robust divergence measure in a generalized Bayesian inference framework. As will be shown in addressing the issue, observation noise mis-specification can generally be understood as likelihood mis-specification regarding variance or shape for the investigated type of problems and is best addressed this way.

Ideally, only after having ensured robustness regarding observation noise mis- specification, we can reliably target mis-specification in the signal noise. The choice of Gaussian noise terms with the true noise being a lot more heavy-tailed will result in sudden jumps and rapid changes of the true signal much more frequent then the assumed noise term may suggest an account for. Accordingly, these jumps are not fully covered by the model, however, may have important implications in practice with popular interpretations being system faults, outside shocks and phase transitions. The tool of choice we employ to detect these instances of impactful signal noise mis-specification is the popular field of sequential change point detection. Inference and forecasting with calibrated uncertainties therefore needs detection of instances of the model failing to adjust in reasonable time.

To summarize, in the popular linear Gaussian Kalman setting the signal and observation noise are subject to potential mis-specification. While signal noise mis-specification in tailedness has meaningful implications for practice, observation noise mis-specification needs being a general concern in Bayesian filtering and may specifically hinders detection of instances of signal mis-specification.

1.2. Contribution

The sketched idea of this thesis is to tackle noise mis-specification in two separate steps to then be combined following a rather basic intuition: Observation noise is hindering, but signal noise is a feature. Accordingly, we will first build provably robust Bayesian inverse inference regarding the observation noise via generalized Bayesian posteriors utilizing weighted score matching as a specific type of diffusion Stein discrepancy. This will be the central result of the first part of this work.

The succeeding second part will investigate sequential probability ratio based detection of signal noise mis-specification instances showing in disruption or jumps at unknown times via CUSUM type sequential change point detection. Either has its own implications for adapted schmes and algorithms with both finally combined into a novel Gaussian mixture model of robust Kalman filters incorporating the previously obtained robust Bayesian inverse inference as well as Gaussian mixtures weighted via plausibility of change at specific time steps.

In other words, the first part provides the method to reduce observation noise influence under mis-specification regrading tailedness and contamination to then enable reliably detecting instances of strong influence of signal noise mis-specification for adjustment. The desired result will be concrete algorithms with supporting theoretical results as adaptations of the popular Kalman filter under considerations of viability of assumptions and computational feasibility in practice. These novel

filters proposes approaches for reliable inference and forecasting under threat of noise mis-specification.
We hereby take recent ideas and results based on Bayesian online change point detection in (Adams and MacKay, 2007) regarding online structure of change point detection, in (Altamirano et al., 2023b) for robust, scalable Bayesian online change point detection and in (Alami et al., 2020) for the restart Bayesian online change point detection procedure and adapt them for the time-varying discrete Bayesian filtering problem with linear dynamics, forward maps and supposed Gaussian noise, so the classical linear Gaussian Kalman filtering setting. This is achieved by deriving recursive formulas with proven robustness to observation noise mis-specification and incorporating them in adapted change detection methods providing uncertainty about change points for inference and forecasting. Related work was done in (Boustati et al., 2020) investigating generalized Bayesian filtering sequential Monte Carlo from a more data science driven perspective and much-less motivated by robustness concerns. The presented master thesis is foremost a stepping stone and careful exploration of combining recent advances in several fields. It will provide novel approaches and results for the larger field of Bayesian filtering under mis-specification as described above, exploring and discussing promising directions for further research such as robust ensemble Kalman methods for noise mis-specification via generalized Bayesian filtering as well as opportunities in open-end sequential probability ratio testing for non-linear, regime-type dynamical systems.

Results in (Boustati et al., 2020) and other recent works on generalized Bayesian inference frequently focused on $\beta-$divergence measures to replace the Kullback-Leibler divergence in classical Bayesian inference. The weighted diffusion score matching Bayesian inference, initially introduced in (Barp et al., 2019) and developed in (Anastasiou et al., 2023) via Stein discrepancies, further adapted and applied in (Altamirano et al., 2023b) and (Altamirano et al., 2023a) provides promising results regarding robustness as well as computational feasibility with the latter being the main downside of β-divergence Bayesian inference. Accordingly, this work aims to utilize these results to develop the described Kalman filter adaptation opening up new directions and connections.
For detection of signal jumps we will mainly focus on adapting recent popular approaches in (Alami et al., 2020) and (Altamirano et al., 2023b) based on (Adams and MacKay, 2007) and (Fearnhead and Liu, 2007) and adapt it to the Kalman setting. To additionally obtain desired properties in reliability via controlled rate of false alarm and detection delay as well as computational feasibility we will utilize classical results such as in (Lai, 1998) on the popular CUSUM procedure. A central result will be in combining ideas from either in the context of Bayesian filtering for a novel approach to change point detection in dynamical systems via conditional evidence. Furthermore, including the previously derived robust posteriors is then also easily incorporated.

To conclude, the contribution of this thesis is not the majority of the body of theory neither the majority of the machinery at work. Much of it is adapted from (Altamirano et al., 2023b), (Alami et al., 2020) and their sources as well as milestone historical results. The central contribution is two-fold: It is the adaptation to the popular

Kalman setting in deriving a closed form solution, providing update formulas, similar to the original Kalman paper (Kalman, 1960), and providing desirable properties such as the global bias robustness in filtering for the first case. Moreover, it is in combining this result as well as results form adjacent fields to obtain the desired inference and forecasting with robustness properties in mis-specification regarding heavy-tails and outliers of the noise terms, that is the Gaussian signal noise and the Gaussian observation noise, assuming the popular Kalman setting. To pick up on the initial motivation, the contributions are in adapting the popular Kalman filter to have calibrated uncertainty for errors of approximation without pattern or system as motivated in (Gelman and Shalizi, 2013) even under mis-specification of the model in challenging aspects, working towards tolerance regarding the fundamental yet only partly feasible assumption of well-specified models. The key enabling tools herein is in bringing together results and arguments from different research communities to enable new, curious perspectives.

1.3. Structure

We aim to foster three layers of understanding with this work. We want everybody with a general interest in mathematical and statistical modelling to gain a basic understanding of the matter and problems presented. Additionally, we want everybody with a background in the natural and engineering sciences to obtain an intuition of the problems as well as their solutions for application. Finally, we want everybody with a background in statistics, mathematics or quantitative research fields to grasp the problems and the proposed solutions, to be convinced by their arguments and to be able to detect connections as well as resulting directions for research in adjacent fields.

The presented master thesis is structured accordingly. We will round up the introduction by providing context and superficial relevance via embedding the investigated problem in the general context of statistical forecasting adn data assimilation. The main body is divided in the two outlined parts already sketched with the additional part bringing both together. Again, the first part will recall the Bayesian inverse inference step in Kalman filtering and derive the diffusion score matching Kalman filter. The second part, investigates connections of classical and modern approaches to sequential change point detection and arising opportunities via Gaussian mixture model filtering evaluating uncertainty of jumps. Major result are presented in the respective sections with the third part then incorporating the results of the previous chapters for reliable detection of signal jumps in the Kalman setting. Each part will provide respective results in concrete algorithms accompanied by theoretical derivations or statements and simulation experiments. The presented work will end in providing a summary of all acquired results and discussing the obtained intuitions and solutions as well as discuss them in their potential regarding the initial context of robust inference and forecasting under potential of abrupt signal jumps.

Each of the three major chapter will have a fairly closed structure on its own. For the first part on observation noise mis-specification we will start with shortly recapping the popular Kalman formula to highlight vital steps of deriving the closed form posterior or analysis step which will later be adapted for the diffusion score matching Bayes posterior. Afterwards we will provide a short introduction to generalised Bayesian inference and more specifically the diffusion score matching Bayes posterior. We will use these results to derive the closed form of the posterior or analysis step of the D_w-Kalman filter via weighted diffusion score matching. With the posterior at hand we can prove the observation outlier robustness of the D_w-Kalman filter under given assumptions for the resulting choice of weight function. Finally, we showcase the results via simulation experiments and briefly discuss the results in the scope of feasibility and the initial motivation regarding observation noise mis-specification as well as chances and limitations.

For the second part, we will introduce the sequential change point problem and provide a notion of frequent reliability criteria and recall arguments of the popular CUSUM procedure. The focus will then be on recent advances in Bayesian online change point detection. The main body of work will be in connecting both exploiting the structure and idea of scenarios in Bayesian online change point detection with the versitility and reliability of CUSUM schemes resulting in suitable adaptations for detecting signal jumps in the Kalman setting. A major focus herein is in maintaining desirable reliability criteria in detection while reformulating approaches to suit quantification of uncertainty about potential jumps in inference and forecasting. To emphasize, while the employed tools are not necessarily new, the idea of utilizing the conditional evidence in Bayesian filtering is. Again, we showcase the results via simulation experiments and discuss the results in the scope of feasibility and the context of signal noise mis-specification as well as chances and limitations.

The third major chapter explores the double mis-specification setting of mis-specification in heavy tailedness in both noise terms. We start with a discussion on compatibility of results in the previous chapters as well as corresponding fallacies, however, the main results will be in simulation experiments. As the third main chapter, chapter 4), introduces no new concepts and is much shorter. Again, it ends with a discussion of insights regarding feasibility, limitations and chances as well as directions for further investigations.

1.4. Relevance and Context Regarding Forecasting

> [We] have a prediction problem.
> We love to predict things
> - and we aren't very good at it
>
> *Nate Silver*

This part will provide a brief introduction to the broader context on mathematical and statistical modelling with a focus on practice in time series forecasting at the intersection to data assimilation.

Starting with an intuitive notion on the process of statistical modelling, we want to pick up on the ideas in (Kokko, 2005). Hereby, creating a statistical or predictive model is compared to drawing a map. The model, just as the map, needs to contain enough detail to be useful, suit its purpose and have practical use. Yet too much detail can be overwhelming, misleading or distract from the relevant information and initial aims. Everything not contained in our map must be negligible for its purpose, i.e. pattern-less error. This intuition generally holds quite well and ties in with higher education teaching on statistical modelling in introductory lectures on mathematical statistics. A basic scheme of what and how things interact in mathematical and statistical modelling is shown in figure (1.1) taken from one such lecture (Husiniga, 2021).

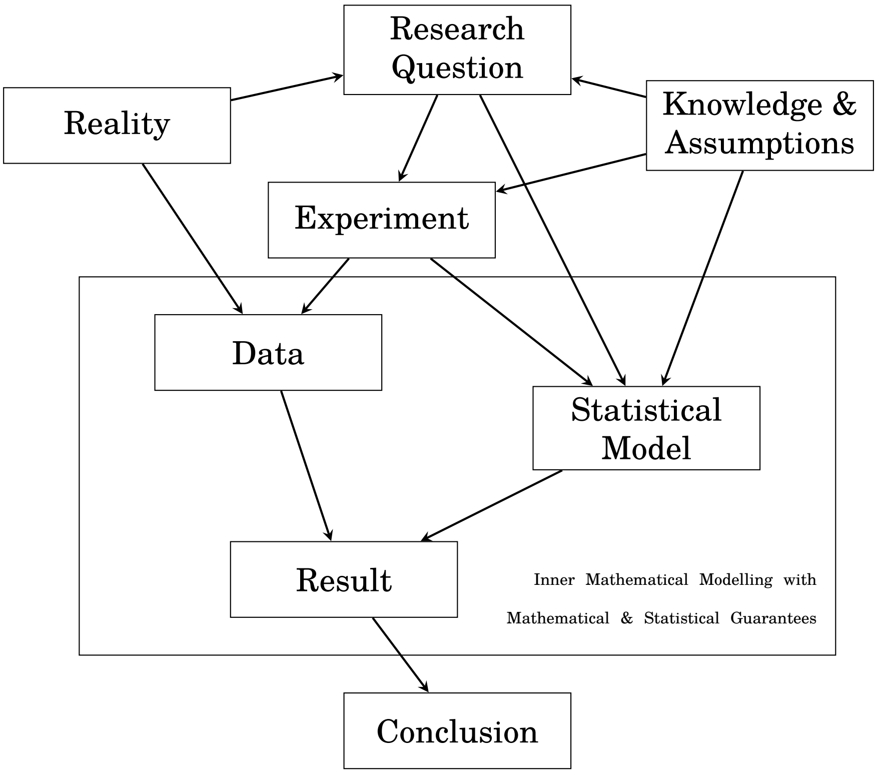


Figure 1.1.: Concept chart of mathematical and statistical modelling as taken from (Husiniga, 2021).

It portrays how Kokko's map ties in with a larger process and how the mentioned crucial degree of detail is necessarily bound to several influences and factors. On a more subtle note, it also adds that a model is part of an inner mathematical framework limited by feasibility in evaluation and computation. The problems addressed in the work at hand are best understood in that bigger picture. Mis-specification in modelling heavy tails or outliers in data as two sides of the same coin frequently matter whenever knowledge and reality end up not matching. Missing knowledge broadly speaking hereby encompasses not knowing about potential for outliers produced by heavy tails, not knowing better about noise specification or simply not being able to adequately evaluate or account for better noise specification with the available tools. For the most part, this last issue is addressed in this work in the context of the popular Kalman filter - while one might even be aware about potential for outliers in measurement and heavy tails of noise distributions, it might be unfeasible and costly to incorporate this knowledge. Accordingly, we want to pick up there and provide easy to implement solutions to reduce the impact of misspecification in heavy-tailedness as previously described and enable adjustments in modelling - we want to provide pencil strokes to add important details to the notion of the map, yet without cluttering it.

Modelling in statistical forecasting is even more so a suitable topic in that regard as it is very strongly bound to its assumptions and limitations. It is a diverse field across disciplines with different takes, approaches and methods in each of them. To provide a general introduction and understanding of the topic we want to recall some intuitions from Hyndman's *Forecasting: Priciples and Practice* (Hyndman and Athanasopoulos, 2018). In his words, the shared aim in forecasting problems is predicting a future scenario as accurate as possible given available information and knowledge. Hereby knowledge and available information combine and match to produce valuable insights - or in other words, we need the right forecasting method and tools for the given question and data. The quality of a forecast, the «predictability», then depends on several aspects that tie in with the upper part of the modeling scheme in figure (1.1). Following (Hyndman and Athanasopoulos, 2018), there are three main aspects:

- How well do we understand the factors that contribute,
- the data availability and
- whether the forecast can impact itself.

The aspects are mainly concerned with outer mathematical modelling as they do not consider our ability to translate our knowledge of factors and data into mathematical models that can be evaluated reliably. Yet, especially the last aspect is interesting regarding the work at hand. Outliers in signal noise may occur as reactions to past predictions not covered by the initial model: Say our forecasting system warns of a traffic jam and we publicly announce it. Consequently, drivers may change their route to a popular alternative producing a traffic jam there but not at the initially location forecast. This issue is addressed in the popular Lucas critique in econometrics (Lucas Jr, 1976) and may also be considered a sudden signal jump, an outlier behavior, not covered by the model. To conclude, Hyndman finishes his introduction with a heads up, saying «forecasters need to be aware of their own limitations and not claim more than is possible» bringing us back to the start.

The main assumption in forecasting is that the way in which an environment is changing will continue into the future, so a past pattern is assumed to repeat (Hyndman and Athanasopoulos, 2018). This idea is deeply intertwined in model based forecasting and even more so for systems with prior model knowledge. The latter are central subjects of this work as we assume there is some knowledge about the dynamics of a system, usually via a state space model of a signal process. While the concepts have long been frequent in engineering since Kalman's ground breaking publication (Kalman, 1960), state space models were only beginning to be widely used by statisticians for forecasting in the early 1980s (Gooijer and Hyndman, 2005). They provide a unifying framework especially suiting linear time series modelling and with Kalman's contribution, a closed form recursive algorithm for computing forecasts via combining prior knowledge about linear system dynamics and observations was easily available. The term *Bayesian Forecasting* arose for the general approach of combining a-priori system knowledge and observations into an a-posteriori forecast (Harrison and Stevens, 1976). Harrison and colleagues continued formulating an essential foundation for the practice of Bayesian forecasting.

- A parametric (or state space) model, not a functional model,
- information on probabilistic properties of the parameters for any given time,
- a sequential model on how parameters evolve through time systematically as well as stochastically and
- general uncertainty in the underlying model itself are required.

This basic set of properties includes the popular Kalman setting for a linear dynamical system with an independent Gaussian signal noise driving term, i.e. white noise or Brownian motion, and observation as linear combinations of the signal with additional independent Gaussian observation noise. Even for a system with little or no prior history, this framework enables quantified forecasts (Harrison and Stevens, 1976). As a note on the assumption of linearity, Harrison and colleagues comment that «linear combinations of linear models are itself a linear model. The obvious is singled out as a principle because of its far reaching implications for the construction of large-scale and apparently complex models. The essential point is that a large linear model may be considered the linear combination of a number of simpler models» (Harrison and Stevens, 1976) arguing for the still far reaching potential and wide applicability of this approach.

Jumping into the much more recent past with (Morzfeld and Reich, 2018), we have the context of Bayesian forecasting embedded in the wider field of data assimilation researching the mathematical and numerical foundation of combining models and data - this is also where this work places itself. Systems, such as frequent in weather forecasting or climate prediction as well as the cognitive sciences have long since become so large, that modelling comes with several challenges such as the popular *curse of dimensionality*. The aim is still about the same:«An elegant way of performing data assimilation is to compute the conditional probabilities, as described above, that describe the mathematical model in view of the data you collected» (Morzfeld and Reich, 2018). For a simple model, as in a linear model, with Gaussian probabilities, data assimilation is straight forward resulting in the previously mentioned Kalman filter. The contemporary, real challenges rarely fit that formula motivating modern research to address this reality. A popular approach also repeatedly picked up in the work at hand is the ensemble Kalman filter as a tool to manage non-linearity (Evensen et al., 2009). Data assimilation requires statements in precise mathematical manner, also about unknown things such as exact error distributions. Recalling Kokko's map and the general approach to mathematical and statistical modelling, this missing knowledge often leads to first resorting to assumptions which simplify the problem (Morzfeld and Reich, 2018). This work is contributing towards this specific challenge in researching how to address the difficulty of only vague knowledge of error or noise distributions via dampening the impact of wrongly simplified assumptions. To close, some claim data assimilation is among the main contributors to improving forecast over the past decade (Bauer et al., 2015). Further, data assimilation enables a unified formulation for a wide variety of different applications in conditional probabilities of the system or signal state given the available observations yet often requiring simplified assumptions for feasibility and lack of knowledge (Morzfeld and Reich, 2018). The relevance of this master thesis lies in its contribution of providing algorithms for data assimilation, akin to the popular Kalman filter and ensemble Kalman filter, with reduced impact of mis-specification in heavy tailedness of the noise distributions, so in simplified assumptions, as opposed to other approaches directly modelling heavy tailed noise terms via t-distributions (see (Bai et al., 2022) and (Tang et al., 2024) for additional details).

2. Addressing Observation Noise Mis-Specification: Robust Bayesian Inverse Inference in Filtering

2.1. Background

2.1.1. Revisiting the Kalman Filter

Recalling the classical Kalman filter, we want to adapt the notation and results in (Stannat, 2023) and (Reich and Cotter, 2015) with some additions from (Słupiński, 2023). We hereby switch fluently between covariance forms and information forms of multivariate Gaussian random variables, generally noting the distribution density function in covariance form $p(x) \sim n(x; m, P)$, the distribution density in information form $p(x) \sim n^{-1}(x; \theta, J)$, covariance P, mean m, precision $J = P^{-1}$ and potential $\theta = Jm \iff m = P\theta$. Additionally, we will frequently use the scaling notation $\propto$ for proportionality up to a constant factor independent of the variable of interest and $\overset{+C}{=}$ for equality up to a constant additional term independent of the variable of interest.

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let X_n be a multivariate random variable to model our noisy signal at discrete time steps $n = \{1, 2 \ldots, N\}$. X_n cannot be observed directly, however, we can measure it via another random variable $Y_n = g_n(X_n, V_n)$, the observation, with V_n denoting the observation noise term. Given the Kalman filter setting we assume X_n and Y_n to be jointly Gaussian with the following linear, time discrete, time-varying signal evolution equation and linear observation equation:

$$\begin{aligned} X_n &= A_n X_{n-1} + C_n W_n \\ Y_n &= H_n X_n + \Gamma_n V_n \end{aligned} \tag{2.1}$$

with

- $X_n : \Omega \to \mathcal{X} = \mathbb{R}^d$ - the d-dimensional signal random vector at time n,
- $Y_n : \Omega \to \mathcal{Y} = \mathbb{R}^p$ - the p-dimensional observation random vector at time n,
- $W_n : \Omega \to \mathbb{R}^d$ and $V_n : \Omega \to \mathbb{R}^p$ - independent standard Gaussian distributed random vectors at time n (white noise) of the corresponding dimensions,

- A_n, C_n, H_n and Γ_n of appropriate dimensions with non-singular $Q_n = C_n C_n^T$ and $R_n = \Gamma_n \Gamma_n^T$ and
- $p(x_0) \sim n(x_0; m_0, P_0)$, the initial Gaussian prior distribution.

The crucial property for the Kalman filter to work is that linear combinations of Gaussian random variables as well as the posterior of the involved Bayesian inverse inference problem remain Gaussian throughout time. Hence deriving a closed form recursive formula for the parameters of the signal forecast and signal posterior incorporating the observations is sufficient (see (Stannat, 2023) for additional details).

We produce the forecast density via forward propagating the current signal distribution given the available observations according to the signal evolution equation:

$$\begin{aligned} p(x_n|y_{1:(n-1)}) &\propto \exp[-\frac{1}{2}(x_n - A_n m_{n-1})^T (A_n P_{n-1} A_n^T + Q_n)^{-1}(x_n - A_n m_{n-1})] \\ &= \exp[-\frac{1}{2}(x_n - m_n^f)^T (P_n^f)^{-1}(x_n - m_n^f)] \\ &\propto \exp[-\frac{1}{2} x_n^T J_n^f x_n + x_n^T \theta_n^f], \end{aligned} \tag{2.2}$$

so $p(x_n|y_{1:(n-1)}) \sim n(x_n; m_n^f, P_n^f)$ or $p(x_n|y_{1:(n-1)}) \sim n^{-1}(x_n; \theta_n^f, J_n^f)$ with forecast covariance $P_n^f = A_n P_{n-1} A_n^T + Q_n$, forecast mean $m_n^f = A_n m_{n-1}$, forecast precision $J_n^f = (P_n^f)^{-1}$ and forecast potential $\theta_n^f = J_n^f m_n^f$ obtained via direct calculation taking (2.1).

The observation likelihood is then given as a conditional distribution on the current signal:

$$\begin{aligned} p(y_n|x_n) &\propto \exp[-\frac{1}{2}(y_n - H_n x_n)^T R_n^{-1}(y_n - H_n x_n)] \\ &= \exp[-\frac{1}{2} x_n^T H_n^T R_n^{-1} H_n x_n + y_n^T R_n^{-1} H_n x_n - \frac{1}{2} y_n^T R_n^{-1} y_n] \\ &\propto \exp[-\frac{1}{2} x_n^T H_n^T R_n^{-1} H_n x_n + x_n^T H_n^T R_n^{-1} y_n], \end{aligned} \tag{2.3}$$

so $p(y_n|x_n) \sim n(y_n; H_n x_n, R_n)$, see (Reich and Cotter, 2015) for additional details.
In Kalman filtering the aim is now to obtain the posterior distribution of the signal $p(x_n|y_{1:n})$ via Bayes theorem utilizing the forecast as a prior of the signal at time n thus solving the involved Bayesian inverse inference problem

$$p(x_n|y_{1:n}) \propto p(x_n|y_{1:(n-1)}) \cdot p(y_n|x_n) \tag{2.4}$$

with $p(x_n|y_{1:n}) \sim n(x_n; m_n, P_n)$.

As stated we desire to express closed form updates of the parameters of the posterior:

$$
\begin{aligned}
p(x_n|y_{1:n}) &\propto p(x_n|y_{1:(n-1)}) \cdot p(y_n|x_n) \\
&\propto \exp[-\frac{1}{2}x_n^T J_n^f x_n + x_n^T \theta_n^f] \cdot \exp[-\frac{1}{2}x_n^T H_n^T R_n^{-1} H_n x_n + x_n^T H_n^T R_n^{-1} y_n] \\
&= \exp[-\frac{1}{2}x_n^T (J_n^f + H_n^T R_n^{-1} H_n) x_n + x_n^T(\theta_n^f + H_n^T R_n^{-1} y_n)] \\
&= \exp[-\frac{1}{2}x_n^T J_n x_n + x_n^T \theta_n]
\end{aligned}
\tag{2.5}
$$

obtaining the Gaussian posterior density via (2.2) and (2.3) with parameters in information form $p(x_n|y_{1:n}) \sim n^{-1}(x_n; \theta_n, J_n)$ and recursive updates of the precision and potential at time n via

$$
\begin{aligned}
J_n &= J_n^f + H_n^T R_n^{-1} H_n \\
\theta_n &= \theta_n^f + H_n^T R_n^{-1} y_n.
\end{aligned}
\tag{2.6}
$$

The remaining challenge is to re-parameterize the obtained Gaussian posterior from information form to covariance form to obtain the classical recursive Kalman update formula. Via employing the *Sherman-Morrison-Woodbury* matrix inversion formula (see (Reich and Cotter, 2015) and (Golub and Van Loan, 2013) for details) we get for the covariance matrix

$$
\begin{aligned}
P_n = J_n^{-1} &= [J_n^f + H_n^T R_n^{-1} H_n]^{-1} \\
&= [(P_n^f)^{-1} + H_n^T R_n^{-1} H_n]^{-1} \\
&= P_n^f - P_n^f H_n^T [R_n + H_n P_n^f H_n^T]^{-1} H_n P_n^f \\
&= P_n^f - K_n H_n P_n^f
\end{aligned}
$$

with Kalman gain matrix

$$
K_n = P_n^f H_n^T [R_n + H_n P_n^f H_n^T]^{-1}.
$$

Using this result as well as repeated applications of the *Sherman-Morrison-Woodbury* matrix inversion formula we then get for the mean vector

$$
\begin{aligned}
m_n = P_n \theta_n &= P_n[\theta_n^f + H_n^T R_n^{-1} y_n] \\
&= P_n[(P_n^f)^{-1} m_n^f + H_n^T R_n^{-1} y_n] \\
&= [P_n^f - K_n H_n P_n^f][(P_n^f)^{-1} m_n^f + H_n^T R_n^{-1} y_n] \\
&= m_n^f - K_n H_n m_n^f + [P_n^f - K_n H_n P_n^f] H_n^T R_n^{-1} y_n \\
&= m_n^f - K_n H_n m_n^f + P_n^f H_n^T [R_n + H_n P_n^f H_n^T]^{-1} y_n \\
&= m_n^f - K_n H_n m_n^f + K_n y_n = m_n^f - K_n(H_n m_n^f - y_n).
\end{aligned}
$$

The overall result is the classical recursive Kalman filter formula updating the parameters in covariance form with forecast step building the prior distribution via propagating the signal

$$
\begin{aligned}
m_n^f &= A_n m_{n-1} \\
P_n^f &= A_n P_{n-1} A_n^T + Q_n
\end{aligned}
\tag{2.7}
$$

and analysis step incorporating the new observation and thus obtaining the posterior distribution of the involved Bayesian inverse inference problem

$$\begin{aligned} K_n &= P_n^f H_n^T [R_n + H_n P_n^f H_n^T]^{-1} \\ m_n &= m_n^f - K_n (H_n m_n^f - y_n) \\ P_n &= P_n^f - K_n H_n P_n^f. \end{aligned} \tag{2.8}$$

There is a large variety of different ways to derive the Kalman filter equations and the one presented is not necessarily the most intuitive or most elegant one. However, the chosen approach is suited best to showcase the changes for the generalized Bayesian adaptation. More specific, it is equation (2.5) where we will apply the generalization and change of divergence from Kullback-Leibler to diffusion score matching.

As a short remark on notation, in literature $\hat{y}_n = H_n m_n^f$ is frequently referred to as the observation forecast mean and $\gamma_n = y_n - \hat{y}_n = y_n - H_n m_n^f$ with $p(\gamma_n | y_{1:(n-1)}) \sim n(\gamma_n; 0, \Sigma_n)$ as innovation, a zero mean distribution with known variance $\Sigma_n = H_n P_n^f H_n^T + R_n$. Furthermore, the conditional marginal distributions of the observation $p(y_n | y_{1:(n-1)}) \sim n(y_n; H_n m_n^f, \Sigma_n)$ is denoted evidence and argued to be a key quantity to evaluated inference in practice with the latent state not available.

2.1.2. Introducing Generalized Bayesian Inference and Diffusion Score Matching

Motivating and Generalizing Bayesian Inference

For this section, we heavily rely on information and results in (Matsubara et al., 2023), (Altamirano et al., 2023b), (Altamirano et al., 2023a) and (Pacchiardi, 2021). We will start briefly highlighting the strong points of regular Bayesian inference before considering when they do not apply and the suggested work-around. We then draw from these results utilizing diffusion score matching in generalised Bayesian inference and prepare them for adaptation to the given setting. Hereby we will mainly use a simplified time-invariant notation to better focus on the changes to the Bayesian inverse inference problem in a specific analysis step via the introduced generalisation to Bayesian inference. Recall Bayes' theorem

$$\begin{aligned} p(x|y) &= \frac{p(x) \cdot p(y|x)}{p(y)} \\ &\propto p(x) \cdot p(y|x) \end{aligned} \tag{2.9}$$

with prior $p(x)$, likelihood $p(y|x)$ and posterior $p(x|y)$ as used in the previous section. For now we want to leave out the evidence $p(y)$ via proportionality as we focus on inference on the random vector X.

The general popularity of Bayesian approaches is tied to its desirable properties as nicely summarized in (Pacchiardi, 2021). For one, the likelihood principle applies in that new observations are incorporated only via the likelihood that contains all relevant information of an observation about the model parameters, which also applies to frequentist methods. Bayesian approaches additionally are optimal in the way they process and integrate information in that they make use of all available information contained by an observation (see (Zellner, 1988) for details). Finally, the *Bernstein-von Mises theorem* applies to Bayes' posteriors as an adapted version of the Central Limit Theorem in that the posterior distribution converges to a Gaussian distribution which is centered in the parameters of the true data generating process with asymptotically vanishing variance. However, these useful properties strictly require the strong assumption, that the statistical model is well specified - that is, the true data generating process is part of or covered by the statistical model. In mathematical terms, let $\pi(\cdot)$ be the true data generating process. Then the statistical model $(\mathcal{Y}, \{p(\cdot|x) : x \in \mathcal{X}\})$ is well specified if and only if

$$\exists x_0 \in \mathcal{X} : p(\cdot|x_0) = \pi(\cdot).$$

The existence of such a true parameter x_0 given the statistical model is the main issue. It is a broad and important discussion that this most fundamental assumption is frequently neglected in Bayesian modelling, recall (Gelman and Shalizi, 2013). Yet, as soon as it is not given, Zellner's optimal information processing no longer holds and the *Bernstein-von Mises theorem* instead tries to recover the model parameter closest to the true data generating process in KL-divergence (Pacchiardi, 2021). This is the key insight regarding mis-specification of the likelihood. As soon as the fundamental assumption is hurt, Bayesian inference aims to recover the best parameter such that $x_0 = \arg\min_{x \in \mathcal{X}} \mathcal{D}_{\mathrm{KL}}(\pi(\cdot), p(\cdot, x))$. While this might still be useful in some cases, it is likely not what is desired in others. Additionally and more important, posterior uncertainties are then also no longer well calibrated. The idea of generalized Bayesian inference starts right here with the question that given mis-specification of the likelihood, which type of divergence do we want to minimize to obtain desired properties given a modelling context.

In the introduced problem of this work, we aim for robustness regarding observation likelihood choice in tail behaviour, outliers and contamination, all of which essentially translating to mis-specification of the observation noise distribution regarding heavy tailedness. As stated in (Jewson et al., 2018) and further inspected in (Altamirano et al., 2023a), the KL-divergence is especially bad in that case by giving large importance and emphasizing tail behavior of distributions. As Pacchiardi explicitly puts it in (Pacchiardi, 2021):

> «With a finite amount of samples, that translates into saying that Bayes' posterior is highly sensitive to outliers in the data.»

This is the exact kind of mis-specification we are interest in and want to focus on for this first part regarding observation noise. Furthermore, we want to focus on how its impact is best avoided. Supposed observation outliers actually produced by heavy tails of the true generating process can heavily distort the sequential inference of the Kalman filter, or for that, Bayesian filters in general. Recalling the second part, for reliably detecting rapid change and jumps in the signal, we have to strongly control the impact of these observation outliers on the signal estimation introduced via the Bayesian inverse inference. At the core, we want to utilize the same approach as in the robust change point detection algorithm in (Altamirano et al., 2023b) with their scalability being an essential requirement, translating into a form of conjugacy of parameters in practice - so similar to the regular Kalman filter. Accordingly, the results are of great value for the general probabilistic forecasting problem involving Bayesian filtering.

Generalized Bayesian inference, as researched in (Grünwald, 2012), (Bissiri et al., 2016), (Jewson et al., 2018), (Pacchiardi and Dutta, 2021) and (Matsubara et al., 2022) among others, is mainly a generalization in producing the posterior via introducing choice in loss function, however, regularly resulting in choosing a discrepancy measure between an assumed likelihood and the true data generating process. Further, this approach thereby dips into the realm of machine learning via introducing additional tuning parameters, such as a learning rate, and often requiring additional computational tools. Adapting the notation in (Altamirano et al., 2023b), the generalised Bayes posterior is given by

$$p(x|y) \propto p(x) \cdot \exp[-\beta \cdot \hat{\mathcal{D}}(\pi(y), p(y|x))]. \tag{2.10}$$

with learning rate $\beta > 0$. Following (Altamirano et al., 2023b), we aim for $\mathcal{D}$ to be a discrepancy measure on probability measures on $\mathcal{Y}$ given a parameter $x \in \mathcal{X}$ and the true data generating process π and $\hat{D}$ its empirical estimator via an observation y. Taking $\beta = 1$ and $\hat{\mathcal{D}}_{\text{KL}}(\pi(y), p(y|x)) = -\log(p(y|x))$ as estimator of the KL-divergence via cross-entropy between model likelihood $p(y|x)$ and the true data generating process as reference distribution π generating the observation y, we recover (2.9), the regular Bayes posterior as stated in (Altamirano et al., 2023b). There is a long list of research applications successfully employing generalised Bayesian posteriors for robustness or computational feasibility in a wide variety of contexts (see (Altamirano et al., 2023b) and (Altamirano et al., 2023a) for examples). The robustness in observation outliers considered in this work will be introduced further down. Next to generalised Bayes posteriors, generalised Bayesian inference also covers other approaches such as non-parametric concepts.

Estimating Divergence via Diffusion Score Matching

The choice of divergence measure here, can be loosely understood as switching from a Shannon information based measure with KL-divergence, so relative entropy, to a Fisher divergence based measure. Again, we hereby follow (Altamirano et al., 2023b) in their argumentation. Starting with the idea of score matching as initially introduced in (Hyvärinen and Dayan, 2005), we want to choose parameters such as to minimise the Fisher divergence between a statistical model and a reference, the true data generating process. Define the score functions as $s_p(y) = \nabla \log(p(y))$ for densities p on $\mathcal{Y}$. The Fisher divergence for the introduced inverse inference is then given via

$$\mathcal{D}_{Id}((\pi(\cdot), p(\cdot|x)) = \mathbb{E}_{Y\sim\pi}[||s_{p(\cdot|x)}(Y) - s_{\pi}(Y)||_2^2]. \tag{2.11}$$

Minimising means therefore matching the statistical model, here the likelihood $p(\cdot|x)$, to the true data generating process $\pi(\cdot)$ with respect to the expected squared L2-distance of their score functions. (Altamirano et al., 2023b) highlights two main reasons supporting this approach. First, the score function can be used for unnormalized likelihoods as $s_p(y) = \nabla \log(p(y)) = \nabla \log(\frac{1}{Z}\tilde{p}(y)) = \nabla \log(\tilde{p}(y))$ since for normalizing constant $Z > 0$ it follows $\nabla \log(\frac{1}{Z}) = 0$. Much more important however, the Fisher divergence can be rewritten not needing to estimate $s_\pi(\cdot)$ for computation under mild constraints in usual regularity conditions on boundary and smoothness. This enables comparably simple implementation, hence score matching has since been further developed and applied to a variety of contexts such as Bayesian model selection or change point detection. Interestingly however, there seems to be no immediate research on score matching for Bayesian inverse inference such as in Bayesian filtering problem.

For the presented problem we want to utilize diffusion score matching as introduced in (Barp et al., 2019) as a generalisation via introducing a weight matrix $w(Y)$ to the score function difference:

$$\mathcal{D}_w(\pi(y), p(y|x)) = \mathbb{E}_{Y\sim\pi}[||w(Y)^T(s_{p(\cdot|x)}(Y) - s_\pi(Y))||_2^2].$$

Hereby $w : \mathcal{Y} \to \mathbb{R}^{p\times p}$ is a point-wise invertible matrix valued function. Following (Anastasiou et al., 2023), w is the name-giving diffusion matrix for interpreting the diffusion score matching divergence in the framework of diffusion Stein discrepancy with diffusion Stein operator w (see (Anastasiou et al., 2023 for details). $\mathcal{D}_w$ is a divergence measure on $\mathcal{Y} = \mathbb{R}^p$ given the expected square difference is finite, so $\int_{\mathcal{Y}} \pi(y)(s_{p(\cdot|x)}(y) - s\pi(y))^2 \mathrm{d}y < \infty$. Additionally assuming appropriate smoothness and boundary conditions on the measures, this can then be relaxed to simply require that $\mathcal{Y}$ is a connected subset in $\mathbb{R}^p$ (see (Liu et al., 2022) and (Zhang et al., 2022)). The idea of introducing the weight function w is hereby rather direct in highlighting areas of $\mathcal{Y}$ in which we want to put more emphasis on matching scores. Moreover, it will be our main tool obtaining the desired robustness via controlling the influence of heavy tailed behavior of the data generating process in comparison to the model likelihood by staunching its impact.

As said before, the main result enabling the work in (Altamirano et al., 2023b) and (Altamirano et al., 2023a) are the insights in (Liu et al., 2022) and (Matsubara et al., 2022) which allow to work around directly estimating $\mathcal{D}_w$. Given the mentioned smoothness and boundary conditions, the expression can be rewritten via integration by parts to not explicitly include $\pi(\cdot)$, the usually unknown true data generating process, via

$$\mathbb{E}_{Y\sim\pi}[||w(Y)^T s_{p(\cdot|x)}(Y)||_2^2 + 2\nabla \cdot (w(Y)w(Y)^T \nabla s_{p(\cdot|x)}(Y))]$$

up to a constant independent of our parameter of interest x - so similar to KL-divergence. The true data generating process $\pi(\cdot)$ is still included implicitly, however, allowing the natural Monte Carlo estimator

$$\hat{\mathcal{D}}_w(y;x) = \hat{\mathcal{D}}_w(\pi(y), p(y|x)) = ||w(y)^T s_{p(\cdot|x)}(y)||_2^2 + 2\nabla \cdot (w(y)w(y)^T \nabla s_{p(\cdot|x)}(y)). \quad (2.12)$$

The mentioned smoothness and boundary conditions are usually given for Gaussian distributions according to (Altamirano et al., 2023b) however, as they also put conditions on the true data generating process π we want to briefly cover them. The likelihood needs to be twice differentiable, this is easily achieved in the Kalman setting. More restrictive, $[\pi w w^T s_{p(\cdot|x)}]$, $[\nabla \cdot (\pi w w^T s_{p(\cdot|x)})] \in L^1(\mathbb{R}^p)$. In other words, we need to assume for the true data generating process to still be measurable after the given transformation or rather for both terms containing π to be measurable functions.

At that point a short comment on the notation as taken from (Altamirano et al., 2023b). While $\nabla f(x)$ is the usual Jacobian matrix with the partial derivatives on a vector field f, $\nabla \cdot f(x)$ is the divergence operator which can be understood as the dot-product of the vector of partial derivative operators and f (or its columns).

The diffusion score matching estimator obtained in (2.12) enables two new insights. For one, it specifies how we obtain our new generalized Bayes posterior (2.10) by specifying the divergence in

$$p_\beta^{\mathcal{D}_w}(x|y) \propto p(x) \cdot \exp[-\beta \cdot \hat{\mathcal{D}}_w(y;x)] \tag{2.13}$$

resulting in the diffusion score matching Bayes posterior as in (Altamirano et al., 2023b). Furthermore, we can also analytically evaluate $\hat{\mathcal{D}}_w(y;x)$ for model observation likelihoods in the exponential family, such as used in the setting of this work. In other words, we can state a closed form of the $\mathcal{D}_w$-posterior, $p_\beta^{\mathcal{D}_w}(x|y)$, for an adaptation of the analysis step in (2.5).

Constructing a Gaussian Posterior

While $\pi(\cdot)$ is still unknown, we can explicitly state the assumed likelihood $p(y_n|x_n)$ as in (2.3) and go from there. Again, the mentioned boundary and smoothness conditions generally hold for our case as is stated in (Altamirano et al., 2023a). As an intermediate step, we want to briefly inspect the score function of the observation likelihood as in (2.3) with

$$\begin{aligned} s_{p(\cdot|x_n)}(y_n) &= \nabla_{y_n} \log(p(y_n|x_n)) \\ &\propto \nabla_{y_n} \log(\exp[-\frac{1}{2} x_n^T H_n^T R_n^{-1} H_n x_n + y_n^T R_n^{-1} H_n x_n - \frac{1}{2} y_n^T R_n^{-1} y_n]) \\ &= \nabla_{y_n} y_n^T R_n^{-1} H_n x_n - \frac{1}{2} \nabla_{y_n} y_n^T R_n^{-1} y_n \\ &= R_n^{-1} H_n x_n - R_n^{-1} y_n. \end{aligned} \tag{2.14}$$

Taking (2.14) and $\hat{\mathcal{D}}_w(y;x)$ as in (2.12), then

$$\hat{\mathcal{D}}_w(y_n;x_n) = \underbrace{||w(y_n)^T s_{p(\cdot|x_n)}(y_n)||_2^2}_{(a)} + 2\underbrace{\nabla \cdot (w(y_n) w(y_n)^T \nabla s_{p(\cdot|x_n)}(y_n))}_{(b)} \tag{2.15}$$

with

$$\begin{aligned}(a) &= ||w(y_n)^T s_{p(\cdot|x_n)}(y_n)||_2^2 \\ &= ||w(y_n)^T (R_n^{-1} H_n x_n - R_n^{-1} y_n)||_2^2 \\ &= x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n x_n + y_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n \\ &\quad - 2 x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n \\ &\overset{+C}{=} x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n x_n - 2 x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n\end{aligned} \tag{2.16}$$

using the symmetry of R_n and its inverse. Further,

$$\begin{aligned}(b) &= \nabla \cdot (w(y_n) w(y_n)^T \nabla s_{p(\cdot|x_n)}(y_n)) \\ &= \nabla \cdot (w(y_n) w(y_n)^T (R_n^{-1} H_n x_n - R_n^{-1} y_n)) \\ &= \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n x_n) - \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} y_n) \\ &\overset{+C}{=} x_n^T \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n)\end{aligned} \tag{2.17}$$

with $\overset{+C}{=}$ referring to equality up to an additive constant independent of x_n, our parameter of interest. We now recombine (2.16) and (2.17) in (2.15) to explicitly evaluate (2.12) for the given setting:

$$\begin{aligned}\hat{\mathcal{D}}_w(y_n; x_n) &= x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n x_n - 2 x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n \\ &\quad + 2 x_n^T \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n) \\ &= x_n^T H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n x_n \\ &\quad + 2 x_n^T (-H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n + \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n)) \\ &= x_n^T \Lambda_n(y_n) x_n + 2 x_n^T \nu_n(y_n)\end{aligned} \tag{2.18}$$

with

$$\begin{aligned}\Lambda_n(y_n) &= H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n \\ \nu_n(y_n) &= -H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n + \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n).\end{aligned} \tag{2.19}$$

The machinery at work here was largely developed in (Altamirano et al., 2023b) and can be adapted for exponential family likelihoods. There are changes mainly concerning the forward map H_n and deriving the multivariate Gaussian case in order to adapt it for the context of this work. Also, the diffusion score matching estimator was reduced to take a single observation at a time suiting the sequential setting at hand, however, it can easily be adapted to take multiple observations in each step via expanding the Monte Carlo estimator for the expectation, so

$$\hat{\mathcal{D}}_w(y_{1:t}; x) = \frac{1}{t} \sum_{i=1}^{t} [||w(y_i)^T s_{p(\cdot|x)}(y_i)||_2^2 + 2 \nabla \cdot (w(y_i) w(y_i)^T \nabla s_{p(\cdot|x)}(y_i))] \tag{2.20}$$

for $t \geq 1$ many observations at every time step. Moreover, this is also the aggregated estimate whenever there is no signal process and the forward propagation is obsolete, i.e. for the regular Bayesian inverse inference over multiple observations.

With a closed form of the diffusion score matching estimator available with (2.18), obtaining the adapted posterior is rather straight forward since (Altamirano et al., 2023b) states that for a squared exponential prior, the posterior will be (truncated) Gaussian. Taking the prior as in (2.2) and the $\mathcal{D}_w$-posterior as in (2.13), then

$$\begin{aligned} p_\beta^{\mathcal{D}_w}(x_n|y_{1:n}) &\propto p(x_n|y_{1:(n-1)}) \cdot \exp[-\beta \cdot \hat{\mathcal{D}}_w(x_n, y_n)] \\ &\propto \exp[-\frac{1}{2}x_n^T J_n^f x_n + x_n^T \theta_n^f] \cdot \exp[-\beta(x_n^T \Lambda_n(y_n) x_n + 2x_n^T \nu_n(y_n))] \\ &= \exp[-\frac{1}{2}x_n^T (J_n^f + 2\beta\Lambda_n(y_n))x_n + x_n^T(\theta_n^f - 2\beta\nu_n(y_n))] \\ &= \exp[-\frac{1}{2}x_n^T J_n^a x_n + x_n^T \theta_n^a], \end{aligned} \tag{2.21}$$

with

$$\begin{aligned} J_n^a &= J_n^f + 2\beta\Lambda_n(y_n) \\ \theta_n^a &= \theta_n^f - 2\beta\nu_n(y_n) \end{aligned} \tag{2.22}$$

resulting in the recursive Gaussian posterior $p_\beta^{\mathcal{D}_w}(x_n|y_{1:n}) \sim n^{-1}(x; \theta_n^a, J_n^a)$ in information form, parallel to (2.5). The remaining challenge, again, is to transform the obtained precision and potential into covariance and mean. As stated in (Altamirano et al., 2023b), we essentially obtain a form of Gaussian conjugacy. Moreover, as we can give a recursive formula of this update step, this renders the resulting $\mathcal{D}_w$-posterior Kalman filter fairly competitive in scalability compared to similar methods - such as the regular Kaman filter. An extension for an ensemble adaptation to avoid the remaining inversion is also right around the corner.

2.2. Deriving a Robust Kalman Filter

The involved Bayesian inverse inference problem of the Kalman filter is the only step we address with major changes for this part. This makes intuitive sense as it is also the only part where observation noise mis-specification matters. While we obtained a recursive update for the diffusion score matching Bayes posterior, we have yet to make it robust via the choice of the point-wise invertible matrix valued function w, the diffusion matrix. We hereby follow the argumentation in (Altamirano et al., 2023b) with additional tools from (Altamirano et al., 2023a).

2.2.1. Global Bias Robustness Regarding Observation Outliers

Following (Altamirano et al., 2023b), when speaking about mis-specification and robustness in the context of observations, we want to refer to the terms as in the framework of ε-contamination models (see (Huber, 2004) for details). Contamination hereby resembles the observation outliers or observations produced by heavy tails of the true observation noise in robustness referring to finite impact on estimation by observations produced via contamination or the other mechanism. In mathematical terms, for a given distribution Q on observation space $\mathcal{Y}$, the ε-contaminated version is described by $\mathbf{Q}_{\varepsilon,y_0} = (1-\varepsilon)\mathbf{Q} + \varepsilon\delta_{y_0}$ with δ_{y_0} the Dirac-measure at point $y_0 \in \mathcal{Y}$ and $\varepsilon \in [0,1]$. In frequentist analysis, we are interested in understanding the impact of ε on an appropriate point estimator T via

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon} ||T(\mathbf{Q}) - T(\mathbf{Q}_{\varepsilon,y_0})||_2, \tag{2.23}$$

the influence function, which reduces under mild assumptions to

$$\frac{\partial}{\partial\epsilon} ||T(\mathbf{Q}_{\varepsilon,y_0})||_2 \bigg|_{\varepsilon=0}. \tag{2.24}$$

It is a fairly classical and established tool to capture outlier robustness of an estimator in both parametric and non-parametric contexts but requires adapting for insights on Bayesian posteriors. The general idea, however, is still the same. To specify, in the Bayesian case the estimator extends to $T : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X})$ with $\mathcal{P}(\cdot)$ describing the respective space of distributions, observation space $\mathcal{Y}$, and parameter space $\mathcal{X}$ (here the signal space). In words, for estimating Bayesian posteriors via observations we are not interested in estimates on $\mathcal{X}$, but require our estimator to map into $\mathcal{P}(\mathcal{X})$, so to estimate non-finite-dimensional objects, distributions, over $\mathcal{X}$. Taking the ansatz in (Altamirano et al., 2023b), we approach this problem in two steps. We define the influence function point-wise over $\mathcal{X}$, introducing an additional sensitivity parameter, to then take the double supremum and check for a bound. Our estimator of choice is the $\mathcal{D}_w$-posterior depending on the likelihood and true data generating process, so $p_\beta^{\mathcal{D}_w}(x|y;\pi) = p_\beta^{\mathcal{D}_w}(x|y)$ and $p_\beta^{\mathcal{D}_w}(x|y;\pi_{\varepsilon,y_0})$, its contaminated version. The point-wise posterior influence function is then given by

$$\mathrm{PIF}(y_0, x, \pi) = \frac{\mathrm{d}}{\mathrm{d}\epsilon} p_\beta^{\mathcal{D}_w}(x|y;\pi_{\varepsilon,y_0}) \bigg|_{\varepsilon=0} \tag{2.25}$$

with sensitivity in parameters x for the estimator and contamination point y_0. The posterior is globally bias-robust if

$$\sup_{x\in\mathcal{X}, y_0\in\mathcal{Y}} \mathrm{PIF}(y_0, x, \pi) < \infty, \tag{2.26}$$

so if for every combination of parameter x and contamination point y_0 the influence on the estimator is finite, hence, the impact of contamination is uniformly bounded both over all parameters and all locations of contamination. Studying the robustness of generalised posteriors via the PIF was introduced in (A. Ghosh and Basu, 2016) and extended in (Matsubara et al., 2022).

Following (Altamirano et al., 2023b), we choose $w(y)$ to guarantee existence of such a bound to obtain the desired robustness to outliers for our posterior. As key result of their work, they show that the double supremum over the PIF is bound, given there is a function $\gamma(x)$ such that

1. $\sup_{y_0\in\mathcal{Y}} |\mathcal{D}_w(y_0; x)| \leq \gamma(x)$,

2. $\sup_{x \in \mathcal{X}} p(x)\gamma(x) < \infty$ and
3. $\int_{x \in \mathcal{X}} p(x)\gamma(x)\mathrm{d}x < \infty$ with prior $p(x)$.

As it is nicely put in (Altamirano et al., 2023a), condition 1 is our tool to design w and ensures that outliers are sufficiently weighted down to obtain robustness. Condition 2 and 3 ensure that our posterior does not blow up in any single point and first moment of the bound given the Bayesian model, so adding the uniform bound in parameter x given a prior.

Designing the Diffusion Matrix w and Bound

Accordingly, we now want to construct w for it satisfying such a boundary function $\gamma(x)$ using the first condition:

$$\begin{aligned} &\sup_{y_0 \in \mathcal{Y}} |\mathcal{D}_w(y_0; x)| = \sup_{y_0 \in \mathcal{Y}} ||w(y_0)^T (s_{p(\cdot|x)}(y_0) - s_\pi(y_0))||_2^2 \leq \gamma(x) \\ \iff &\sup_{y_0 \in \mathcal{Y}} [\underbrace{||w(y_0)^T s_{p(\cdot|x)}(y_0)||_2^2}_{(a)} + 2 \underbrace{\nabla \cdot (w(y_0) w(y_0)^T \nabla s_{p(\cdot|x)}(y_0))}_{(b)}] \leq \gamma(x) \end{aligned} \tag{2.27}$$

via (2.12). We want both parts to be bound depending on x to then obtain $\gamma(x)$. Starting with

$$\begin{aligned} (a) &= ||w(y_0)^T s_{p(\cdot|x)}(y_0)||_2^2 \\ &= ||w(y_0)^T [R^{-1} H x - R^{-1} y_0]||_2^2 \\ &= \sum_{i=1}^{p} (w(y_0)^T R^{-1} [Hx - y_0])_i^2 \end{aligned} \tag{2.28}$$

via (2.16), we again use a non-time varying simplified version of our likelihood in this step. The term in (a) is bound, if every additive term is bound, so if $|(w(y_0)^T R^{-1}[Hx - y_0])_i| < \infty$ for every $i \in \{1, 2, \dots, p\}$. Choose $w(y_0) = R^{\frac{1}{2}} \tilde{w}(y_0) \iff w(y_0)^T = \tilde{w}(y_0) R^{\frac{1}{2}}$ with $\tilde{w} : \mathcal{Y} \to \mathbb{R}^{p \times p}$ a new point-wise invertible *diagonal* matrix valued function with positive entries, the weights. Let $M_{\max} = \underset{i,j \in \{1,2,\dots,p\}}{max} |M_{ij}|$ for arbitrary matrices M, then

$$\begin{aligned} |w(y_0)^T R^{-1} [Hx - y_0]|_i &= |\tilde{w}(y_0) R^{\frac{1}{2}} R^{-1} [Hx - y_0]|_i = |\tilde{w}(y_0) R^{-\frac{1}{2}} [Hx - y_0]|_i \\ &\overset{+C}{=} |\tilde{w}(y_0) R^{-\frac{1}{2}} y_0|_i = |\sum_{j=1}^{p} \tilde{w}(y_0)_{ii} R_{ij}^{-\frac{1}{2}} (y_0)_j| \\ &\overset{\triangle}{\leq} \tilde{w}(y_0)_{ii} \sum_{j=1}^{p} |R_{ij}^{-\frac{1}{2}}| |(y_0)_j| \leq R_{\max}^{-\frac{1}{2}} \tilde{w}(y_0)_{ii} \sum_{j=1}^{p} |(y_0)_j| \\ &= R_{\max}^{-\frac{1}{2}} \tilde{w}(y_0)_{ii} ||y_0||_1 \leq \sqrt{c} R_{\max}^{-\frac{1}{2}} \tilde{w}(y_0)_{ii} ||y_0||_2 \\ &\propto \tilde{w}(y_0)_{ii} ||y_0||_2. \end{aligned} \tag{2.29}$$

We hereby utilize several things implicitly. R is symmetric and positive definite and has exactly one decomposition $R^{\frac{1}{2}}R^{\frac{1}{2}} = R$ with $R^{\frac{1}{2}}$ also symmetric and positive definite, the positive square root of R. Further, $R^{-1} = (R^{\frac{1}{2}}R^{\frac{1}{2}})^{-1} = R^{-\frac{1}{2}}R^{-\frac{1}{2}}$ with $R^{\frac{1}{2}}R^{-1} = R^{-\frac{1}{2}}$ as used above. In the last equation we use the inequality of L1- and L2-norms in $||y||_2 \leq ||y||_1 \leq \sqrt{c}||y||_2$ for $y \in \mathbb{R}^p$.
Accordingly, each diagonal entry of $\tilde{w}(y_0)$ must weight down y_0 by at least $\frac{1}{1+||y_0||_2}$ with the additional 1 ensuring non-zero denominator. This is akin to the conditions in (Altamirano et al., 2023a) for robustness in their univariate Gaussian process regression setting. Therefore, we want to apply similar arguments for the choice of weight function of the diagonal entries of $\tilde{w}(\cdot)$ in that we want plausible observations to have high weights with weights decreasing for less plausible observations. So what is plausibility given the Kalman filter setting? Assuming an observation is an outlier produced by heavy tails of the true data generating process but not covered in the observation likelihood of the model - a realization of the mis-specification - we want to lean more towards the forecast estimator via the prior instead of the empirical estimator from the observation to not distort the posterior or analysis mean (see (Stannat, 2023) for the intuition on the Kalman gain). Accordingly, we want to employ a notion of distance of an observation from the observation forecast mean via the forward map, i.e the euclidean distance in observation space $||y_0 - (Hm^f)||_2$ in a simplified notation. Additionally, we need to ensure the weight matrix is point-wise invertible and smooth for cheap differentiability. Finally, we want outliers weighted down without being made obsolete, robustness yet also maintaining appropriate relevance, so some form of heavy tailed behavior should be included in the weight kernel keeping approximately linear weights. For the one dimensional problem (Altamirano et al., 2023a) suggest *inverse multi-quadratic kernels* (IMQ) as they satisfy all desired properties. They are smooth for differentiability which is not the case for direct linear down-weighting yet they have heavier tails and slower decay compared to a (squared) exponential kernel. Sharing their intuition as well as desired properties, we also want to utilize IMQ-kernels in the diffusion matrix $w(\cdot)$ via

$$\begin{aligned}
&\hat{w}(y_0) = (1 + \frac{\langle y_0 - Hm^f, y_0 - Hm^f\rangle}{q^2})^{-\frac{1}{2}} = (1 + \frac{||y_0 - Hm^f||_2^2}{q^2})^{-\frac{1}{2}},\\
&\tilde{w}(y_0) = \hat{w}(y_0) \cdot \mathbf{1}_{p\times p} \quad \text{and} \\
&w(y_0) = \hat{w}(y_0)R^{\frac{1}{2}}
\end{aligned} \tag{2.30}$$

with outlier thresholds $q > 0$ steering the kurtosis of the kernel and $\mathbf{1}_{p\times p}$ the $p \times p$ identity matrix. The resulting weight matrix $\tilde{w}(y_0)$ is diagonal, the point-wise inverse exists as $||y - Hm^f||_2^2 \geq 0$ and is easily obtained, just like its square and their partial derivatives. Further, $0 < \hat{w}(y_0) \leq 1$ with $w(y_0)_{ij} \leq R_{ij}^{\frac{1}{2}}$.

Returning to (2.28), recalling $M_{\max} = \underset{i,j\in\{1,2,\ldots,p\}}{max} |M_{ij}|$, we can now find a bound on (a)

via

$$
\begin{aligned}
(a) &= \sum_{i=1}^{p} (w(y_0)^T R^{-1}[Hx - y_0])_i^2 \\
&= \sum_{i=1}^{p} (\tilde{w}(y_0) R^{\frac{1}{2}} R^{-1}[Hx - y_0])_i^2 = \sum_{i=1}^{p} (R^{-\frac{1}{2}} \hat{w}(y_0)[Hx - y_0])_i^2 \\
&= \sum_{i=1}^{p} (\hat{w}(y_0) \sum_{j=1}^{p} R_{ij}^{-\frac{1}{2}} [Hx - y_0]_j)^2 \overset{\triangle}{\leq} \sum_{i=1}^{p} (\hat{w}(y_0) \sum_{j=1}^{p} |R_{ij}^{-\frac{1}{2}}|(|Hx|_j + |y_0|_j))^2 \\
&\leq R_{\max}^{-1} \sum_{i=1}^{p} (\sum_{j=1}^{p} \hat{w}(y_0)(|Hx|_j + |y_0|_j))^2 \leq R_{\max}^{-1} \sum_{i=1}^{p} (\sum_{j=1}^{p} (|Hx|_j + \tilde{c}^{(a)}))^2 \\
&= R_{\max}^{-1} \sum_{i=1}^{p} (||Hx||_1 + p\tilde{c}^{(a)})^2 = \tilde{c}_1^{(a)}(||Hx||_1 + \tilde{c}_2^{(a)})^2 \leq c_1^{(a)}(||Hx||_2 + c_2^{(a)})^2 = \gamma^{(a)}(x)
\end{aligned}
\tag{2.31}
$$

with $\hat{w}(y_0) \cdot |y_0|_i \leq \tilde{c}^{(a)} < \infty$ for all $i \in \{1, 2, \dots, p\}$ by construction.
In order to obtain the bound $\gamma(x) \geq \gamma^{(a)}(x) + 2\gamma^{(b)}(x)$ to satisfy the first condition as in (2.27) we still need to inspect the second part, (b), regarding the choice of w. For that step as well as later derivations, we want to briefly inspect two terms:

$$
\begin{aligned}
R^{-1} w(y_0) w(y_0)^T R^{-1} &= R^{-1} \hat{w}(y_0) R^{\frac{1}{2}} R^{\frac{1}{2}} \hat{w}(y_0) R^{-1} \\
&= \hat{w}^2(y_0) R^{-1} R R^{-1} = \hat{w}^2(y_0) R^{-1} \quad \text{and} \\
w(y_0) w(y_0)^T R^{-1} &= \hat{w}(y_0) \hat{w}(y_0) R R^{-1} \\
&= \hat{w}^2(y_0),
\end{aligned}
\tag{2.32}
$$

then

$$
\begin{aligned}
(b) &= \nabla \cdot (w(y_0) w(y_0)^T \nabla s_{p(\cdot|x)}(y_0)) \\
&= \nabla \cdot (w(y_0) w(y_0)^T R^{-1}[Hx - y_0]) \\
&= \nabla \cdot (\hat{w}^2(y_0)[Hx - y_0]) \\
&= \sum_{i=1}^{p} \frac{\partial}{\partial (y_0)_i} (\hat{w}^2(y_0)[Hx - y_0])_i \\
&= \sum_{i=1}^{p} \frac{\partial}{\partial (y_0)_i} \hat{w}^2(y_0)(Hx)_i - \sum_{i=1}^{p} \frac{\partial}{\partial (y_0)_i} \hat{w}^2(y_0)(y_0)_i,
\end{aligned}
\tag{2.33}
$$

so in order to obtain a bound $\gamma^{(b)}(x)$ on (b), we have to show a bound on the partial derivatives. We have

- $\hat{w}^2(y_0) = (1 + \frac{||y_0 - Hm^f||_2^2}{q^2})^{-1} \in (0, 1]$,
- $|\frac{\partial}{\partial (y_0)_i} \hat{w}^2(y_0)| = |\frac{-2(y_0 - Hm^f)_i}{q^2}(1 + \frac{||y_0 - Hm^f||_2^2}{q^2})^{-2}| \leq \tilde{c}^{(b)} \hat{w}(y_0) \leq \tilde{c}_1^{(b)}$ and
- $|\frac{\partial}{\partial (y_0)_i}(\hat{w}^2(y_0) \cdot (y_0)_i)| \leq |\frac{\partial}{\partial (y_0)_i} \hat{w}^2(y_0)| \cdot |y_0|_i + |\hat{w}^2(y_0)| \leq \tilde{c}^{(b)}(\hat{w}(y_0) \cdot |y_0|_i) + 1 \leq \tilde{c}_2^{(b)}$

with appropriately chosen constants via the down-weighting properties of the IMQ-kernel regarding y_0. Then

$$\begin{aligned}(b) &= \sum_{i=1}^{p} \frac{\partial}{\partial(y_0)_i}\hat{w}^2(y_0)(Hx)_i - \sum_{i=1}^{p} \frac{\partial}{\partial(y_0)_i}\hat{w}^2(y_0)(y_0)_i \\ &\overset{\triangle}{\leq} \sum_{i=1}^{p} |\frac{\partial}{\partial(y_0)_i}\hat{w}^2(y_0)||Hx|_i + \sum_{i=1}^{p} |\frac{\partial}{\partial(y_0)_i}\hat{w}^2(y_0)(y_0)_i| \\ &\leq \tilde{c}_1^{(b)} \cdot \sum_{i=1}^{p} |Hx|_i + p \cdot \tilde{c}_2^{(b)} \\ &\leq \tilde{c}_1^{(b)} \cdot ||Hx||_1 + c_2^{(b)} \leq c_1^{(b)} \cdot ||Hx||_2 + c_2^{(b)} = \gamma^{(b)}(x).\end{aligned} \tag{2.34}$$

Bringing the gathered insights back to 2.27, we have

$$\begin{aligned}\sup_{y_0 \in \mathcal{Y}} |\mathcal{D}_w(y_0; x)| &= \sup_{y_0 \in \mathcal{Y}} [\underbrace{||w(y_0)^T s_{p(\cdot|x)}(y_0)||_2^2}_{(a)} + 2 \underbrace{\nabla \cdot (w(y_0)w(y_0)^T \nabla s_{p(\cdot|x)}(y_0))}_{(b)}] \\ &\leq \gamma^{(a)}(x) + 2\gamma^{(b)}(x) = c_1^{(a)}(||Hx||_2 + c_2^{(a)})^2 + 2(c_1^{(b)}||Hx||_2 + c_2^{(b)}) \\ &= c_1^{(a)}||Hx||_2^2 + 2c_1^{(a)}c_2^{(a)}||Hx||_2 + c_1^{(a)}(c_2^{(a)})^2 + 2c_1^{(b)}||Hx||_2 + 2c_2^{(b)} \\ &= c_1||Hx||_2^2 + c_2||Hx||_2 + c_3 \propto (||Hx||_2 + c)^2 = \gamma(x)\end{aligned} \tag{2.35}$$

with appropriately chosen constants for each c. With the Gaussian prior in the Kalman setting, thus we have that each of its moments exists as well as squared exponential decay in the tails, we can now satisfy conditions 2 and 3 in (2.2.1) for the resulting $\gamma(x)$.

For a slightly more sophisticated approach on condition 3,

$$\begin{aligned}\mathbb{E}[(||HX||_2 + c)^2] &= \mathbb{E}[||HX||_2^2] + 2c\mathbb{E}[||HX||_2] + c^2 \\ &\leq \mathbb{E}[||HX||_2^2] + 2c\sqrt{\mathbb{E}[||HX||_2^2]} + c^2\end{aligned} \tag{2.36}$$

with

$$\begin{aligned}\mathbb{E}[||HX||_2^2] &= \mathbb{E}[(HX)^T(HX)] = \mathbb{E}[\mathrm{Tr}((HX)^T(HX))] \\ &= \mathbb{E}[\mathrm{Tr}((HX)(HX)^T)] = \mathrm{Tr}(\mathbb{E}[(HX)(HX)^T]) \\ &= \mathrm{Tr}(\mathrm{Cov}(HX) + \mathbb{E}[HX]\mathbb{E}[HX]^T) \\ &= \mathrm{Tr}(HP^fH^T + Hm^f(m^f)^TH^T) < \infty\end{aligned} \tag{2.37}$$

via linearity of the expectation, using Jensen's inequality for the expectation of the concave square root and the trace trick (the expectation of a trace is the trace of the expectation) this way showing a direct bound for condition 3 in $\mathbb{E}_{X\sim p(x)}[\gamma(X)] < \infty$.

Resulting Considerations

Given the choice of diffusion matrix w as above, we obtain the desired global bias robustness of the posterior distribution via the uniform bound on the posterior influence function. We achieved this without major additional constraints, however, so far we introduced two new parameters. As said, the learning rate β drags the approach into the realm of machine learning, however, since learning rates in generalised Bayesian inference as in (2.10) are fairly unresolved according to (Altamirano et al., 2023a) referring to (Lyddon et al., 2019), (Wu and Martin, 2023) and (Frazier et al., 2023) among others, we will fix β as a default at 1 for now and instead focus on the outlier threshold q as it has a direct interpretation for the IMQ-kernel in kurtosis. For $q \to \infty$ and $\beta = 1$ we recover $R^{\frac{1}{2}}$, the positive root of the original observation noise covariance, as diffusion matrix independent of $y \in \mathcal{Y}$. For finite choice of $q^2 > 0$ we get the desirable property that y_0 with a distance $||y_0 - Hm^f||_2^2$ exceeding q^2 is increasingly weighted down. In (Altamirano et al., 2023a) it is suggested to choose the q as the corresponding empirical quantile absolute deviation centered on the prior mean. This is only of limited use in our case, so instead we choose to pick q^2 related to the popular Mahalanobis distance (see (Mahalanobis, 2018) and (De Maesschalck et al., 2000) for details). This shares similarity with the approach in (Chang, 2014) for robust Kalman filtering via covariance scaling as will be discussed later. Recall $Y - Hm^f \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = HP^fH^T + R$. Currently, the IMQ-kernel considers the centered, squared euclidean distance $(y - Hm^f)^T(y - Hm^f)$ of an observation $y_0 \in \mathcal{Y}$. Re-scaling this distance with the knowledge about observation covariance in Σ via $z = (y - Hm^f)^T\Sigma^{-1}(y - Hm^f)$ leads to a standardized distance akin to the Mahalanobis distance with known distribution $Z \sim \chi^2(p)$ for $\mathcal{Y} = \mathbb{R}^p$. On a side note, this needs assuming H^T has full column rank for ensuring Σ is positive definite. Using this result we can build a heuristic for the outlier threshold q^2 via the resulting Chi-square distribution, however, needing to adapt the euclidean distance estimation in the IMQ-kernel. Choosing

$$(y - Hm^f)^T\Sigma^{-1}(y - Hm^f) = \langle y - Hm^f, y - Hm^f\rangle_{\Sigma^{-1}} = ||y - Hm^f||_{\Sigma^{-1}}^2 \tag{2.38}$$

instead of $||y - Hm^f||_2^2$ does only change the constants in the derivation of the bounds, but nothing on the desired outcomes regarding robustness as it is only a reweighting independent of the parameters of interest - the down-weighting property still holds in

$$\sup_{y_0 \in \mathcal{Y}} y_0 \cdot (1 + \frac{||y_0 - Hm^f||_{\Sigma^{-1}}^2}{q^2})^{-\frac{1}{2}} < \infty \tag{2.39}$$

for all $q > 0$. Accordingly, we now want to choose $q^2 = \mathbb{E}[\chi^2(p)] = p$ or $q^2 = \chi^2_{1-\alpha}(p)$ for a desired confidence threshold with confidence value $\alpha \in (0, 1)$.

In essence, we make the covariance matrix of the observation noise, R, out to be the main source of error introduced via the mis-specification. Accordingly we want to dynamically adjust its effect using $\hat{w}$ and thus the *inverse multi-quadratic* kernel to control its impact and adapt to outliers.

At this moment, it may seem somewhat counter-intuitive to choose $R^{\frac{1}{2}}$ over R, however, the first leads to better properties down the line as the diffusion matrix only appears in squared form. Similarly, choosing the IMQ-kernel currently leads to a range of values in $[\frac{1}{\sqrt{2}}, 1]$ for observation distances within the confidence threshold, so $||y_0 - Hm^f||_{\Sigma^{-1}}^2 < q^2$. Appropriate re-scaling by factor 2 and squaring the diffusion matrix yields more intuitive values recovering $2w^2(y_0) \approx R$ for $||y_0 - Hm^f||_{\Sigma^{-1}}^2 \approx q^2$ motivating the choice of $q^2 = p$, the expected squared Mahalanobis distance.

On a side note, the choice of diffusion matrix w made here is by no means unique. Really, the key insight is that it needs to achieve the desired scaling in $w(y_0) \propto \frac{1}{||y_0||}$ given the knowledge that all norms are equivalent in $\mathcal{Y} = \mathbb{R}^p$ and the additional constraints on point-wise non-singularity and smoothness. The respective bounds, here (a) and (b), are then obtained fairly straight forward. Accordingly more daring choices of diffusion matrix should be explored and may improve performance.
Additionally, using similar arguments as before but for the regular Bayesian posterior, superficially inspecting

$$\sup_{y_0 \in \mathcal{Y}} |-\log(p(y_0|x))| = \sup_{y_0 \in \mathcal{Y}} [\frac{1}{2}(y_0 - Hx)^T R^{-1}(y_0 - Hx)] \tag{2.40}$$

yields that there can not be such a bound $\gamma(x)$ to provide global bias robustness via this approach. This agrees with similar results in (Altamirano et al., 2023a).
As a remark, alternatively borrowing concepts from Tsallis statistical mechanics (see (Umarov et al., 2008) for additional details) with Tsallis information via q-logarithms $\log_q(p) = \frac{p^{1-q}-1}{1-q}$ (equivalent to Box-Cox transformation in (Box and Cox, 1964)), a parallel line of work for robust posteriors via Tsallis information might be worth exploring as

$$\sup_{y_0 \in \mathcal{Y}} |-\log_q(p(y_0|x))| < \infty \tag{2.41}$$

for all $q \in (0, 1)$ and $y_0 \in \mathcal{Y}$ with $p(y_0|x) < \infty$. However, further exploration will not be subject of this work.

Before formulating the results in theorems and propositions as well as stating the $\mathcal{D}_w$-Kalman formula in the next section, we want to use the obtained results on w to evaluate the divergence term in ν in (2.19). The first term on the right hand side in

$$\nu_n(y_n) = -H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n + \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n)$$

is similar to updating the potential in the regular Kalman filter by $H_n^T R_n^{-1} y_n$, see (2.6), which now reduces to $-H_n^T[\hat{w}_n^2(y_n)R_n^{-1}]y_n$. It is the second part that is new and in the divergence operator also somewhat more difficult to evaluate. However, parts of it are already known from deriving part of the bound for robustness in (2.33). Recall $(\nabla \cdot g(y))_k = \sum_{i=1}^p \frac{\partial g_{ik}}{\partial y_i}(y)$ for $k \in \{1, 2, \ldots, d\}$ as $g(y) = w(y)w(y)^T R^{-1} H$ maps from $\mathcal{Y}$ to $\mathcal{Y} \times \mathcal{X}$. So $(\nabla \cdot g(y))_k = \nabla \cdot g_k(y)$ with $g_k(y)$ is the k-th column of $g(y)$. Again, we want to use a simplified time-invariant notation for that step. Inspecting $g(y)$, we find that

$$\begin{aligned} g_{ik}(y) &= (w(y)w(y)^T R^{-1} H)_{ik} \\ &= (\hat{w}^2(y) R^{\frac{1}{2}} R^{\frac{1}{2}} R^{-1} H)_{ik} \\ &= (\hat{w}^2(y) H)_{ik} = \hat{w}^2(y) H_{ik} \end{aligned} \tag{2.42}$$

via previous arguments. Further, adapting previous results on the derivative, we now have partial derivatives of quadratic forms via the introduction of the standardization resulting in

$$\begin{aligned}\frac{\partial \hat{w}^2(y)}{\partial y_i} &= -\frac{\sum_{j=1}^{p}\Sigma_{ij}^{-1}(y-Hm^f)_j + \sum_{j=1}^{p}(y-Hm^f)_j\Sigma_{ji}^{-1}}{q^2}(\hat{w}^2(y))^2 \\ &= -\frac{2\sum_{j=1}^{p}\Sigma_{ij}^{-1}(y-Hm^f)_j}{q^2}(\hat{w}^2(y))^2 = -\frac{2\langle\Sigma_{i\bullet}^{-1},(y-Hm^f)\rangle}{q^2}\hat{w}^4(y)\end{aligned} \tag{2.43}$$

with $\Sigma = HP^fH^T+R$ as above and $M_{i\bullet}$ the i-th row of a matrix M. When additionally utilizing that Σ^{-1} is symmetric, then

$$\begin{aligned}\frac{\partial g_{ik}}{\partial y_i}(y) &= \frac{\partial \hat{w}^2(y)}{\partial y_i}H_{ik} \\ &= -\frac{2\langle\Sigma_{i\bullet}^{-1},(y-Hm^f)\rangle}{q^2}\hat{w}^4(y)H_{ik}.\end{aligned} \tag{2.44}$$

Taking everything together, we obtain

$$\begin{aligned}(\nabla\cdot(w(y_n)w(y_n)^TR_n^{-1}H_n))_k &= \sum_{i=1}^{p}\frac{\partial\hat{w}^2(y)}{\partial y_i}H_{ik} \\ &= -\frac{2\hat{w}^4(y)}{q^2}\sum_{i=1}^{p}\langle\Sigma_{i\bullet}^{-1},(y-Hm^f)\rangle H_{ik} \\ &= -\frac{2\hat{w}^4(y)}{q^2}\sum_{i=1}^{p}H_{ik}(\sum_{j=1}^{p}\Sigma_{ij}^{-1}(y-Hm^f)_j) \\ &= -\frac{2\hat{w}^4(y)}{q^2}\sum_{i=1}^{p}H_{ki}^T(\Sigma^{-1}(y-Hm^f))_i \\ &= -\frac{2\hat{w}^4(y)}{q^2}(H^T\Sigma^{-1}(y-Hm^f))_k \quad \text{or} \\ &= \sum_{i=1}^{p}H_{ki}^T\frac{\partial\hat{w}^2(y)}{\partial y_i}\end{aligned} \tag{2.45}$$

leading to two strains of thought in

$$\begin{aligned}\nabla\cdot(w(y_n)w(y_n)^TR_n^{-1}H_n) &= -\frac{2\hat{w}^4(y)}{q^2}H^T\Sigma^{-1}(y-Hm^f) \\ &= -\frac{2\hat{w}^4(y)}{q^2}H^T(HP^fH^T+R)^{-1}(y-Hm^f) \quad \text{or} \\ &= H^T\nabla\hat{w}^2(y)\end{aligned} \tag{2.46}$$

with $\nabla\hat{w}^2(y)\in\mathcal{Y}$, and $\hat{w}^4(y) = (1+\frac{||y-Hm^f||^2_{\Sigma^{-1}}}{q^2})^{-2}$. The first derivation includes the innovation precision $\Sigma^{-1} = (HP^fH^T+R)^{-1}$ in the potential update introduced via the adaptation of the distance which may offer opportunities for the resulting recursive solution. However, the second derivation with the gradients of the weight kernel is much more general and provides a more direct interpretation.

2.2.2. Deriving the $\mathcal{D}_w$-Kalman Filter

The only thing left before we can finally state the recursive update formula is the parameter transformation from information form to covariance form in the analysis step. Picking up at (2.22) we want to work alongside (2.6) to (2.8). Let

$$\begin{aligned} \Sigma_n &= H_n P_n^f H_n^T + R_n, \\ \hat{w}_n(y_n) &= (1 + \frac{||y_n - H_n m_n^f||^2_{\Sigma_n^{-1}}}{q^2})^{-\frac{1}{2}}, \\ w_n(y_n) &= \hat{w}_n(y_n) R_n^{\frac{1}{2}} \quad \text{and} \\ N_n(y_n)^{-1} &= 2\beta R_n^{-1} w_n(y_n) w_n(y_n)^T R_n^{-1} = 2\beta \hat{w}_n^2(y_n) R_n^{-1} \end{aligned} \tag{2.47}$$

with $N_n(y_n)$ symmetric and pointwise-invertible as R_n is symmetric and invertible. To obtain the covariance of the diffusion score matching Bayes posterior, we basically take the same approach as before but substituting R_n^{-1} with $N_n^{-1}(y_n)$. Again, the key is the *Sherman-Morrison-Woodbury* matrix inversion formula (see (Golub and Van Loan, 2013) for details). The resulting covariance is

$$\begin{aligned} P_n^a &= (J_n^a)^{-1} = [J_n^f + 2\beta\Lambda_n(y_n)]^{-1} \\ &= [J_n^f + 2\beta(H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n)]^{-1} \\ &= [(P_n^f)^{-1} + H_n^T N_n(y_n)^{-1} H_n]^{-1} \\ &= P_n^f - P_n^f H_n^T [N_n(y_n) + H_n P_n^f H_n^T]^{-1} H_n P_n^f \\ &= P_n^f - \tilde{K}_n(y_n) H_n P_n^f \end{aligned} \tag{2.48}$$

with adjusted Kalman gain

$$\begin{aligned} \tilde{K}_n(y_n) &= P_n^f H_n^T [N_n(y_n) + H_n P_n^f H_n^T]^{-1} \\ &= P_n^f H_n^T [\frac{1}{2\beta \hat{w}_n^2(y_n)} R_n + H_n P_n^f H_n^T]^{-1}. \end{aligned} \tag{2.49}$$

We hereby note for the diffusion weight to appear via the previously mentioned squared form with factor 2. Additionally note

$$\begin{aligned} P_n^a &= P_n^f - \tilde{K}_n(y_n) H_n P_n^f \\ &= [\mathbf{1}_{d\times d} - \tilde{K}_n(y_n) H_n] P_n^f \\ \iff \quad P_n^a (P_n^f)^{-1} &= \mathbf{1}_{d\times d} - \tilde{K}_n(y_n) H_n. \end{aligned} \tag{2.50}$$

As before, we use the result on the analysis covariance and repeated applications of the *Sherman-Morrison-Woodbury* matrix inversion formula to transform from potential to mean with

$$\begin{aligned} m_n^a &= P_n^a \theta_n^a = P_n^a [\theta_n^f - 2\beta\nu_n(y_n)] \\ &= P_n^a [\theta_n^f - 2\beta(-H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n + \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n))] \\ &= P_n^a [\theta_n^f + H_n^T N_n(y_n)^{-1} y_n - 2\beta \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n)] \\ &= \underbrace{P_n^a [(P_n^f)^{-1} m_n^f]}_{(a)} + \underbrace{P_n^a [H_n^T N_n(y_n)^{-1} y_n]}_{(b)} - \underbrace{P_n^a [2\beta \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n)]}_{(c)}. \end{aligned} \tag{2.51}$$

Part (a) is the same as for the regular Kalman filter apart from the adapted Kalman gain, so

$$(a) = m_n^f - \tilde{K}_n(y_n)H_n m_n^f. \tag{2.52}$$

Similar for part (b) with

$$\begin{aligned}(b) &= P_n^a[H_n^T N_n(y_n)^{-1} y_n] \\ &= P_n^a[H_n^T N_n(y_n)^{-1} y_n] \\ &= [P_n^f - \tilde{K}_n(y_n)H_n P_n^f][H_n^T N_n(y_n)^{-1} y_n] \\ &= P_n^f H_n^T[N_n(y_n) + H_n P_n^f H_n^T]^{-1} y_n \\ &= \tilde{K}_n(y_n) y_n.\end{aligned} \tag{2.53}$$

This leaves part (c) containing the divergence operator. This part was already prepared in 2.54 and for each choice of formulation we recover different, yet equivalent, formulations.

$$\begin{aligned}(c) &= P_n^a[2\beta\nabla\cdot(w(y_n)w(y_n)^T R_n^{-1}H_n)] \\ &= 2\beta P_n^a[-\frac{2\hat{w}_n^4(y_n)}{q^2}H_n^T(H_n P_n^f H_n^T + R_n)^{-1}(y_n - H_n m_n^f)] \\ &= P_n^a[-\frac{4\beta\hat{w}_n^4(y_n)}{q^2}H_n^T(H_n P_n^f H_n^T + R_n)^{-1}(y_n - H_n m_n^f)] \\ &= [P_n^f - \tilde{K}_n(y_n)H_n P_n^f][-\frac{4\beta\hat{w}_n^4(y_n)}{q^2}H_n^T(H_n P_n^f H_n^T + R_n)^{-1}(y_n - H_n m_n^f)] \\ &= \frac{4\beta\hat{w}_n^4(y_n)}{q^2}\tilde{K}_n(y_n)H_n P_n^f H_n^T(H_n P_n^f H_n^T + R_n)^{-1}(y_n - H_n m_n^f) \\ &\quad - \frac{4\beta\hat{w}_n^4(y_n)}{q^2}P_n^f H_n^T(H_n P_n^f H_n^T + R_n)^{-1}(y_n - H_n m_n^f) \\ &= \frac{4\beta\hat{w}_n^4(y_n)}{q^2}\tilde{K}_n(y_n)H_n K_n(y_n - H_n m_n^f) - \frac{4\beta\hat{w}_n^4(y_n)}{q^2}K_n(y_n - H_n m_n^f) \\ &= \frac{4\beta\hat{w}_n^4(y_n)}{q^2}[\tilde{K}_n(y_n)H_n - \mathbf{1}_{d\times d}]K_n(y_n - H_n m_n^f) \\ &= \frac{4\beta\hat{w}_n^4(y_n)}{q^2}[\mathbf{1}_{d\times d} - \tilde{K}_n(y_n)H_n]K_n(H_n m_n^f - y_n) \\ &= \frac{4\beta\hat{w}_n^4(y_n)}{q^2}P_n^a(P_n^f)^{-1}K_n(H_n m_n^f - y_n)\end{aligned} \tag{2.54}$$

or

$$\begin{aligned}(c) &= P_n^a[2\beta\nabla\cdot(w(y_n)w(y_n)^T R_n^{-1}H_n)] \\ &= P_n^a H_n^T\nabla(2\beta\hat{w}_n^2(y_n)) = (\nabla(2\beta\hat{w}_n^2(y_n))H_n P_n^a)^T\end{aligned} \tag{2.55}$$

with $K_n = P_n^f H^T(H_n P_n^f H_n^T + R_n)^{-1}$ the usual Kalman gain. Again, notice the scaling factor and square on the diffusion weight. The first result is very interesting in its structure showing a relation to the regular Kalman gain for this particular choice of kernel with standardization via Mahalanobis distance. Furthermore, as similar result may be expected for any kernel employing Mahalanobis distance in some inner function via a chain rule in the derivative. Next to the explicit diffusion weight in the factor, it is hereby also implicitly contained in the analysis covariance matrix. For its simplicity and interpretation we want to maintain and highlight the second form including the gradient .

Taking everything together for the analysis mean formula, we obtain

$$\begin{aligned}
m_n^a &= P_n^a \theta_n^a \\
&= m_n^f - \tilde{K}_n(y_n) H_n m_n^f + \tilde{K}_n(y_n) y_n - 2\beta P_n^a H_n^T \nabla \hat{w}_n^2(y_n) \\
&= m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n) - 2\beta P_n^a H_n^T \nabla \hat{w}_n^2(y_n), \\
&= m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n) \\
&\quad - \frac{4\beta \hat{w}_n^4(y_n)}{q^2} P_n^a (P_n^f)^{-1} K_n (H_n m_n^f - y_n) \quad \text{or} \\
&= m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n) - P_n^a H_n^T \nabla(2\beta \hat{w}_n^2(y_n)),
\end{aligned} \tag{2.56}$$

for equivalent forms of the mean update recovering the usual Kalman filter analysis mean update up to the adjusted observation-dependent Kalman gain and with the additional diffusion term.

With covariance and mean available, we can now state the update formula for the $\mathcal{D}_w$-Kalman filter as in (2.7) and (2.8) with forecast step

$$\begin{aligned}
P_n^f &= A_n P_{n-1}^a A_n^T + Q_n \\
m_n^f &= A_n m_{n-1}^a
\end{aligned} \tag{2.57}$$

and analysis step

$$\begin{aligned}
q^2 &= \mathbb{E}[\chi^2(p)] = p \\
\Sigma_n &= H_n P_n^f H_n^T + R_n \\
\hat{w}_n(y_n) &= \Big(1 + \frac{||y_n - H_n m_n^f||^2_{\Sigma_n^{-1}}}{q^2}\Big) \\
\frac{\partial \hat{w}_n^2(y_n)}{\partial (y_n)_i} &= -\frac{2\langle \Sigma_{n,i\bullet}^{-1}, (y_n - H_n m_n^f)\rangle}{q^2} \hat{w}_n^4(y_n) \\
w_n(y_n) &= R_n^{\frac{1}{2}} \hat{w}_n(y_n) \\
N_n(y_n)^{-1} &= 2\beta \hat{w}_n^2(y_n) R_n^{-1} \iff \frac{1}{2\beta \hat{w}^2(y_n)} R_n \\
\tilde{K}_n(y_n) &= P_n^f H_n^T [N_n(y_n) + H_n P_n^f H_n^T]^{-1} \\
P_n^a &= P_n^f - \tilde{K}_n(y_n) H_n P_n^f \\
m_n^a &= m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n) - P_n^a H_n^T \nabla(2\beta \hat{w}_n^2(y_n)).
\end{aligned} \tag{2.58}$$

The result is the full desired recursion formula for updating the Gaussian mean and covariance .

Gathering Obtained Results

The overall result of this part now consists of two propositions and an algorithm - the first proposition presents the diffusion score matching posterior in the Kalman filter setting, the second proposition presents the global bias robustness property and its conditions and finally everything is combined into the $\mathcal{D}_w$-Kalman algorithm.

Recall the Kalman setting:
Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let X_n be a multivariate random variable to model our noisy signal at discrete time steps $n = \{1, 2 \ldots, N\}$. X_n cannot be observed directly, however, we can measure it via another random variable $Y_n = g_n(X_n, V_n)$, the observation, with V_n denoting the observation noise. Given the Kalman filter setting we assume X_n and Y_n to be jointly Gaussian with the following linear time discrete, time varying signal evolution equation and linear observation equation:

$$\begin{aligned} X_n &= A_n X_{n-1} + C_n W_n \\ Y_n &= H_n X_n + \Gamma_n V_n \end{aligned} \tag{2.59}$$

with

- $X_n : \beta \to \mathcal{X} = \mathbb{R}^d$ - the d-dimensional signal random vector at time n,
- $Y_n : \beta \to \mathcal{Y} = \mathbb{R}^p$ - the p-dimensional observation random vector at time n,
- $W_n : \beta \to \mathbb{R}^d$ and $V_n : \beta \to \mathbb{R}^p$ - independent standard Gaussian distributed random vectors at time n (white noise) of the corresponding dimensions,
- A_n, C_n, H_n and Γ_n of appropriate dimensions with non-singular $Q_n = C_n C_n^T$ and $R_n = \Gamma_n \Gamma_n^T$ and
- $p(x_0) \sim n(x_0; m_0, P_0^a)$, the initial prior distribution.

Proposition 1 ***The Diffusion Score Matching Bayes Posterior***
Given the above setting, then

$$\begin{aligned} p_\beta^{\mathcal{D}_w}(x_n|y_{1:n}) &\propto p(x_n|y_{1:(n-1)}) \cdot \exp[-\beta \cdot \hat{\mathcal{D}}_w(y_n; x_n)] \\ &\propto \exp[-\frac{1}{2} x_n^T J_n^a x_n + x_n^T \theta_n^a] \end{aligned}$$

for

$$\begin{aligned} J_n^a &= J_n^f + 2\beta \Lambda_n(y_n) \\ \theta_n^a &= \theta_n^f - 2\beta \nu_n(y_n) \end{aligned}$$

with

$$\begin{aligned} \Lambda_n(y_n) &= H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} H_n \\ \nu_n(y_n) &= -H_n^T R_n^{-1} w(y_n) w(y_n)^T R_n^{-1} y_n + \nabla \cdot (w(y_n) w(y_n)^T R_n^{-1} H_n) \end{aligned}$$

leading to a Gaussian distribution $p_\beta^{\mathcal{D}_w}(x_n|y_{1:n}) \sim n(x; m_n^a, P_n^a)$ *with*

$$
\begin{aligned}
&\Sigma_n = (H_n P_n^f H_n + R_n),\ q > 0\\
&\hat{w}_n(y_n) = (1 + \frac{||y_n - H_n m_n^f||^2_{\Sigma_n^{-1}}}{q^2})^{-\frac{1}{2}}\\
&w_n(y_n) = R_n^{\frac{1}{2}} \hat{w}_n(y_n)\\
&N_n(y_n)^{-1} = 2\beta \hat{w}_n^2(y_n) R_n^{-1}\\
&\tilde{K}_n(y_n) = P_n^f H_n^T [N_n(y_n) + H_n P_n^f H_n^T]^{-1}\\
&P_n^a = P_n^f - \tilde{K}_n(y_n) H_n P_n^f\\
&m_n^a = m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n) - 2\beta P_n^a H_n^T \nabla \hat{w}_n^2(y_n).
\end{aligned}
$$

The *proof* is given via the construction in section 2.1.2 and the beginning of 2.2.2.

Proposition 2 ***Global Bias Robustness***
Given the above setting, then $p_\beta^{\mathcal{D}_w}(x_n|y_{1:n})$ *is globally bias robust if* $w : \mathcal{Y} \to \mathbb{R}^{p \times p}$ *is chosen such that*

$$w_n(y_n) = R_n^{\frac{1}{2}} \hat{w}_n(y_n)$$

with

$$\hat{w}_n(y_n) = (1 + \frac{||y_n - H_n m_n^f||^2_{(H_n P_n^f H_n + R_n)^{-1}}}{q^2})^{-\frac{1}{2}}$$

and $q > 0$.

The *proof* is given in section 2.2.1 via the bound on the double supremum of the posterior influence function. Adaptations of the diffusion matrix w are possible and need exploring.

2.3. Interpretation and Extension

2.3.1. Understanding Novel Terms

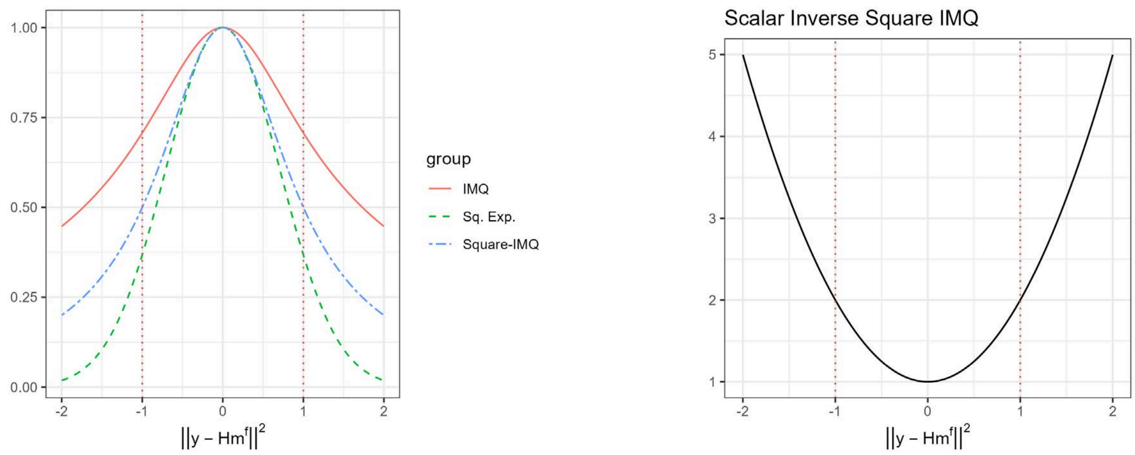
With proposition 2 providing the desired robustness property and algorithm 1 summarizing the necessary parameters and computations, we now want to focus on concrete interpretations of the novel terms and changes with some of them implicit in

Algorithm 1 The Diffusion Score Matching Kalman Filter

```
  Input:
    • initial condition $p(x_0) \sim n(x_0; m_0^a, P_0^a)$,
    • signal model $X_n = A_n X_{n-1} + Q_n W_n$,
    • observation model $Y_n = H_n X_n + R_n V_n$,
    • learning rate $\beta := 1 > 0$ (default) and
    • outlier threshold $q^2 := \mathbb{E}[\chi^2(p)] = p > 0$ (default).
  Output:
    • signal forecast at time $n$
          $p(x_n | y_{1:(n-1)}) \sim n(x_n; m_n^f, P_n^f)$
      and
    • signal posterior at time $n$
          $p(x_n | y_{1:n}) \sim n(x_n; m_n^a, P_n^a)$
for $n \geq 1$ do
  forward step:
      $P_n^f = A_n P_{n-1}^a A_n^T + Q_n$
      $m_n^f = A_n m_{n-1}^a$
  receive observation: $y_n$
  analysis step:
      $\Sigma_n = H_n P_n^f H_n^T + R_n$
      $\hat{w}_n(y_n) = (1 + \frac{||y_n - H_n m_n^f||_{\Sigma_n^{-1}}^2}{q^2})^{-\frac{1}{2}}$
      $N_n(y_n) = \frac{1}{2\beta \hat{w}_n^2(y_n)} R_n$
      $\tilde{K}_n(y_n) = P_n^f H_n^T [N_n(y_n) + H_n P_n^f H_n^T]^{-1}$
      $P_n^a = P_n^f - \tilde{K}_n(y_n) H_n P_n^f$

      $K_n(y_n) = P_n^f H_n^T [R_n + H_n P_n^f H_n^T]^{-1}$
      $m_n^a = m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n)$
          $- \frac{4\beta \hat{w}_n^4(y_n)}{q^2} P_n^a (P_n^f)^{-1} K_n (H_n m_n^f - y_n)$
      or
      $\frac{\partial 2\beta \hat{w}_n^2(y_n)}{\partial (y_n)_i} = -\frac{4\beta \langle \Sigma_{n,i\bullet}^{-1}, (y_n - H_n m_n^f) \rangle}{q^2} \hat{w}_n^4(y_n)$
      $m_n^a = m_n^f - \tilde{K}_n(y_n)(H_n m_n^f - y_n) - P_n^a H_n^T \nabla(2\beta \hat{w}_n^2(y_n))$
end for
```

the algorithm but explicit in proposition 1.
There are no changes in the forecast step which may leave room for usual adaptations to non-linearity in the signal propagation. We only adapted the Bayesian inverse inference for the posterior involved in the analysis step. The newly introduced terms are the substitute $N_n(y_n) \propto R_n$ for the observation noise covariance in the adapted Kalman gain and the divergence term introduced with the analysis mean update. Both share an interesting relationship in their interaction with the diffusion weight and provide a good intuition on their contribution to estimating mean and covariance. On a side note, the diffusion score matching Kalman gain is now explicitly dependent on the observation. While this might be highly desirable in some contexts, it also disables pre-computation of the Kalman gain and covariance matrices. Further, adaptations similar to the steady state Kalman filter (see (Stannat, 2023) for details) are less accessible.



(a) Inverse multi-quadratic kernel and squared exponential kernel.

(b) Scalar local inverse of the square of the IMQ kernel.

Figure 2.1.

Recall the IMQ-kernel as in figure (2.1a)

$$\hat{w}_n(y_n) = (1 + \frac{||y_n - H_n m_n^f||^2_{\Sigma_n^{-1}}}{q^2})^{-\frac{1}{2}}$$

with threshold $q^2 > 0$, its square and the scalar inverse of the square (see figure (2.1b). It was chosen due to its desirable behavior in fairly heavy tails to not straight up delete outliers, especially compared to squared exponential kernels, yet weight them down for satisfying robustness and keeping required properties in easily accessible derivatives. The threshold q^2 allows for easy control of the kurtosis and this way balancing which observations are considered outliers and to what degree. Yet, it only sparsly appears directly in the diffusion score matching Kalman equations. In the new analysis step it only directly appears squared, so in scale of the variance which makes sense considering its objective. Recall the new Kalman gain

$$\tilde{K}_n(y_n) = P_n^f H_n^T [N_n(y_n) + H_n P_n^f H_n^T]^{-1}$$

with

$$N_n(y_n)^{-1} = 2\beta \hat{w}_n^2(y_n) R_n^{-1} \iff N_n(y_n) = \frac{1}{2\beta \hat{w}_n^2(y_n)} R_n.$$

Here, the re-scaled inverse IMQ kernel in $N_n(y_n)$ directly scales the covariance matrix R_n. For outliers, so $||y_n - H_n m_n^f||_{\Sigma^{-1}}^2 \gg q^2$, this term blows up and inflates the observation noise covariance matrix, see figure 2.1b, and accordingly decreases the Kalman gain to a minimum putting all the trust in the forecast mean. For observations with Mahalanobis distance close to the expected error, so $||y_n - H_n m_n^f||_{\Sigma^{-1}}^2 \approx q^2$, $N_n(y_n)$ approximately recovers the the regular observation noise covariance matrix R_n and thus the regular Kalman filter behavior. An interesting new insight is in that for observations and observation forecast mean aligning, so $||y_n - H_n m_n^f||_{\Sigma^{-1}}^2 \ll q^2$, the original covariance matrix is scaled down by up to a factor of $\frac{1}{2\beta}$. In practice, this translates to putting more trust in these observations via increasing the adapted Kalman gain compared to the regular Kalman gain. This has interesting implications on the long term stability and need further exploring in that regard. The update of the analysis or posterior mean and covariance directly use this observation dependent calibration. Looking at the analysis mean update in more detail, for observations close to the forecast mean, so within the confidence interval, the adapted Kalman gain increases and the innovation term is falling much more into weight for the mean update. In short, the adapted Kalman gain utilizes the observation to dynamically re-scale the variance according to its reliability. Outliers are weighted down so their innovation term does not distort the mean estimate while plausible observations y_n are processed similar to the regular Kalman filter with small observation noise covariance R_n. This is directly reflected in the analysis covariance update. Suspected outliers lead to a small $\tilde{K}_n(y_n)$ and hence to a small reduction from forecast covariance to analysis covariance while plausible observations lead to a reasonable adjustment that may surpass the reduction in signal covariance matrix of the regular Kalman filter.

The newly introduced divergence term makes also use of this notion of plausibility. Taking the second formulation in (2.54) $P_n^a H^T \nabla \hat{w}_n^2(y_n)$. The resulting vector additionally steers the mean update in the analysis step processing the observation implicitly via the analysis covariance and explicitly via the gradient $\nabla \hat{w}_n^2(y_n)$. The actual adjustment on the analysis mean is best understood entry-wise:

$$\begin{aligned}(P_n^a H_n^T \nabla \hat{w}_n^2(y_n))_k &= \sum_{i=1}^{d} (P_n^a H_n^T)_{ki} (\nabla \hat{w}^2(y_n))_i \\ &= \langle (P_n^a H_n^T)_{k,\bullet}, \nabla \tilde{w}^2(y_n) \rangle \\ &= \langle (H_n P_n^a)_{\bullet,k}, \nabla \tilde{w}^2(y_n) \rangle\end{aligned} \tag{2.60}$$

recalling $A_{k,\bullet}$ denoting the respective k-th row or column vector of some matrix A. Writing the divergence vector like this provides insight on what is happening. Looking at the partial derivatives of the squared IMQ-kernel in 2.43, it explicitly introduces the plausibility of an observation y_n as the driving force via the gradient. On the other hand, $(H_n P_n^a)_{ik} = \sum_{j=1}^{d} (H_n)_{ij} (P_n^a)_{jk}$, so the i-th entry of the k-th column is the combination of all analysis covariances involving the k-th dimension of the signal effecting the i-th dimension of the observation via the forward map H_n. Accordingly, the column vector $(H_n P_n^a)_{\bullet,k}$ collects the entry-wise summed effects of the analysis covariances involving entry k in P_n^a over H_n, so $\mathrm{Cov}((H_n[X_n|y_{1:n}])_i, [X_n|y_{1:n}]_k)$ - a notion of total variation effects on entry i of the observation from signal entry k. The dot product now sums over the collected effects of the covariances involving signal entry k in P_n^a produced by H_n and the partial derivatives, resulting in the observed impact of the divergence on the mean estimate. To put it different, the step size $(H_n P_n^a)_{ik}$ and the gradient involving y_n describe a vector field of flow summarizing all impacts on the k-th entry of the signal analysis mean caused by covariances in P_n^a given all observations $y_{1:n}$ via the forward map H_n, $\sum_{i=1}^{p} \frac{\partial \hat{w}^2(y_n)}{\partial (y_n)_i} \mathrm{Cov}((H_n[X_n|y_{1:n}])_i, [X_n|y_{1:n}]_k)$, so the sum over the corresponding partial derivatives inferring a notion of plausibility of the covariance effects via the observation, to describe the total change in flow density for that dimension - the initial divergence in a classical sense as change of density of a liquid. To conclude, the new divergence term combines all information we have about change in the k-th entry of the signal analysis mean encoded in the analysis covariances and combines them with direction and impulse strength via plausibility of a observation y_n to nudge and adjust the mean estimate. For plausible observations, the entries of P_n^a will be smaller due to the adjusted Kalman gain allowing for reduction in analysis covariance compared to the forecast covariance, and the gradients will have larger values (see figure 2.1a). For suspected outliers it is the other way around. The key insight is in that the divergence term takes its information from where it deems most reliable, either the gradient or the forecast variance with the sign of the gradient flow always decided by the observation regardless of plausibility to steer the mean update even for supposedly unreliable observations.

2.3.2. Considerations on Long-Term Stability and Learning Rate

Approaches to Long-Term Stability

With the changes in the recursive formula well understood, especially regarding the novel covariance update, we want to address the long term stability of the filter. The approach as with the regular Kalman filter in solving an algebraic Riccati-equation does not immediately work here due to the introduced stochasticity via the observation dependent Kalman gain. Again, we want to utilize the concept of ε-contamination to describe observations produced by heavy tails and outliers for a basic intuition. Generally speaking, the result for the regular Kalman filter still holds for observations with Mahalanobis distance within the χ^2 expectation and even improves with the down-scaling in noise covariance for observations close to the observation forecast mean. The regular observations with large Mahalanobis distance as well as observations produced by contamination are what is falling into weight and destabilize the covariance. However, similar issues have already arisen with the variety of other adaptive filtering techniques and alongside them approaches to formulate adjusted stability requirements. In works such as (Solo, 1996), (Zhen-Wei and Hai-Tao, 2013) and (Gan and Liu, 2020) the long-term stability of the Kalman filter with random or faulty coefficients is investigated. This motivates an interpretation of the adapted observation noise covariance $N_n(y_n)$ to be such a random parameter with certain aspects of pattern and scaling known. The key insight is then that $N_n(y_n)$ replacing R_n in the discrete algebraic Riccati equation occurring in the stability analysis is only stochastically bound for observation distributions with finite second moments and cross-terms. Deriving the weak stochastic bound on $N(Y)$ for arbitrary random vectors Y leads to deriving a weak stochastic bound on $\frac{1}{\hat{w}^2(Y)}$, a uni-variate non-negative random variable. Accordingly, a bound can then be derived using Markov's inequality with a simplified notation in

$$\begin{aligned}\mathbb{P}[\frac{1}{\hat{w}^2(Y)} \leq b] = \mathbb{P}[(1 + \frac{||Y||^2_{\Sigma^{-1}}}{q^2}) \leq b] \geq \frac{\mathbb{E}[(1 + \frac{||Y||^2_{\Sigma-1}}{q^2})]}{b} &= \frac{1}{b} + \frac{1}{bq^2}\mathbb{E}[||Y||^2_{\Sigma^{-1}}] \\ &\overset{+C}{\propto} \frac{1}{b}\mathbb{E}[Y^T\Sigma^{-1}Y] = \frac{1}{b}\mathbb{E}[\sum_{i=1}^{p}\sum_{j=1}^{p} Y_i\Sigma^{-1}_{ij}Y_j] \\ &\propto \frac{1}{b}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}[Y_iY_j]\end{aligned} \tag{2.61}$$

$$\begin{aligned}&\iff \mathbb{P}[\frac{1}{\hat{w}^2(Y)} \geq b] \leq \frac{\mathbb{E}[||Y||^2_{\Sigma^{-1}}]}{b} \\ &\implies \lim_{b\to\infty}\mathbb{P}[\frac{1}{\hat{w}^2(Y)} \leq b] \leq \lim_{b\to\infty}\frac{\mathbb{E}[||Y||^2_{\Sigma^{-1}}]}{b} = 0\end{aligned} \tag{2.62}$$

for $\mathbb{E}[||Y||^2_{\Sigma^{-1}}] < \infty$.
This cheap bound only provides limited insight and is by no means exhaustive, however, it still conveys a first idea: If all entries of the SSCP matrix YY^T are finite, the expected Mahalanobis distance is finite and we obtain a weak stochastic bound which then is sufficient for long-term stability of the filter with results in (Solo, 1996) under usual assumptions and conditions for stability in detectability and stabilizability of the Kalman filter components. However, it is the class of super-heavy tailed distributions such as Cauchy distributions or t-distributions with $\mathrm{df} \leq 2$ which have non-finite second moments and therefore non-finite expected Mahalanobis distance we are also interested in. They are likely not stochastically bound and their impact on the stochastic discrete algebraic Riccati equation needs further investigation. To summarize, we observe two cases: If $N_n(y_n)$ has a weak stochastic bound, i.e via the Markov inequality with finite second moments, we can substitute R_n with this bound in usual stability arguments with conditions on detectability and stabilizabilty of the signal to recover long-term stability. If there is no such weak stochastic bound on $N_n(y_n)$, there is essentially no reliable reduction of the covariance in the analysis step and stability solely depends on the asymptotic stability of the signal process with usual conditions. For rigorous analysis of long-term stability of the Kalman filter with the respective conditions see (Anderson and Moore, 2012) and (Stannat, 2023).

This problem may then be generalized to the more broad discussion of stability in combination with robustness via diffusion score matching posteriors. In order for the likelihood to be of use, an observation y_0 needs at least linear order representation. To obtain robustness in the posterior influence function, the loss function needs to weight down observations of appropriate order of y_0 in the likelihood, also considering the derivative in the score matching loss. This is what makes the approach in (Altamirano et al., 2023b) strong as they restrict their investigated exponential family members to linear order representations of the observation in the likelihood. As long as a weight balancing with at least linear order then appears in squared scalar inverse form in the covariance update, we necessarily end up with existence requirements for higher order moments such as here regarding finite expected Mahalanobis distance. In short, the price we pay for obtaining provable robustness via the posterior influence function while still aiming to obtain as many information from an observation as possible is likely at least partially paid in stability. If outliers are too frequent, so the respective moments of the true data generating process are not finite, stability has to come solely with system dynamics. This brief and superficial intuition is by no means rigorous or exhaustive and thorough sophisticated investigation has yet to be done.

Tuning the Learning Rate β

We also want to pick up on discussing the learning rate β. In the final equations in proposition 1 it only appears coupled to the squared diffusion weight. A default choice of $\beta = 1$ is reasonable to recover the desired intuitions as scaling of the observation noise covariance is related to expected Mahalanobis distance. Yet, a brief discussion is in place. A large learning rate causes the inverse substitute $N_n(y_n)$ to be toned down in value reducing its impact on the adjusted Kalman gain and thus supports trust in observation - likely at risk of over-fitting fr too large choices of β. Regarding the analysis covariance this leads to a larger reduction from the forecast covariance and similarly to a larger impact of the innovation term on the analysis mean. Further, a large learning rate increases the impact of the divergence term, however, this also needs considering the then smaller analysis covariance. On the other hand, a too small learning rate $0 < \beta \ll 1$ is prone to over-smoothing and under-fitting in that it further inflates the observation covariance matrix and reduces weight of recent observations. Accordingly, the learning part can be taken to be an important parameter that needs considering.

Recall, tuning of the learning rate is an open problem in generalised Bayesian inference motivating several approaches with ongoing research (see (Lyddon et al., 2019), (Matsubara et al., 2023) and (Wu and Martin, 2023) among others for details). The issue is picked up in (Altamirano et al., 2023b) stating for the learning rate β to offset sensitivity in q^2. This makes intuitive sense in that tuning β much different from its default makes the choice in threshold q^2 somewhat arbitrary. In that case then, the choice of q^2 is fairly robust while requiring strong arguments for tuning β. In (Altamirano et al., 2023a) they suggest fixing β and solely focusing on q^2 via directly coupling both parameters. Our setting allows to do so as well to similar extend, yet we want to also discuss the approach to tuning β in (Altamirano et al., 2023b) regardless. There they argue that most developed methods are computationally expensive and usually strive to satisfy frequentist, asymptotic properties. These do not apply for the overarching purpose of detecting change in the signal caused by mis-specification of the signal noise. Instead they suggest matching the uncertainty of the diffusion score matching posterior to that of the regular posterior counterpart ideally on a controlled/un-contaminated set of N^* observations. Depending on the context, it may be feasible to simulate a certain number of N^* observations from the assumed model not taking contamination into consideration and matching both posteriors on the simulated data. This is explicitly possible for the Kalman setting as we have access to an assumed data generating process to simulate from and we will therefore implement it this way in the experiments. In (Altamirano et al., 2023b) Kullback-Leibler divergence is suggestedthe suggested measure for matching as it is reliable in the absence of outliers, with the authors claiming to produce well calibrated uncertainties this way. Further, the regular Bayesian posterior and the regular Kalman filter are optimal in that case and therefore offer an easy choice as reference. To formalise, we want to pick β^* such that

$$\beta^* = \arg\min_{\beta>0} \mathcal{D}_{\mathrm{KL}}(p(x_{N^*}|y_{1:N^*})||p_\beta^{\mathcal{D}_w}(x_{N^*}|y_{1:N^*})). \tag{2.63}$$

This argument makes heuristic sense for the Kalman setting and furthermore, as we deal with Gaussian distributions for both the diffusion score matching posterior and the regular posterior, the evaluation can be done via the closed formula following (Duchi, 2007):

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(p(x_{N^*}|y_{1:N^*})||p_\beta^{\mathcal{D}_w}(x_{N^*}|y_{1:N^*})) &= \mathbb{E}_{X\sim p(\cdot|y_{1:N^*}))}[\log(p_\beta^{\mathcal{D}_w}(x_{T^*}|y_{1:N^*})) - \log(p(x_{T^*}|y_{1:N^*}))] \\ &= \frac{1}{2}\log(\frac{\det(P^a_{N^*})}{\det(P_{N^*})}) - d \\ &\quad + \mathrm{Tr}(J^a_{N^*}P_{N^*}) + (m^a_{N^*} - m_{N^*})^T J^a_{N^*}(m^a_{N^*} - m_{N^*}).\end{aligned} \tag{2.64}$$

The convexity of the KL divergence for fixed reference distribution is hereby an additional useful feature for optimization.
Instead of minimizing Kullback-Leibler divergence, maximizing mutual information also seems like a reasonable heuristic. While the general problem in application is that it is an open problem to estimate it reliably with good uncertainty quantification, there are analytic derivations of the mutual information of two multivariate Gaussian random variables (see (Carrara and Ernst, 2020) for details). The main issue arising then with the result applicable to the case at hand is in requiring the covariance matrix of the joint distribution of $p(x_{N^*}|y_{1:N^*})$ and $p_\beta^{\mathcal{D}_w}(x_{N^*}|y_{1:N^*})$. So while it might not be as suitable for tuning the diffusion score matching Kalman filter right away, it has potential via sampling and estimating the joint covariance from joint samples.

Another, likely more fruitful approach to choosing the learning rate β arises in the Kalman setting via aiming to recover the regular Kalman filter and its learning rate on average, i.e. $\mathbb{E}[2\beta\hat{w}^2(Y)] \approx 1$, assuming no contamination. A much more sophisticate approach would then employ this result for an analysis on $\mathbb{E}[\tilde{K}(Y)]$. The case here is mainly a preliminary result in that regard and lays ground for future work.

A straight forward observation lies then in that we recover this desired result for a choice of $\beta = 1$ up to tightness of a Jensen's inequality. Let $p(y) = n(y; Hm^f, \Sigma)$, $\mathcal{Y} = \mathbb{R}^p$ and $q^2 = p$, then

$$\begin{aligned} 2\hat{w}^2(y) &= 2(1 + \frac{||y - Hm^f||^2_{\Sigma^{-1}}}{p})^{-1} \\ &= \frac{2}{1+z} = f(z) \end{aligned} \tag{2.65}$$

with $||Y - Hm^f||^2_{\Sigma^{-1}} \sim \chi^2(p)$ hence $Z \sim \frac{||Y-Hm^f||^2_{\Sigma^{-1}}}{p} = \text{Gamma}(k = \frac{p}{2}, \theta = \frac{2}{p})$ for shape k and scale θ with $\mathbb{E}[Z] = 1$ and $Var(Z) = \frac{2}{p}$. Further, f is convex and $\frac{\mathrm{d}^2 f(z)}{\mathrm{d}^2 z} = \frac{4}{(1+z)^3} > 0$ for $z \geq 0$. Accordingly, the expectation of interest can be reformulated to $\mathbb{E}[f(Z)] \geq f(\mathbb{E}[Z]) = 1$, hence the motivation of $\beta = 1$ up to that tightness of a Jensen's inequality. Utilizing results in (Liao and Berg, 2018) for sharpness of the inequality exploiting Taylor approximations and curvature, we obtain the valuable insight

$$\begin{aligned} 0 = 0 \cdot Var(Z) \leq \mathbb{E}[f(Z)] - 1 \leq \frac{1}{2}Var(Z) = \frac{1}{p} \\ \iff f(\mathbb{E}[Z]) \leq \mathbb{E}[f(Z)] \leq f(\mathbb{E}[Z]) + \frac{1}{p} \end{aligned} \tag{2.66}$$

via $\inf$ and $\sup$ of a help function $h(z) = \frac{f(z) - f(\mathbb{E}[Z])}{[z - f(\mathbb{E}[Z])]^2} - \frac{f'(\mathbb{E}[Z])}{z - f(\mathbb{E}[Z])} = \frac{f(z)-1}{(z-1)^2} - \frac{f'(1)}{z-1}$.

For large observation dimensions p, we can take $1 = f(\mathbb{E}[Z]) \approx \mathbb{E}[f(Z)]$ and easily justify the choice of $\beta = 1$. However, for smaller dimensions it can make sense to evaluate the resulting integral in $\mathbb{E}[f(Z)] = \int_0^\infty f(z)p(z)\mathrm{d}z$ numerically. Taking the least ideal case in $p = 1$, we then obtain

$$\mathbb{E}[f(Z)] = \frac{2}{\sqrt{2\pi}} \int_0^\infty (1+z)^{-1} z^{-\frac{1}{2}} \exp[-\frac{z}{2}]\mathrm{d}z = \sqrt{2\exp(1)\pi}\,\mathrm{erfc}(\frac{1}{\sqrt{2}}) \approx 1.31, \tag{2.67}$$

with erfc the complementary error function and choose then $\beta \approx \frac{3}{4}$ to approximately recover $\mathbb{E}[2\beta\hat{w}^2(Y)] \approx 1$. This integral can be numerically evaluated for other small values of p, however, with the intuition on approaching equality with increasing observation dimension, we can conclude to obtain a learning rate $\beta \in [\frac{3}{4}, 1]$ to approximately recover expectation. For better understanding, additional investigation on the choice of learning rate regarding variance of the approximation is required and will be subject of future work next to embedding the obtained result for investigations on the expected adjusted Kalman gain.

2.3.3. Expanding to Ensemble Implementation

One more immediate adaptation we want to address is the extension to the field of sequential Monte Carlo methods, particle filters or ensemble methods. The ensemble Kalman filter (EnKF) with perturbed observations is popular and favored for a variety of applications over the regular Kalman filter due to not needing analytical evaluation of the forecast covariance. The diffusion score matching Kalman filter can easily be extended to suit the EnKF framework of a particle approximation that recovers the analytic filter in mean field limit under slight adjustments. Following (Reich and Cotter, 2015), we want to start with an ensemble $\{x_{n-1}^{a,(l)}\}_{l\in\{1,2,\dots,M\}}$ of size M sampled from $p(x_{n-1}|y_{1:(n-1)})$, so $X_{n-1}^{a,(1:M)} \sim_{iid} p(x_{n-1}|y_{1:(n-1)})$. The ensemble is propagated according to the signal model to produce the forecast ensemble $\{x_n^{f,(l)}\}_{l\in\{1,2,\dots,M\}}$. Again, up to this point we want to follow the regular EnKF procedure with the diffusion score matching EnKF as we only make changes in the next step to the Bayesian inverse inference problem. In the regular EnKF with perturbed observations the ensemble is updated with the new observation via

$$x_n^{a,(l)} = x_n^{f,(l)} - \hat{K}_n(H_n x_n^{f,(l)} + \epsilon_n^{(l)} - y_n) \tag{2.68}$$

for $l \in \{1,2,\dots,M\}$ with $\{\epsilon_n^{(l)}\}_{l\in\{1,2,\dots,M\}}$ independent and identically distributed draws from $\mathcal{N}(0,R_n)$ resulting in the updated ensemble $\{x_n^{a,(l)}\}_{l\in\{1,2,\dots,M\}}$. The corresponding mean and covariance are then evaluated empirically from the ensemble after the respective step providing estimates $\hat{m}_n^f$, $\hat{P}_n^f$, $\hat{m}_n^a$ and $\hat{P}_n^a$. Dependent quantities such as $\hat{K}_n$ or $\hat{\Sigma}_n$ are then calculated via these estimates. The EnKF with perturbed observations as shortly given here can be understood as a Monte Carlo implementation of a best, linear, unbiased estimator given the Kalman setting. The empirical estimators hold in the mean field approximation and recover the regular Kalman filter.

The main change for the diffusion score matching EnKF with perturbed observations from the regular D_w-KF is in substituting the forecast mean m_n^f with the individual members of the forecast ensemble $\{x_n^{f,(l)}\}_{l\in\{1,2,\dots,M\}}$ in the novel analysis step. This majorly influences $\hat{w}$ in that it now needs taking both y_n and $x_n^{f,(l)}$ as arguments and thus also impacts every computation implicitly and explicitly taking $\hat{w}$. The resulting equations with changes are then

$$\begin{aligned}
&\hat{w}_n(y_n;x_n^{f,(l)}) = (1+\frac{||y_n - H_n x_n^{f,(l)}||^2_{\Sigma_n^{-1}}}{q^2})^{-\frac{1}{2}}\\
&N_n(y_n;x_n^{f,(l)})^{-1} = 2\beta\hat{w}^2(y_n;x_n^{f,(l)})R_n^{-1}\\
&\tilde{K}_n(y_n;x_n^{f,(l)}) = \hat{P}_n^f H_n^T[N_n(y_n;x_n^{f,(l)}) + H_n\hat{P}_n^f H_n^T]^{-1}\\
&\tilde{P}_n^a(y_n;x_n^{f,(l)}) = \hat{P}_n^f - \tilde{K}_n(y_n;x_n^{f,(l)})H_n\hat{P}_n^f\\
&x_n^{a,(l)} = x_n^{f,(l)} - \tilde{K}_n(y_n;x_n^{f,(l)})[H_n x_n^{f,(l)} + \epsilon_n^{(l)} - y_n) - 2\beta\underset{(*)}{\tilde{P}_n^a(y_n;x_n^{f,(l)})}H_n^T\nabla_{y_n}\hat{w}_n^2(y_n;x_n^{f,(l)})
\end{aligned} \tag{2.69}$$

for $l \in \{1,2,\dots,M\}$. The corresponding estimators are then again taken empirically for mean and covariance after the forecast and analysis step. As the empirical analysis covariance in $(*)$ is not yet available, we need taking the individual analysis covariance matrix computed in the previous equation to update each ensemble member. The $\{\epsilon_n^{(l)}\}_{l\in\{1,2,\dots,M\}}$ are hereby now independent samples via $\epsilon_n^{(l)} \sim \mathcal{N}(0, N_n(y_n;x_n^{f,(l)}))$. Again, the result recovers the regular diffusion score matching Kalman filter in the mean field limit.

What is the intuitive upside of this approach? Next to not having to propagate the signal covariance from analysis to forecast, the ensemble members seem to explore the signal space and the signal dynamic much more. Ensemble members reasonably close to an observation are steered towards it while ensemble members much further away stay close to where they were propagated to in the forecast step. For the linear case we expect no major differences from the regular diffusion score matching Kalman filter. However, the approach is promising in managing non-linear signal dynamics under observation mis-specification concerns. Observations with large deviation from the forecast mean produced by chaotic behavior of the non-linear dynamic can easily be confused with outliers, however, the exploration of this chaotic behavior via the ensemble members can account for that to reasonable extend under uncertainty. The fairly direct problem is the loss of Gaussianity of the forecast distributions and thus the analysis distributions drawing into the realm of particle filters and sequential Monte Carlo. However, the EnKF is surprisingly helpful even then and the diffusion score matching EnKF with perturbed observation ideally keeps this property. Assuming not needing to tune the learning rate by keeping the default choice of $\beta = 1$, this avoids a major issue as the approach for the linear case in matching β in KL divergence to the regular Bayesian case is no longer reliable in non-linear system.

2.4. Experiments: Linear and Non-Linear Simulations

To showcase the theoretical approaches and results we conduct four experiments A-D each aiming to provide a different focus and insight. Hereby we are mainly interested in mean estimates after the analysis step as there were no changes done to the forecast step.

Experiment A focuses on the influence of the learning rate in a simple implementation of a one-dimensional Ornstein-Uhlenbeck process with contaminated observations. *Experiment B* implements the $\mathcal{D}_w$-Kalman filter for a four-dimensional target tracking model with contaminated observations to provide insight on unobserved velocities highlighting the robustness. *Experiment C* implements a linear Rössler-system with contaminated partial observations. The generated signal is estimated with an EnKF with perturbed observations and the $\mathcal{D}_w$-EnKF with perturbed observations with small ensemble size for direct comparison. Lastly, *Experiment D* employs the popular chaotic Lorenz-63 model and again puts the EnKF and the $\mathcal{D}_w$-EnKF next to each other while also providing first insights at performance of the $\mathcal{D}_w$-EnKF for non-linear systems.

The experiments are partially inspired and taken from (Reich and Cotter, 2015) and (Evensen et al., 2022). Similarly, we either use a forward Euler or Euler-Murayama scheme for simulation. Additional graphs on each experiment are provided in the appendix.

All simulations were done in R version 4.2.2.

2.4.1. Experiment A: One-Dimensional Ornstein-Uhlenbeck-Process

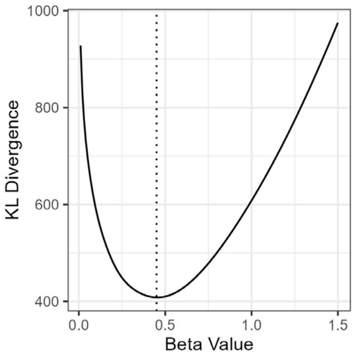


Figure 2.2.: Kullback-Leibler divergence of Kalman Filter and $\mathcal{D}_w$-Kalman filter with learning rate β after $T^* = 1.5$.

The first experiment aims to investigate the effect of tuning learning rates as well as showcasing the general robustness against observation contamination. For the model we chose a simple uni-variate Ornstein-Uhlenbeck-process discretized and simulated via an Euler-Murayama scheme - essentially resulting in a uni-variate AR(1)-process. The underlying model is given as

$$\mathrm{d}x_t = -\theta x_t \mathrm{d}t + \sigma \mathrm{d}W_t \tag{2.70}$$

with state x_t, initial value $x_0 = 5$, mean reversion parameter $\theta = 2.5$, volatility $\sigma = 2$ and standard Brownian motion W_t. We simulate over an interval $[0, 1.5]$ with step size $\delta t = 0.005$ for the discretized model

$$x_{t+1} = x_t - \delta t \theta x_t + \delta t \sigma \epsilon_t \tag{2.71}$$

with ϵ_t iid standard Gaussian noise. The contaminated observations are generated at $t_{out} = 1$, so at every time step via

$$y_t = x_t + \varepsilon_t^\lambda \tag{2.72}$$

with contaminated observation noise $\varepsilon_t^\lambda \sim \mathcal{N}(0, \sqrt{\delta t R}) + \lambda \mathcal{N}(0, \sqrt{5 \cdot \delta t R})$, observation variance $R = 5$ and contamination parameter $\lambda = 0.15$. In practice this resulted in generating y_t via drawing $q \sim \mathrm{Ber}(p = \lambda)$ with $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$ standard Gaussian noise and $\varepsilon_t^\lambda = \delta t R \varepsilon_t^{(1)} + q \delta t 10 R \varepsilon_t^{(2)}$.

The learning rate β was tuned by generating observations without contamination and matching the posterior of the $\mathcal{D}_w$-KF to to the one of the regular KF in KL-divergence at $T^* = 1.5$. The obtained curve is nicely convex leading to an approximate $\beta^* \approx 0.45$ (see figure 2.2). Notice hereby that this values is fairly different from the suggested choice of $\beta \approx \frac{3}{4}$ from the considerations on approximately recovering $\mathbb{E}[2\beta\hat{w}^2(Y)] \approx 1$. We also included the posterior mean estimates and variance for a default choice of $\beta = 1$ for comparison.

The results are shown in figure 2.3 (see figure A.1 and figure A.2 in the appendix for additional graphs). As can be seen, they align nicely with the considerations in the previous section on the learning rate. Especially comparing the choice of $\beta = 1$ with the tuned choice β^* we see only little impact on the adapted Kalman gain and the divergence term. The smaller learning rate leads to potentially larger values of $N_n(y_n)^{-1}$ and thus to smaller values of the adapted Kalman gain showing in slightly larger variance and reduced impact of the observation. The divergence term can then only provide little impulses. Yet, although tuning may improve performance of the $\mathcal{D}_w$-Kalman filter, the effect is mostly negligible and the desired robustness is also acquired with the default choice of $\beta = 1$. More important, the results agree with (Altamirano et al., 2023b) in that they produce well-calibrated uncertainties - fairly regardless of learning rates via balancing mean estimates and variances.

Regarding the main emphasis of this chapter, the outlier robustness or impact of mis-specification, this first experiment convincingly portrays the theoretical results in that contaminated observations do not majorly impact the mean estimates. This also reflects in the naive MSE evaluation over all estimates and the squared error plots (see A.3 and A.4), however, we have to keep in mind that the MSE is sensible especially to outliers itself and it was only aggregated over time, not over repeated simulations.

2.4.2. Experiment B: Two-dimensional Target Tracking

For the second experiment we choose a simplification of the popular target tracking task with signal space $\mathcal{X} = \mathbb{R}^4$ containing x-position, x-velocity, y-position as well as y-velocity, and observation space $\mathcal{Y} = \mathbb{R}^2$ containing the measured x-position, as well as y-position. We design the signal model discrete and instead focus for this experiment on the robustness in the unobserved velocity dimensions. Given the usual Kalman setting up to contamination, we have

$$\begin{aligned} X_{n+1} &= A_n X_n + Q_n^{\frac{1}{2}} W_n \\ Y_n &= H_n X_n + R_n^{\frac{1}{2}} V_n^{\lambda} \end{aligned} \tag{2.73}$$

with W_n standard Gaussian noise and V_n^{λ} contaminated noise in the respective dimensions. As usual for these kind of models, yet much more simplified, we choose $A_n = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, $H_n = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ and signal noise covariance $Q_n = 0.1 \cdot \begin{pmatrix} 1 & 0.5 & 0.5 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.5 & 0 & 1 & 0.1 \\ 0 & 0 & 0.1 & 1 \end{pmatrix}$. The initial signal is chosen as $X_0 = (0, 1, 0, 0)^T$. The contaminated observations are generated via $R_n^{\frac{1}{2}} V_n^{\lambda} \sim \mathcal{N}(0, R_n) + \lambda \mathcal{N}(0, 100 R_n)$ with $R_n = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}$ and $\lambda = 0.2$. The contaminated observations are generated as before. We simulate X_n for n in $[0, 2]$ with step-size $\mathrm{d}n = 0.01$ resulting in 200 positions with $n_{out} = 1$, so 200 observations. The learning rate β was chosen as its default 1. The object trajectory and analysis mean estimates for the regular Kalman filter simulation can be seen in figure A.5 and figure A.6 in the appendix.

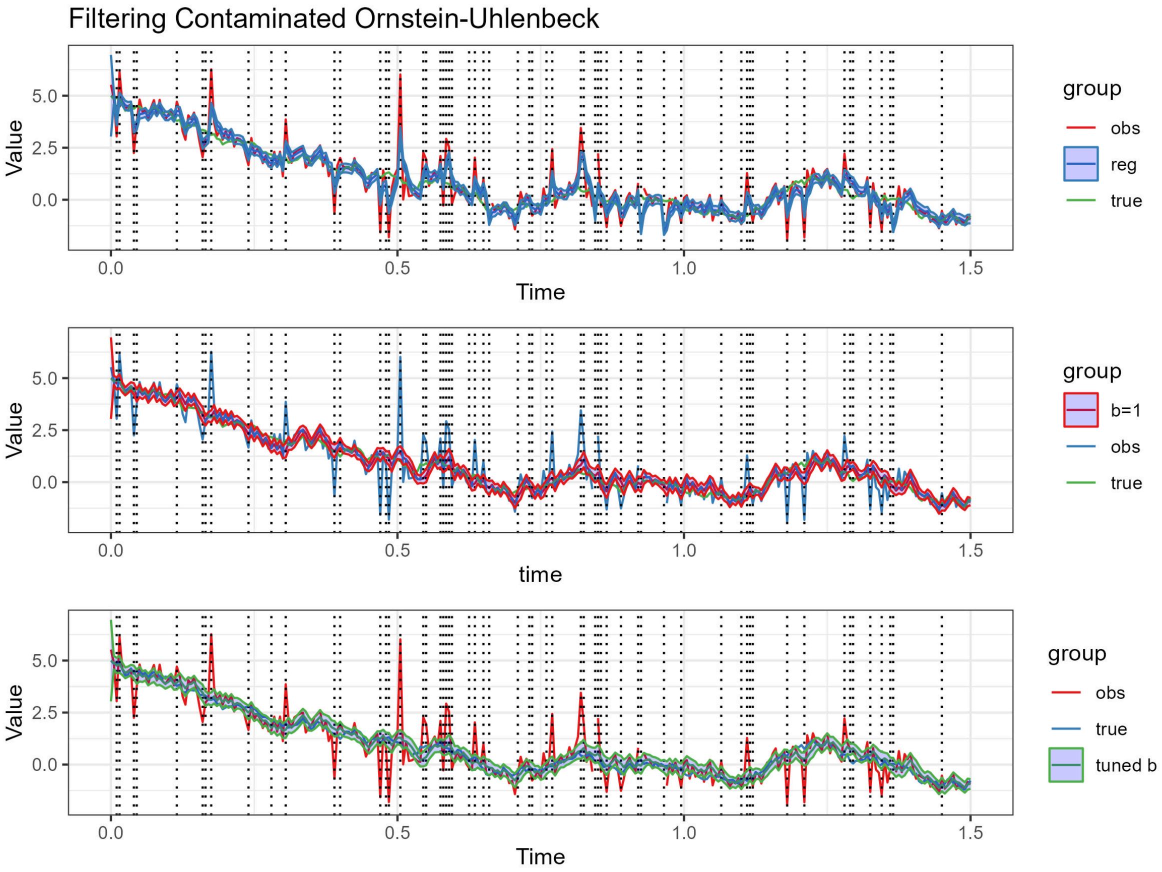


Figure 2.3.: Side-by-side graph comparison of the simulated signal and contaminated observations with the analysis/posterior mean estimates of the regular KF (top) and the $\mathcal{D}_w$-KF for $\beta = 1$ (middle) and tuned β^* (bottom) each with the respective 95%-CI. Dotted lines signal instances of observation contamination.

The results of the main experiment presented in figure 2.4 and figure 2.5 support the theoretical robustness of the $\mathcal{D}_w$-Kalman filter as the estimated trajectory is much more reliable in following the true simulated trajectory. Further, while the velocity dimensions are in much smaller scale compared to the positions, the impact of the contamination on the estimation of the regular Kalman filter shows clearly with the $\mathcal{D}_w$-Kalman filter being much more stable. Accordingly, the diffusion score matching posterior approach seems reliable under contamination of the observations or observation noise mis-specification.
Additional graphs for the regular Kalman filter with contaminated observations are provided in figure (A.5) and (A.6) in the appendix.

2.4.3. Experiment C: Rössler Model and Ensemble Simulation

With experiments A and B providing evidence for the theoretical results on robustness and the considerations on the effect of the learning, we want to focus on in-

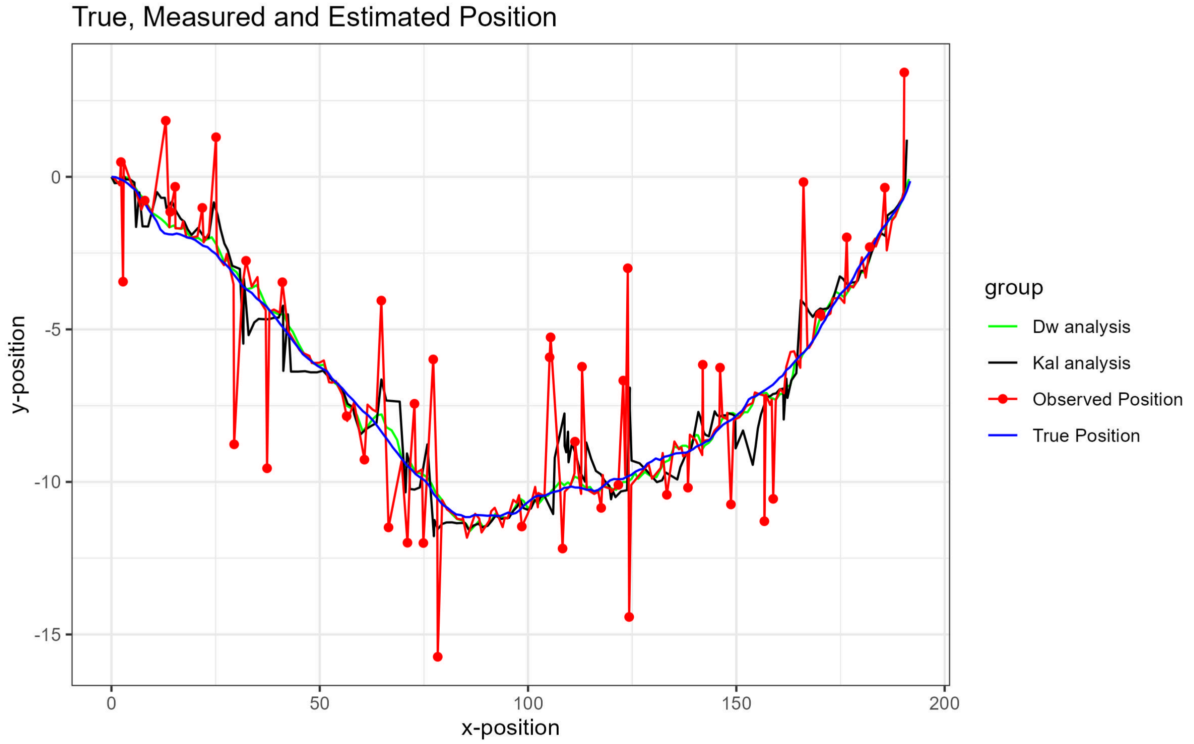


Figure 2.4.: True object trajectory in x and y-coordinate (blue), measured object trajectory (red) and estimated object trajectory for the regular KF (black) and tuned $\mathcal{D}_w$-KF (green).

vestigating the sketched diffusion score matching ensemble Kalman filter with perturbed observations with experiments C and D. A major issue hereby may lie in tuning the learning rate, however, with the previous results supporting good performance for a default choice of $\beta = 1$, we will repeat the choice here as well.

As an initial test for the $\mathcal{D}_w$-EnKF we chose the popular Rössler model with parameter choices such that it results in a linear system. We take this example from (Evensen et al., 2022) and implement it in similar fashion. We start with the linear Rössler matrix $M = \begin{pmatrix} 0 & -1 & -1 \\ 1 & a & 0 \\ 0 & 0 & -c \end{pmatrix}$ with $a = 0.1$ and $c = 0.05$ resulting in an oscillatory trajectory in the x-y-plane with reverting behavior in the z-coordinate. Again, we discretize and simulate the system via an Euler-Murayama scheme resulting in a signal propagation via $A = \mathrm{Id} - \delta t M$ and

$$X_{t-1} = AX_t + \delta t Q^{\frac{1}{2}} W_t \tag{2.74}$$

with $Q = 0.05 \begin{pmatrix} 1 & 0.1 & 0 \\ 0.1 & 1 & 0 \\ 0 & 0 & 0.005 \end{pmatrix}$, W_t standard Gaussian noise of appropriate dimension, step size $\delta t = 0.1$ and initial value $X_0 = (1, 1, 0)^T$. The observations are generated at $t_{out} = 10$ by

$$Y_t = HX_t + R^{\frac{1}{2}} V_t^{\lambda} \tag{2.75}$$

with $R = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ so that the second dimension cannot be observed directly and V_t^{λ} is contaminated standard Gaussian noise as before with variance

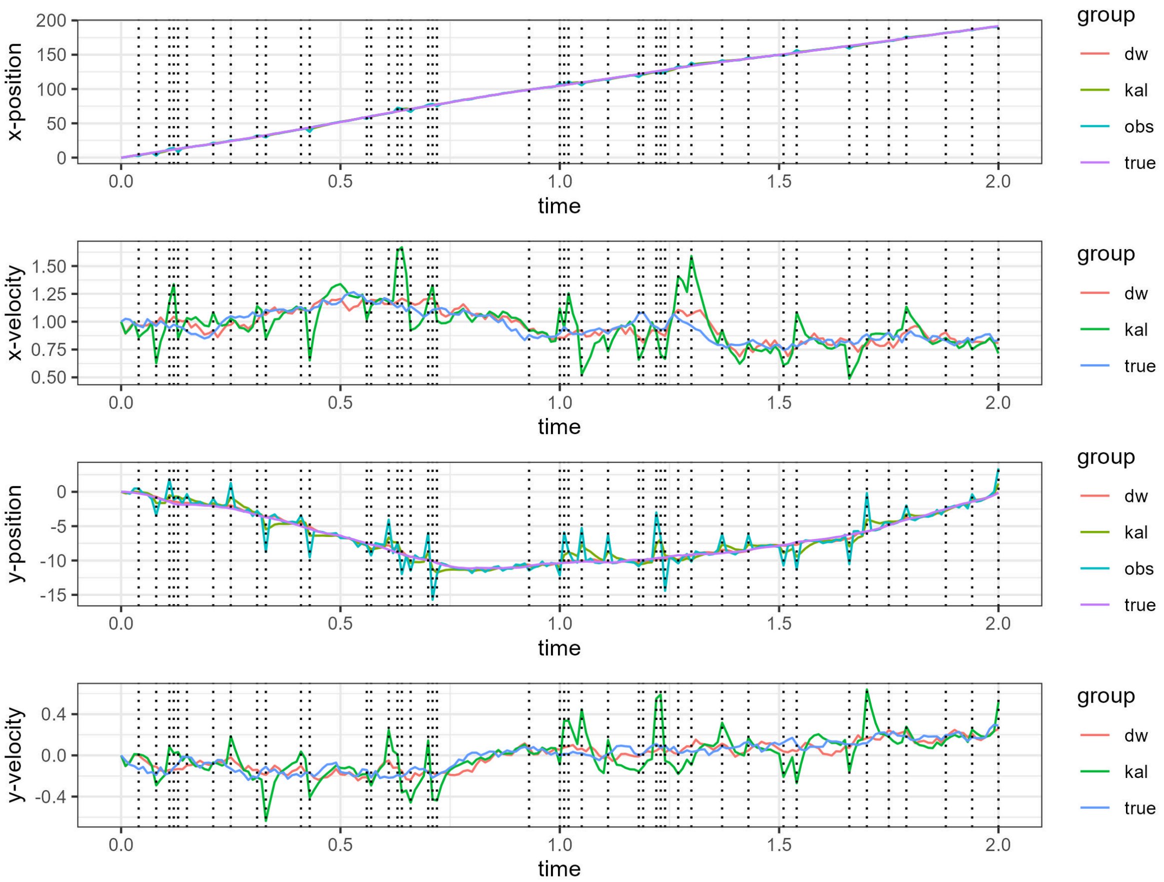


Figure 2.5.: Side-by-side graph comparison of the simulated signal (true), generated observations (obs), the KF analysis mean (kal) and the tuned $\mathcal{D}_w$-KF analysis mean (dw). Dotted lines signal instances of observation contamination.

factor $25\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}$ and $\lambda = 0.25$.
As we wanted to investigate the capabilities of the $\mathcal{D}_w$-EnKF also regarding small ensemble sizes, we picked $M = 10$. An initial investigation of the $\mathcal{D}_w$-EnKF to inspect general performance given the defautl learning rate with no contamination was done over a window $[0, 40]$. The actual simulation with contaminated observations was done over a window $[0, 50]$.

The main results shown in 2.6 agree with especially the unobserved dimension benefiting from the robustness of the diffusion score matching approach. The results look very promising for further investigations on ensemble implementation of the approach also for unfeasible or very challenging forecast covariance propagation. A comparison on uncontaminated observations as well as the individual graphs for each dimension can be found in figure (A.7) to (A.13) in the appendix.

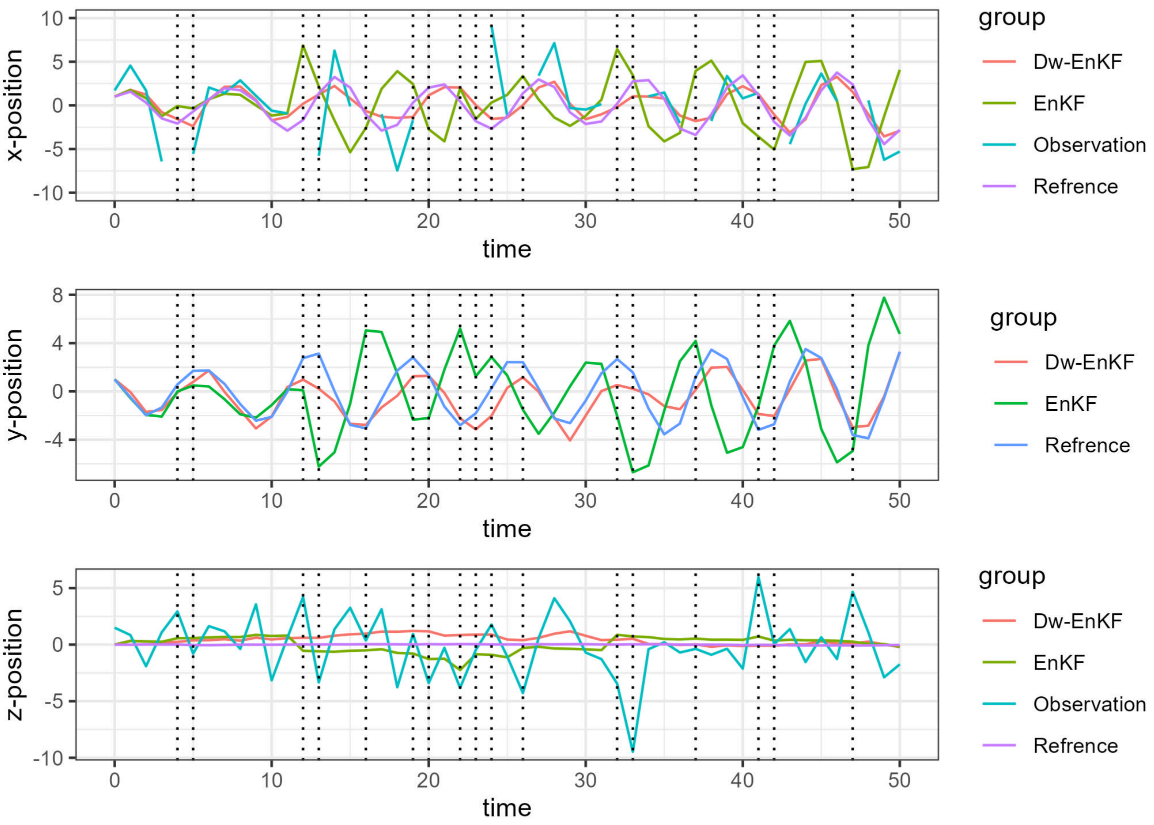


Figure 2.6.: Side-by-side graph comparison of the simulated signal (Refrence), generated observations (Observation), the EnKF analysis mean (EnKF) and the $\mathcal{D}_w$-EnKF analysis mean (Dw-EnKF). Dotted lines signal instances of observation contamination.

2.4.4. Experiment D: The Lorenz-63 Test

Finally, we want to test the performance of the sketched $\mathcal{D}_w$-EnKF with perturbed observations for the chaotic Lorenz-63 model. We hereby take the simulation mainly from (Reich and Cotter, 2015). Recall, let z be the signal variable, we then have the vector field f given by

$$f(z) := \begin{pmatrix} \sigma(z_2 - z_1) \\ z_1(\rho - z_3) - z_2 \\ z_1 z_2 - \beta z_3 \end{pmatrix} \tag{2.76}$$

with parameters $\sigma = 10$, $\rho = 28$ and $\beta = \frac{8}{3}$. We chose a step-size of $\delta t = 0.001$ and the initial value $z_0 = (-0.587, -0563, 16.870)^T$. To implement a forward Euler scheme as numerical approximation, we include a non-autonomous forcing term g_n that essentially comes down to a tent map iteration. We set $a = (\delta t)^{-\frac{1}{2}}$ and the initial forcing term as $g_0 = (a(2^{-\frac{1}{2}} - \frac{1}{2}), a(3^{-\frac{1}{2}} - \frac{1}{2}), a(5^{-\frac{1}{2}} - \frac{1}{2})$ with the entry-wise recursive definition

$$g_{n+1,i} = \begin{cases} 1.99999 g_{n,i} + \frac{a}{2} & \text{if } g_{n,i} < 0 \\ -1.99999 g_{n,i} + \frac{a}{2} & \text{otherwise.} \end{cases} \tag{2.77}$$

The signal is then propagated via

$$z_{n+1} = z_n + \delta t(f(z_n) + g_n) \tag{2.78}$$

over a window $[0, 5]$ for investigating performance for the default choice of tuning parameter and $[0, 10]$ for the experiment. Observations were generated at $t_{out} = 100$ via

$$Y_n = H_n z_n + \sqrt{2} V_n^{\lambda} \tag{2.79}$$

with $H_n = \left(\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \end{smallmatrix}\right)$ and V_n^{λ} contaminated standard Gaussian noise as before with variance factor 10 and $\lambda = 0.1$. We chose an ensemble size of $M = 5$ to mimic application for highly expansive forward models. While a default choice of $\beta = 1$ generally results in good performance, it can still be improved as shown in the first experiment. As the EnKF is no longer optimal or that reliable for the non-linear system, the tuning heuristic via matching KL-divergence no longer holds. The default learning rate managed to recover results of the regular EnKF and no contamination (see figure A.14 to figure A.17 in the appendix for graphs of the uncontaminated simulation).

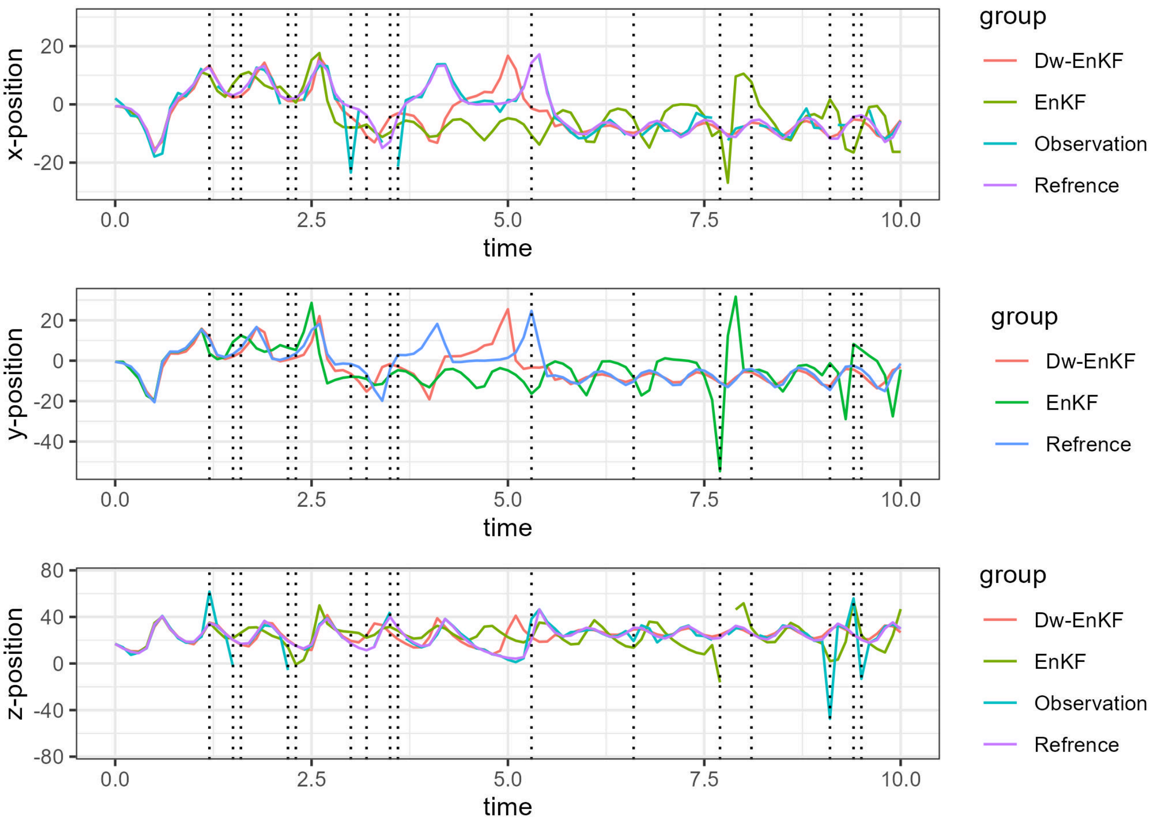


Figure 2.7.: Side-by-side graph comparison of the simulated signal (Refrence), generated observations (Observation), the EnKF analysis mean (EnKF) and the $\mathcal{D}_w$-EnKF analysis mean (Dw-EnKF). Dotted lines signal instances of observation contamination.

The result of a single simulation can only reach so far. Accordingly conclusions should be drawn with great care. Still, from the presented trajectory it seems that while both filters eventually get off track, the $\mathcal{D}_w$-EnKF with perturbed observations manages to noticeably recover contrary to the regular EnKF. The contaminated observations hereby seem to throw the EnKF off-track much earlier compared to the simulation with no contamination. Similarly, the $\mathcal{D}_w$-EnKF is thrown off in the interval $[4, 5]$, yet, manages to catch on again. The hypothesis that the robustness helps in dealing with non-linearity still stands and needs further investigation.

Additional graphs for uncontaminated observations as well as individual dimensions can be found in figure (A.14) to (A.20) in the appendix.

2.5. Discussion and Conclusion

Let us take a step back at this point. The aim of this part was to adapt the popular Kalman filter to achieve robustness regarding observation noise mis-specification regarding heavy tails and measurement outliers. The overarching goal is still in reducing impact of mis-specification, including instances of signal noise mis-specification showing in jumps and rapid change of the signal.

To achieve the desired robustness in observation outliers, we introduced the concept of generalized posteriors and more specific diffusion score matching posteriors given the usual Kalman setting. We proved robustness of the diffusion score matching posterior for Gaussian prior and likelihood via the framework of $\varepsilon-$contaminated observation distribution and a bound on the double supremum over the posterior influence function under mild constraints. In the process we chose an adapted IMQ-kernel utilizing Mahalanobis distance and the positive root of the observation noise covariance for the diffusion matrix and managed to provide a recursive formula for closed form updating of the posterior, so a form of conjugacy, alike to the regular Kalman filter. We discussed the ideas and workings of the adapted and novel terms as well as addressed potential in tuning the learning rate parameter β. We extended on the results by addressing the open problem of long-term stability of the posterior and outlined an ensemble based implementation of the recursive formula, however, with further investigation needed. The intuitions and theoretical results were supported via experiments A and B with experiments C and D providing first ideas about the chances and challenges regarding the particle approach and non-linearity.

It is reasonable to conclude that the approach of the $\mathcal{D}_w$-Kalman filter is a success in its main idea of targeting observation noise mis-specification and thus opens up a variety of directions to go from here - the main direction of this work is in next targeting signal noise mis-specification to combine both later. We are still aware of the initial motivation by the philosophical approach in (Gelman and Shalizi, 2013) on the challenges in Bayesian statistics and modelling. Further, also taking the perspectives in (Morzfeld and Reich, 2018) into consideration, having to assume a model to be well-specified is a strong constraint yet it can often only be addressed to limited extend also for reasons of practicability. The presented results contribute towards reducing the need of this assumption while maintaining this practicability. For further research apart from detecting instance of signal noise mis-specification, we propose investigations on particle approaches similar to what was sketched here, also recalling first results in (Boustati et al., 2020). Further, the theoretical results should also hold for truncated Gaussian distributions on bound subsets of real vector spaces, however, likely need to adapting some details. Next to the idea of filtering, an adaptation along the lines of the popular Rauch-Tung-Striebel smoother seems rigth around the corner in investigating joining backwards propagation and the previous posteriors of the dynamical system. Finally, the long term stability is of major concern. While briefly discussed it still needs more sophisticated evaluation. While the approach presented here produced good results, there are several aspects that can be changed and adjusted with an emphasis on the design of diffusion matrix $w(y)$ in combination with tuning the learning rate β.

This leads to the limits of this work. With the explicit dependency on the observations in the Kalman gain and thus the analysis/posterior covariance, we lose the ability to pre-compute and thus need to put a stronger focus on numerical complexity. The main bottleneck hereby then lies in the inversion required for the Kalman gain, however, this is similar for the regular Kalman filter. Apart from that, all required operations are straight forward with linear complexity of the observation dimension $\mathcal{Y}$ with some even reducing to simple matrix operations depending on the implementation. This agrees with the results in (Altamirano et al., 2023b) claiming the diffusion score matching approach as scalable and cheap compared to previous robust posteriors.

This work is not exhaustive in its comparison to other approaches for robust Kalman filtering such as in (Zhu et al., 2002), (Agamennoni et al., 2011), (Chang, 2014), (Wang et al., 2018) and (Li et al., 2020). Approaches hereby cover Huber's M-estimators, concepts from robust regression and explicitly modelling heavy tailed noise via t-distributions among others. However, due to the ambiguity of the term *robustness* in literature comparison is not as straight forward. To give an example, (Wang et al., 2018) proposes an adaptation to the Kalman filter regarding outlier robustness. While their problem may be similar to mis-specification of observation noise in practice, they aim for a *detect-and-reject* approach explicitly not incorporating outliers in their filtering while the proposed approach of this work aims to still get as many information as possible from them. Either approach might be desired for some applications and lacking in others depending on context. The majority of approaches produce a recursive formula and usually confirm their robustness via simulation results, but rarely in a robustness framework such as contaminated observation distributions.

To close this first part, the research area seems lacking a unified understanding of its terms and problems as well as an overview of established results. While topics like change point detection in the next chapter are also scattered across research fields, they frequently provide summaries and overviews for shared understandings of the similar problem varieties. Accordingly, this work contributes to the body of literature on robust Kalman filtering, but is yet to be placed within. An overview of methods addressing a similar problem formulation as we did is given in (Das et al., 2021) experimentally comparing them for an application in wheel odometry. In their experiments, they compared a Huber Kalman filter utilizing M-estimators as MLEs for robust loss functions, a covariance scaling Kalman filter also employing Mahalanobis distance and variational filters such as the previously mentioned one, however, their experimental results show no clear edge in RMSE of one over the other. This puts emphasis on tuning giving a slight edge to the covariance scaling approach as it only takes a confidence level while all other methods require several tuning parameters with some needing empirical tuning.

The mentioned covariance scaling Kalman filter introduced in (Chang, 2014) is similar to the presented approach in that it also introduces Mahalanobis distance to address modelling errors. Still, there are several significant differences especially in derivation and implementation. Chang utilizes the χ^2-distribution to test for normality and only after rejecting the null-hypothesis of normality a scaling factor is employed to adjust the observation noise covariance via Newton's method until the testing criterion can no longer be rejected with the adjusted covariance. The resulting algorithm is then evaluated empirically. This is fundamentally different from the presented approach via generalised Bayesian posteriors, employing diffusion score matching with Mahalanobis distance utilized in the IMQ-kernel and the observation noise covariance scaling as a result of this fundamental change of the inner workings. Accordingly, the presented approach here is a bottom-up derivation addressing the inherent source of outlier sensitivity resulting from the Kullback-Leibler divergence while the approach in (Chang, 2014) being a top down adjustment to the symptoms.

The diffusion score matching Kalman filter derived in this first part is a rigorous and novel result with a high potential for further investigation.

3. Addressing Signal Noise Mis-Specification: Change Point Inference and Inference Under Change

Constructing an Approach

Change point detection is a multifaceted, divers and widely spread research topic across disciplines. The various notions of change are hereby as different as what may change in a statistical model and reach far beyond the scope of this work. However, the general idea is similar all across - breaking with certain in repetitive structure of distribution in statistical modelling via relaxations allowing for some form of change at an unknown point in a data generating process. In practice, this may frequently translate to not assuming identical distribution throughout a whole sample. For the scope of this second part, we want to pick up the intuition put forward in (Adams and MacKay, 2007) in change point detection as «the identification of abrupt changes in the generative parameters of sequential data». This short and on point notion puts emphasizes on the idea of detecting the presence of change and recovering the temporal location of an instance of change in at least one aspect of the data generating process from an observations sequence. This adequately suits the introduced Kalman setting with the signal as the single most central aspect of the data generating process in that it directly translates into the observation mean for the lienar Guassian setting, and thus, into the resulting challenge of detecting sudden jumps in the signal via detecting change in mean from sequential observations. As initially introduced, we want to interpret these jumps to be results of mis-specification in heavy tailedness of the signal noise, so realization of theses heavy tails not accounted for by Gaussian signal noise. In the scope of this chapter we will introduce them as additional driving terms $\delta_{\eta,k}u_k$ in the signal process at unknown, yet not necessarily random times η, the temporal change point locations, with $\delta_{\eta,n}$ providing some form of emergence criteria such as $\delta_{\eta,n} = \mathbf{1}\{\eta = n\}$ or $\delta_{\eta,n} = \mathbf{1}\{\eta \geq n\}$ in what will be introduced as the *minimax* detection problem, or $\delta_{\eta,k} = \pi_n$ with π_n non-negative, discrete random variable, e.g. $\pi_n \sim_{iid} \mathrm{Ber}(\varphi_n)$, in the Bayesian setting, similar to the ε-contamination in the previous chapter regarding observation noise mis-specification. To put in different terms, $\delta_{\eta,k} \in \{0, 1\}$ is an indicator for the additional driving term u_n depending on $\eta \in \mathbb{N}$ indicating instances of change and the regular discrete time index $n \in \mathbb{N}$.

As said, the field of change point detection is vast and divers. Accordingly, we want to hone in on methods and results relevant for the aim of this section. Detecting instances of abrupt jumps, or change, in the signal is hereby a crucial stepping stone towards the central aim in accounting for signal noise mis-specification in inference and forecasting procedures - therefore, similar to the previous section, the desired result will be an adaptation of the popular Kalman filter incorporating potential signal jumps under considerations of uncertainty. In the scope of this chapter we will not consider observation noise mis-specification and only focus on challenges arising with signal noise mis-specification for the regular Kalman setting and corresponding required additional assumptions. To suit the Kalman setting, we are therefore mainly interested in sequential and online approaches to detecting change in mean of stochastic sequences.

This chapter will first provide a brief introduction to the broader setting of sequential change point detection with its most central results providing required tools for considerations further down. The main references hereby are recent reviews and popular monographs as well as single key publications. The popular CUSUM approach related to sequential hypothesis testing and its variations, frequent optimality criteria and adaptations to different assumptions and better computational feasibility will hereby be the main focus. Next, we will introduce the more recent, successful approach of Bayesian online change point detection via run-length posteriors as well as an adaptation in restarted BOCPD and adapt them and their intuition for exploiting results on CUSUM schemes in deriving tuning criteria satisfying notions of asymptotic optimality. The resulting scheme, denoted CUSUM restarted BOCPD (CR-BOCPD) is hereby the central result of this chapter. Afterwards, we want to finish by deriving current state-of-the-art via χ^2 strategies for linear Gaussian systems. Both will be related back to detecting abrupt signal jumps in Kalman filtering and utilized in constructing Gaussian mixture model filters under relaxed assumptions.
Taking everything together in a concrete road-map: We provide valuable results on CUSUM strategies. We modify BOCPD and R-BOCPD to exploit these results on CUSUM strategies. We explore current state-of-the-art strategies for detecting change in linear Gaussian systems. We propose strategies for inference under change in the Kalman setting via Gaussian mixture models of multiple Kalman filters.

3.1. Motivating CUSUM Strategies in Sequential Change Point Detection

As put in (Niu et al., 2016), «there exists a massive number of research papers on change-point detection or closely related topics». The field enjoys relevance across research areas, e.g. in engineering, climatology, the bio-sciences and linguistics among others, with frequent terminology including anomaly, event, outlier and irregularity detection (Namoano et al., 2019). The sequential change point detection approaches we are interested in are closely related to the broader problem of sequential probability ratio testing (SPRT) of hypotheses and therefore have their roots in pioneering work in (Wald, 1947) and (Girshick and Rubin, 1952) on sequential decision making. Our focus will be with the celebrated and widely used CUSUM approach introduced in Page's line of work in (E. S. Page, 1954) and (E. Page, 1955) which was later related to open-end SPRTs in milestone works such as (Lorden, 1971). As it is a crucial contribution to the field of sequential change point detection, we want to adapt the notation in (Lai, 1998) of the cumulative sum (CUSUM) procedure for non-iid random variables via conditional density functions. Moreover, its also the results in (Lai, 1998) providing the most relevant results of this section and the ideas of this chapter.

Let $Y_{1:(\eta-1)}$ be random vectors with a common density function f_0 and let $Y_{\eta:n}$ be RVs with common density function f_1. $\mathbb{P}_\eta$ and $\mathbb{E}_\eta$ denote the probability measure and expectation for change at $1 \leq \eta \leq n$ and $\mathbb{P}_0$ and $\mathbb{E}_0$ denote the probability measure and expectation for no change, so $\eta = \infty$. We are explicitly interested in non-independent random variables. Let $f_0(\cdot|Y_{1:(n-1)})$ be the conditional density function of Y_n under $\mathbb{P}_0$ for $n \geq 1$ with the conditional density function under $\mathbb{P}_\eta$ given by $f_0(\cdot|Y_{1:(n-1)})$ for $n < \eta$ and $f_1(\cdot|Y_{1:(n-1)})$ for $n \geq \eta$. As a key quantity, introduce the log-likelihood ratio (LLR) statistic

$$S_n = \log[\frac{f_1(Y_n|Y_{1:(n-1)})}{f_0(Y_n|Y_{1:(n-1)})}] = \log[f_1(Y_n|Y_{1:(n-1)})] - \log[f_0(Y_n|Y_{1:(n-1)})]. \tag{3.1}$$

Note hereby the role of Kullback-Leibler divergence as expectation of the LLR. The generalized CUSUM rule is a stopping time

$$\begin{aligned} N &= \inf\{n \geq 1 : \max_{1\leq k\leq n} \sum_{s=k}^{n} S_s \geq c\} \\ &= \inf\{n \geq 1 : \max_{1\leq k\leq n} \sum_{s=k}^{n} \log[\frac{f_1(Y_s|Y_{1:(s-1)})}{f_0(Y_s|Y_{1:(s-1)})}] \geq c\} \end{aligned} \tag{3.2}$$

for a threshold c chosen such that it achieves desired criteria. Define $\inf\{\} = \infty$, so that for the empty set we maintain the assumption of no change. Further, for $n \geq \eta$ we assume that $\frac{1}{n-\eta+1} \sum_{s=\eta}^{n} S_s$ converges in probability under $\mathbb{P}_\eta$ to some constant I. In the iid-case, so assuming independence of the corresponding distributions, this constant is given by the Kullback-Leibler divergence of the pre- and post change density functions. Another frequent sequential change point detection procedure, however not focus of this work, is the Shiryaev-Roberts-Pollak procedure also based on the LLR and with similar properties.

3.1.1. Considering Notions of Optimality

The popularity of the CUSUM procedure is rooted in its simplicity, especially for independent observations, as well as its performance regarding optimality criteria popular in some research communities . In (Xie et al., 2021), authors provided a well written summary which will be the basis for a brief introduction to these notions of optimality here. In sequential change point detection procedures are subject to a trade-off. Overly sensitive methods quickly fall victim to a high false alarm rate in that they flag instances of change when there is non present. On the other hand, more conservative methods may be subject to long delay times between emergence and detection of change. In essence, both quantities, probability of false alarm (PFA) and detection delay resemble the usual type-I and type-II errors in statistics in that we work with the null hypothesis of no change versus multiple alternative hypothesis of change at respective instances. Accordingly, the PFA, so the type-I error of supposedly detecting change when there is non, is what we want to control via a fixed bound with detection delay, akin to the type-II error of not detecting a present change, minimised or equivalently maximised in power for a given sample size. The PFA constraint generally translates into the detection threshold value of a procedure such as the value c in (3.2). Therefore, the choice of a threshold value directly influences the detection delay in the sketched trade off. The central challenge is in finding procedures balancing both regarding some notion of detection delay via finding threshold values minimising this detection delay for a given, fixed probability of false alarm. Hereby two settings are distinguished: The setting of the minimax or non-Bayesian change point detection problem assumes the change point $\eta \geq 1$ to be deterministic and unknown. The Bayesian change point detection setting on the other hand assumes a change point to be a realisation of an integer valued random variable with non-negative support.

Taking the Frequentist Perspective

Minimax optimality faces the problem that both probability of false alarm and optimal detection delay are more challenging to precisely define. This is to an even greater extent the case when considering the stretched body of literature with different notions present in different research communities. Following (A. G. Tartakovsky, 2009), we want to start with a basic notion of loss measured via probability of false alarm (PFA) $\mathbb{P}_\eta[N < \eta]$ or $\mathbb{E}_\eta[N \cdot \mathbf{1}\{N < \eta\}]$, the expected time to false alarm. We observe that controlling these quantities for all $\eta \geq 1$ is equivalent to controlling the corresponding quantities under the null hypothesis in $\mathbb{P}_0[N < \infty]$ and $\mathbb{E}_0[N]$. A central insight is then that $\mathbb{P}_0[N < \infty] \leq \alpha_0^*$ for $0 < \alpha_0^* < 1$ leads to $\mathbb{E}_0[N] = \infty$ and vice versa $\mathbb{E}_0[N] < N^*$ for $N^* < \infty$ leads to $\mathbb{P}_0[N < \infty] = 1$. Either offers their own approaches and challenges and has arguments for and against. For the literature and line of work we want focus on, the latter in choice of a constraint in finite expected time to false alarm, so $\mathbb{P}_0[N < \infty] = 1$, is considered. It indicates that the stopping rule, such as (3.2), will necessarily activate in finite time. A central reason hereby lies in choosing constant threshold values of detection procedures to obtain theoretical results in notions of asymptotic optimality of interest. Addressing this issue (Borovkov, 1999) and (Alami et al., 2020), among others, investigated curvilinear thresholds achieving similar notions of asymptotic optimality under additional strict assumptions while maintaining $\mathbb{P}_0[N < \infty] \leq \alpha_0^*$ for $0 < \alpha_0^* < 1$.
Back to (Xie et al., 2021), a frequent quantity given the limitations is the average

run length $ARL(N) = \mathbb{E}_0[N]$. Its reciprocal is then what is generally referred to as false alarm rate $\mathrm{FAR}(N) = \frac{1}{ARL(N)} = \frac{1}{\mathbb{E}_0[N]}$ with the corresponding interpretation via rate of false alarms occurring for repeated procedures. We want to think of it as a substitute for the PFA given from a choice of constant threshold c. While we know that the process will necessarily terminate with probability 1, we still want to impose some sort of control on the frequency of such an event this way. Analogue to the hypothesis testing framework, we want to find a minimax most powerful change detection procedure minimizing measures of detection delay over all procedures T satisfying $\mathrm{FAR}(T) \leq \alpha_0$ with $\alpha_0 \in (0, 1]$. Over time, two central measures of detection delay in a minimax sense have been established. The first was introduced by Lorden in (Lorden, 1971) relating CUSUM procedures to open-end SPRTs. His idea was in a notion of worst average detection delay via the supremum of the average detection delay conditioned on the worst possible realizations:

$$\mathrm{WADD}(T) = \sup_{n \geq 1} \operatorname{ess\,sup} \mathbb{E}_n[(T-n)^+ | Y_{1:(n-1)}]. \tag{3.3}$$

Deemed an overly pessimistic measure of detection delay in considering the worst possible pre-change sample, the second measure was brought forth by Pollak in (Pollak, 1985):

$$\mathrm{CADD}(T) = \sup_{n \geq 1} \mathbb{E}_n[T - n | T \geq n]. \tag{3.4}$$

For any procedure T we observe $\mathrm{WADD}(T) \geq \mathrm{CADD}(T)$. As the exact evaluation of both measures is highly challenging, mainly first-order asymptotically optimal solutions were investigated. To specify, first order asymptotic optimality hereby refers to the fraction of optimal and achieved detection delay approaching 1 as $\alpha_0 \to 0$, e.g. $\frac{\mathrm{CADD}(N)}{\inf_T \mathrm{CADD}(T)} \to 1$ as $\alpha_0 \to 0$. In his milestone work (Lai, 1998), Lai proved the asymptotic lower bound for the CADD, and thus also for the WADD, in the general non-iid setting introduced above via

$$\mathrm{WADD}(T) \geq \mathrm{CADD}(T) \geq \frac{|\log(\alpha_0)|}{I}(1 + o(1)) \tag{3.5}$$

with positive constant I from the above convergence constraint and reducing to $I = \mathcal{D}_{KL}(f_0 || f_1)$ in the iid-case, for all procedures T with $\mathrm{FAR}(T) \leq \alpha_0$ as $\alpha_0 \to 0$. Furthermore, Lai proved that the CUSUM procedure (3.2) attains this asymptotic lower bound for appropriately chosen threshold $c \sim |\log(\alpha_0)|$ and is first-order asymptotic minimax optimal in both the sense of Lorden's WADD and Pollak's CADD (Lai, 1998).

As an important remark, the introduced notions of minimax optimality here are just two among many introduced in literature under a large variety of constraints and assumptions. It grew from a classical line of work, however, has not seen much development in recent years. We mainly want to focus on this specific notion of first-order asymptotic optimality under constraints on the rate of false alarms as it has previous investigation and results regarding change in linear Gaussian state space models. Other notions of minimax optimality for this context need exploring with additional concepts for shared optimality regarding inference combining filtering and change point detection.

Taking the Bayesian Perspective

The notion of Bayesian optimality is fairly more direct in comparison, as location of change can be quantified via their random character. As given in (Xie et al., 2021), let η be an integer valued random variable with non-negative support and probability mass function $\pi_n = \mathbb{P}[\eta = n]$. To give an example, a frequent assumption is then for η to follow a geometric distribution via $\pi_n = \mathbb{P}[\eta = n] = \varphi(1-\varphi)^{n-1}\mathbf{1}\{n \geq 1\}$ with $0 < \varphi < 1$ and $\pi_0 = 0$. Note that this intuition is more sophisticated than the idea sketched in the introduction, yet resembles the same idea in practice. Given a detection procedure T, the average detection delay (ADD) and the probability of false alarm (PFA) are then given via

$$\begin{aligned} ADD(T) &= \mathbb{E}[(T-\eta)^+] = \sum_{n=0}^{\infty} \pi_n \mathbb{E}_n[(T-\eta)^+] \\ \mathrm{PFA}(T) &= \mathbb{P}[T < \eta] = \sum_{n=0}^{\infty} \pi_n \mathbb{P}_n[T < \eta]. \end{aligned} \tag{3.6}$$

Again, we want to focus on all detection procedures with $\mathrm{PFA}(T) \leq \alpha_0^*$ for fixed $\alpha_0^* \in (0,1)$. In (Lai, 1998), Lai proved an asymptotic lower bound for the ADD given the constraint on the PFA and proved first-order asymptotic optimality of the CUSUM procedure in (3.2) for adequately chosen threshold c and certain additional conditions.

Against this Background we want to critically emphasize the concept of this probability mass function π_n. While it is as valuable and highly useful tool for theoretical analysis, there is usually no reliable access to it in practice. It frequently needs assuming in corresponding strategies for application, however, we then necessarily face the issue that we are very likely going to be wrong about this random quantity loosing theoretical results and guarantees. Accordingly, we want to lean more towards the minimax setting as introduced here to not rely on such a crucial assumption.

3.1.2. Versatility of the CUSUM Rule

A key takeaway lies in the incredible power and versatility of the CUSUM procedure. It achieves first-order asymptotic optimality in all three popular senses for corresponding choices of threshold value c. The main approach hereby is to utilizes Doob's sub-martingale inequality (see (Basseville, Nikiforov, et al., 1993) for additional details) and choice of $c \sim |\log(\alpha_0)|$.

Given all desirable properties, where are the limitations, there are two major drawbacks which are both also addressed in (Lai, 1998). The first lies in computational cost. For the iid-case, the log likelihood ration statistic S_n reduces to a simple recursion in $S_n = (S_{n-1} + \log(\frac{f_1(Y_n)}{f_0(Y_n)}))^+$ with detection rule $N = \inf\{n \geq 1 : S_n \geq c\}$. However, for the generalized CUSUM rule in (3.2) no such recursion is available. Instead, it requires maximization in index $k \in \{1, 2, \ldots, n\}$ at time n and for every time step leading to linear increasing computational complexity in the observations. The popular workaround initially introduced in (A. S. Willsky and Jones, 1974) utilizes a window-limited adaptation only tracking the last m many time points as potential candidates for change, so $k \in \{m - n, m - n + 1, \ldots, n\}$.

A second limitation lies in that the CUSUM procedure as given in (3.2) requires explicit knowledge of the post change density f_1. To tackle this wide reaching challenge in application, there are two popular adaptations of the CUSUM procedure. Following (Lai, 1998), instead of assuming an explicit post-change density f_1, we instead want to assume the post-change density to be a member of some parametric family f_θ with $\theta \in \Theta$. Accordingly, we also want to adapt the probability measures to $\mathbb{P}_{\eta,\theta}$ and the corresponding expectation to $\mathbb{E}_{\eta,\theta}$. Addressing the resulting estimation problem, we either want to assume some distribution of the unknown parameter or need estimating it with every step. For the first approach, let G be a probability distribution on Θ. We then obtain the mixture likelihood ratio statistics

$$\tilde{S}_{k:n} = \frac{\int_\Theta \prod_{t=k}^n f_\theta(Y_t|Y_{1:(t-1)}) \mathrm{d}G(\theta)}{\prod_{t=k}^n f_0(Y_t|Y_{1:(t-1)})} \tag{3.7}$$

and accordingly the weighted CUSUM procedure

$$\tilde{N} = \inf\{n \geq 1 : \max_{1 \leq k \leq n} \tilde{S}_{k:n} \geq \exp(c)\}. \tag{3.8}$$

If no such probability distribution on the space of the unknown post-change density parameter is available or reasonable, the second approach wants to employ maximum likelihood estimation. Define the parameter dependant likelihood ratio statistics

$$S_n(\theta) = \log(\frac{f_\theta(Y_n|Y_{1:(n-1)})}{f_0(Y_n|Y_{1:(n-1)})}). \tag{3.9}$$

Combining both estimation of $\theta \in \Theta$ and testing then results in the generalized likelihood ratio (GLR) CUSUM procedure

$$\hat{N} = \inf\{n \geq 1 : \max_{1 \leq k \leq n} \sup_{\theta \in \Theta} \sum_{s=k}^n S_s(\theta) \geq c\}. \tag{3.10}$$

Both adaptations maintain the first-order asymptotic optimality via attaining the asymptotic lower bound in detection delay for $c \sim |\log(\alpha_0)|$ and $\alpha_0 \to 0$ as proven in (Lai, 1998) under additional assumptions and constraints.

Making CUSUM Computationally Feasible

Going back to the issue of numerical complexity, the CUSUM procedure (3.2), the weighted CUSUM procedure (3.8) and most of all the GLR CUSUM procedure (3.10) are affected by this issue. The window-limited scheme was developed for the GLR CUSUM approach regarding detection of abrupt jumps in linear dynamical systems, such as we are interested in, as the corresponding community has an additional emphasize on computational feasibility. Introduced in (A. S. Willsky and Jones, 1974) and (A. Willsky and Jones, 1976) to keep complexity to a predetermined finite limit, it was initially lacking statistical theory. As put forward and proven in (Lai, 1998), when choosing the window size $m = m(\alpha_0)$ carefully as a function of the rate of false alarm α_0 additionally satisfying strict constraints, the window-limited CUSUM strategies maintain desired properties in their first-order asymptotic optimality. Carefully chosen hereby requires for $m(\alpha_0)$ that

$$\begin{aligned}
&\frac{m(\alpha_0)}{|\log(\alpha_0)|} \to \infty \\
&\log(m(\alpha_0)) = o(\log(\alpha_0)) \iff \frac{\log(m(\alpha_0))}{|\log(\alpha_0)|} \to 0 \\
&\text{as } \alpha_0 \to 0.
\end{aligned} \tag{3.11}$$

One such choice of $m(\alpha)$ meeting the criteria is therefore given by a poly-logarithmic function $m(\alpha_0) = |\log(\alpha_0)|^l$ with $l > 1$.

To summarize, the CUSUM procedures attain the asymptotic lower bound under all three established detection delay notions with respective constraints. Moreover, they do so also for their limited size window variants. Let $c \sim |\log(\alpha_0)|$ and $m = m(\alpha_0)$ such that (3.11) holds. We can then adapt the previous procedures such that $\max\limits_{1\leq k\leq n}$ is replaced by $\max\limits_{(n-m)\leq k\leq n}$ reducing to the $m(\alpha_0)$ most recent hypotheses of change, e.g

$$N = \inf\{n \geq 1 : \max_{(n-m)\leq k\leq n} \sum_{s=k}^{n} S_s \geq c\} \tag{3.12}$$

for (3.2), resulting in powerful tools for sequential change point detection - this is, assuming we can match their requirements on density functions.

3.2. Grasping Bayesian Online Change Point Detection

Effective online implementation of change point detection procedures is of major concern in practice. As given in (Xie et al., 2021), the introduced sequential procedures allow for that to only some degree. Taking the GLR CUSUM procedure for example, even for the window-limited adaptations, the parameter estimation with incorporating novel observations poses a major challenge for effective implementation.

Addressing this hurdle, approaches from sequential learning were investigated. Among them, two highly cited works exploited forward message passing structure and Bayesian computation resulting in the popular Bayesian online change point detection (BOCPD) scheme. The approach was developed independently in (Adams and MacKay, 2007) and (Fearnhead and Liu, 2007) with both approaches foundation for several follow up publications. Among them are (Altamirano et al., 2023b) and (Alami et al., 2020) as key results to the work at hand. For this section we will focus on the formulation of Bayesian online change point detection as introduced in (Adams and MacKay, 2007) and sequential notation as well as simplifications similar to analysis in (Alami et al., 2020).

Core Ideas in BOCPD

The central intuition of BOCPD lies in evaluating different scenarios about supposed instances of change via run-length posteriors counting time steps since a last presumed instance of change. Using Bayes theorem, a prior probability of change is combined with a parametric model for a conditional observation likelihood of specific instance of supposed change to obtain the desired posterior on time-steps since the last change. The utilized structure hereby is very similar to a forward message-passing algorithm for hidden Markov models. Alongside, conjugacy properties of exponential family distributions are exploited as observation likelihoods for easy and implicit online estimation of post-change parameters. In the original approach in (Adams and MacKay, 2007), assumed change points at k are transformed to run-lengths r_n of a regime since that last assumed change up to current time n via $r_n = n - k$.

For a latent parameter $\theta \in \Theta$ subject to change, assume the conjugated likelihood-prior pair $\mathbb{P}[y_n|\theta]$ and $\mathbb{P}[\theta]$ with posterior $\mathbb{P}[\theta|y_{1:(n-1)}]$ and cheap access to the corresponding marginal distribution $\mathbb{P}[y_n|y_{1:(n-1)}]$, also called model predictive in the BOCPD context. Accordingly, the pre- and post-change distributions are assumed to be recovered for some $\theta_0, \theta_1 \in \Theta$. Bayes theorem is utilized to obtain the discrete posterior distribution of run-lengths r_n, so a notion of plausibility about a change point r_n time steps ago, via

$$\mathbb{P}[r_n|y_{1:n}] = \frac{\mathbb{P}[r_n, y_{1:n}]}{\mathbb{P}[y_{1:n}]}. \tag{3.13}$$

Hereby, we can decompose the joint distribution of run-lengths and available observations in

$$\begin{aligned}\mathbb{P}[r_n, y_{1:n}] &= \sum_{r_{n-1}} \mathbb{P}[r_n, r_{n-1}, y_{1:n}] \\ &= \sum_{r_{n-1}} \mathbb{P}[r_n, y_n|r_{n-1}, y_{1:(n-1)}]\mathbb{P}[r_{n-1}, y_{1:(n-1)}] \\ &= \sum_{r_{n-1}} \mathbb{P}[r_n|r_{n-1}]\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}]\mathbb{P}[r_{n-1}, y_{1:(n-1)}].\end{aligned} \tag{3.14}$$

In (Adams and MacKay, 2007), the conditional observation likelihood is reduced to $\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}] = \mathbb{P}[y_n|r_{n-1}, y_{(n-r_n):(n-1)}]$, so only taking recent observations $y_{(n-r_n):(n-1)}$, thus combining both conditions. The intuition lies in that the implicit Bayesian distribution of the post-change latent parameter θ_1 in $\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}]$ via the underlying conjugacy will be most accurate when taking the most amount of observations available from the post-change distribution with the least amount of observations from the pre-change distribution. So for the ideal case of $r_n = n - \eta$ for an actual instance of change at η, the observation y_n is most likely for $\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}]$ conditioned only on the observations $y_{\eta:(n-1)}$. For $n = \eta$ this reduces to the marginal distribution over the prior.

The implementation is a recursive update taking the joint distribution at the previous time step $\mathbb{P}[r_{n-1}, y_{1:(n-1)}]$, the conditional likelihood of observing y_n given the observations and the corresponding run-length $\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}]$ and the change point prior $\mathbb{P}[r_n|r_{n-1}]$. The main challenges with BOCPD are in formulating the conditional observation likelihood and in justifying a suitable change point prior. Both contribute to making the approach fast, however, we deem the latter to impose a strong assumption. Accordingly, it is also this assumption we want to address with modification.

The understanding as the model predictive of the marginal or conditional observation likelihood also arises with its use in predicting a novel observation y_{n+1} via marginalizing over all current run-lengths in

$$\mathbb{P}[y_{n+1}|y_{1:n}] = \sum_{r_n} \mathbb{P}[y_{n+1}|r_n, y_{1:n}]\mathbb{P}[r_n|y_{1:n}]. \tag{3.15}$$

Moreover, it is exactly this weighted sum we are interested in for adjusting the Kalman filter regarding abrupt jumps in the signal. It is part of what makes BOCPD attractive for our purpose of estimation and forecasting under potential signal noise mis-specification next to similarity in structure via conjugacy.

As a short remake on notation, we are currently using $\mathbb{P}$, staying with the original notation in (Adams and MacKay, 2007). It hereby refers respectively to the continuous density function on y_n, the discrete probability measure on r_n or the mixture for the joint measure over both. Later on we will adapt notation to align with the previous chapter.

As stated in (Adams and MacKay, 2007), the conditional change point prior is emphasized as a main contributor for computational efficiency of the approach. It is essentially a binary random variable providing conditional probabilities of the run-length to either increasing by one, so no change, or directly being set to zero, an instance of change occurring, given the current run-length. More precise, let $H(\tau) = \frac{\phi(\tau)}{1-\Phi(\tau)}$ be some hazard function of the run-length with discrete density function ϕ and cumulative distribution function Φ. Let $\mathbb{P}[r_n = 0|r_{n-1}] = H(r_{n-1} + 1)$, $\mathbb{P}[r_n = r_{n-1} + 1|r_{n-1}] = 1 - H(r_{n-1} + 1)$ and 0 otherwise. Choosing a geometric density function for ϕ with parameter $\tilde{\varphi}$ yields the popular memory-less hazard function $H(\tau) = \frac{1}{\tilde{\varphi}} = \varphi$ - a constant probability for an instance of change independent of time. Note hereby the similarity to the previous notions of probabilistic change. In the message-passing algorithm, the probability of each positive run-length is only in

extending a previous run-length while the probability of the run-length dropping to zero, so change at that instance, is aggregated over all run-lengths at the previous time step.
The resulting approach requires an initial condition for the observation likelihood and for the run-length at time zero. When there are no past observations of the sequence available, (Adams and MacKay, 2007) suggests setting the initial run-length to zero. If there are past observations with instances of change available, the run-length prior may be adapted based on the survival function. In the case of the memory-less hazard function it makes no difference.

To summarise, Bayesian online change point detection combines probability of a run-length r_n at time n describing the time steps since a assumed last instance of change at k with an observation likelihood conditioned on past observation and change occurring a given run-length ago, so $\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}]$. Via marginalizing over all run-lengths we obtain posterior probabilities of run-lengths given the observations. In parallel to message-passing, we assume run-length probabilities to distribute along a trellis according to the change point prior, tracking all previous possible instances of change.
Herein also arises a central limitation. Similar to the CUSUM procedure, at time step n, we have to track $n + 1$ different run-lengths via $r_n \in \{0, 1, \ldots, n\}$. The solution suggested in (Altamirano et al., 2023b), is in pruning the observed run-lengths either with a threshold or by only tracking a specified number of the most probable run-lengths. Both may distort the posterior probabilities in the iterative algorithm, however, seem to have little impact in practical application judging by popularity.

A Note on Bayesian Prediction of Change Points

A direct extension alongside the idea of modelling the change point prior via a hazard function is to then investigate a notion of time until the next instance of change. Given the intuition of the run-length as a survival process, we want to evaluate probabilities of when the survival process described by the hazard function might die - the next instance of change given the current run-lengths. In (Agudelo-España et al., 2020), the authors develop this idea calling it Bayesian online prediction of change points. The number of time steps until the next change point is herein denoted as residual time. Let o_n be this residual time expressed via the run-lengths as the event

$$r_{n+1}, r_{n+2}, \ldots, r_{n+o_n} > 0 \text{ and } r_{n+o_n+1} = 0. \tag{3.16}$$

The residual time posterior proposed in (Agudelo-España et al., 2020) is then the probability of this event via accumulating the hazard function given a current run-length, so $\mathbb{P}[o_n|r_n] = H(r_n + o_n) \prod_{\tau=r_n}^{r_n+o_n-1}(1 - H(\tau))$, and marginalized over all current run-length posteriors $\mathbb{P}[r_n|y_{1:n}]$, resulting in

$$\mathbb{P}[o_n|y_{1:n}] = \sum_{r_n} \mathbb{P}[o_n|r_n]\mathbb{P}[r_n|y_{1:n}]. \tag{3.17}$$

Note hereby, the switch from additive term in the index of the event to additive term of the run-length for evaluation of the probability of the event is intended.

In a remark, the authors emphasize the ability to pre-compute $\mathbb{P}[o_n|r_n]$ as it does not depend on the observations given the implicit assumption on independence of observation model regarding residual time . Further, for the memory-less hazard function of a geometric random variable, it is observed that the event in (3.16) reduces to the somewhat trivial case $\mathbb{P}[o_n|r_n] = \varphi(1-\varphi)^{o_n}$ which may only hold limited insight.

In (Agudelo-España et al., 2020) a specific hidden semi-Markov structure is assumed for transitions between regimes of the data generating process. While adapting to the context at hand is possible and potentially interesting, it is not a concern here. Moreover, the same holds for the broad idea of Bayesian online change point prediction. It may be interesting depending on application and can likely be incorporated with little additional effort, yet it requires explicit knowledge or assumptions on the change point prior. As discussed, this is a strong and limiting assumption we want best avoided.

Still, we want to make use of this idea whenever there is access to such run-length prior for a one-step ahead forecast of run-length posteriors, so $o_n = 1$. Similar to the intuition of the forecast step in the Kalman filter, we want to adjust the equation in (3.14) to obtain

$$\mathbb{P}[r_n|y_{1:(n-1)}] \propto \sum_{r_{n-1}} \mathbb{P}[r_n|r_{n-1}]\mathbb{P}[r_{n-1}|y_{1:(n-1)}], \tag{3.18}$$

the run-length forecast for a point in time n based on the available past observations - essentially leaving out the model predictive in (3.14) as observation y_n is not yet available.

Putting BOCPD in Perspective

At this point, there are three major observations about BOCPD we want to point out. As already said, the essence of the approach lies in tracking posteriors of different scenarios resembling instances of change at specific time steps via $\mathbb{P}[y_n|r_{n-1}, y_{1:(n-1)}]$. Taking the notion of scenarios in (Alami et al., 2020) with $r_n = n - k \iff k = n - r_n$, we want to interpret the conditional observation likelihood in the sense of a hypothetical scenario of presumed instances of change having occurred at time $k \in \{1, 2 \ldots, n\}$ with $\mathbb{P}[y_n|y_{1:(n-1)}, k]$ but no difference in computation. In line with that change in interpretation, we want to adapt the idea of run-length posteriors to the concept of scenario weights.

The second observation was also pointed out in (Alami et al., 2020). The approach does not perform change point detection in a classical sense of making an explicit decision about the presence of change. Instead, it provides a notion of plausibility of an assumed change at a time point $k = n - r_n$ given the available observations via the run-length posterior for r_n or scenario weight. The BOCPD approach avoids concrete decision making, however, it introduces the weighted sum useful in estimation and prediction of observations and the latent parameter of the underlying

Bayesian model via (3.15), quantifying uncertainty induced by considering change. Lastly, as stated in the concluding remarks of (Adams and MacKay, 2007), the run-length posteriors are exact in the sense that all performed computations are exact. However, that is only so, assuming the change point prior is a good representation of the actual occurrence of change points and up to computational reductions such as pruning. The approach explicitly requires both for application, yet especially the run-length prior is something we deem difficult to obtain for the applications in mind.

BOCPD was well received and enjoys large popularity regardless. Yet, as stated in (Alami et al., 2020), there was no sophisticated, thorough analysis of BOCPD in some classical sense of sequential change point detection introduced in the beginning of this chapter with measures of false alarm and detection delay. The modification in (Knoblauch and Damoulas, 2018) and (Altamirano et al., 2023b) via addressing volatility to outliers of the conjugacy with robust posteriors addressed a well known issue of a tendency of false alarms of BOCPD, but again results are only evaluated empirically. Detection in the sense of decision making about instances of change is only done implicitly by tracking the maximum a-posteriori (MAP) estimator of run-lengths in (Altamirano et al., 2023b) and avoided in (Adams and MacKay, 2007).

Some of the observations go hand in hand. In order to evaluate the approach with measures of false alarm and detection delay, an explicit decision making rule is required. Further, exactly evaluating such a rule is highly difficult with the trellis like structure passing down part of the scenario weights, the run-lengths posteriors, at every time step. Accordingly, the authors in (Alami et al., 2020) opted for an adaptation of the BOCPD approach to what they call restarted BOCPD (R-BOCPD) via modifying major components and introducing an explicit detection rule. For their adapted procedure, they proved first-order asymptotic minimax optimality regarding a adjusted measure of detection with the constraint of fixed probability of false alarm $\alpha_0^* \in (0,1)$ both for data generated from Bernoulli processes and a corresponding likelihood model in Laplace predictors. Further, they showed for R-BOCPD to outperform the regular BOCPD strategy with the same detection rule and to compare favourably with an improved GLR-CUSUM strategy introduced in (Maillard, 2019) for their specific problem via evaluation of simulation studies. We construct an approach similar to their adaptation via utilizing results in (Lai, 1998) and making use of their wide-reaching implications.

3.3. Modifying BOCPD for CUSUM Versatility

The main ideas in Bayesian online change point detection we want to exploit are in its intuitive way of implicit estimation of pre- and post-change parameters via conjugacy to access conditional pre- and post-change density functions as well as the aggregation of uncertainty for different scenarios via a weighted sum. It is only in

that sense that the proposed approach is still Bayesian. There seems to be a promising similarity to the Kalman setting via utilizing cheap conjugacy and conditional marginal distribution of the observations available. Furthermore, the marginals in (3.15) share a strong resemblance in interpretation to the observation evidence in the Kalman setting up to the condition in run-length r_n or equivalently in scenario k. The evidence in the Kalman context, akin to the model predictive in the BOCPD context, is hereby often taken to be the best quantity for evaluating online performance in practice with the latent signal not available. Accordingly, it makes intuitive sense to investigate BOCPD for inference in the Kalman setting when considering change, i.e. via signal noise mis-specification. However, as discussed, requiring access to the change point prior is a strong and restrictive assumption. As applicability is a central concern for Bayesian filtering, we want this requirement best avoided, especially given that we already consider specifying the signal noise a challenging task.
Moreover, recalling the versatility of the CUSUM procedures, its access to evaluation regarding reliability and its modification via the window-limitation to control computational complexity, we ideally want to make use of all of these properties. The idea then lies in working towards a formulation similar to BOCPD in exploiting conjugacy and marginals as well as uncertainty aggregation while losing requiring a change point prior and instead employing the versatility of the CUSUM strategies and its properties in reliability and computational feasibility.
The desired result will then be what we denote CUSUM restarted BOCPD (CR-BOCPD). It utilizes ideas in BOCPD for cheap access to pre- and post-change conditional density functions of the observations, the restart rule and ideas in adaptation from R-BOCPD and the statistical guarantees base on open-end SPRTs from non-iid CUSUM strategies.

Expressing BOCPD in Loss

Starting along the lines in (Alami et al., 2020), we also want to initially alter BOCPD similar to their R-BOCPD procedure. For choosing the geometric change point prior as above with parameter $\varphi \in (0, 1)$, we observe for the run-length posterior with adapted notation that

$$\begin{aligned} p(r_n \neq 0|y_{1:n}) &\propto (1-\varphi)p(y_n|r_{n-1}, y_{1:(n-1)})p(r_{n-1}|y_{1:(n-1)}) \\ p(r_n = 0|y_{1:n}) &\propto \varphi \sum_{r_{n-1}} p(y_n|r_{n-1}, y_{1:(n-1)})p(r_{n-1}|y_{1:(n-1)}), \end{aligned} \tag{3.19}$$

so utilizing proportionality again. Similarly, the run-length forecasts in (3.18) can be rewritten leaving out the model predictive in (3.19) to obtain $p(r_n|y_{1:(n-1)})$. We already observed $r_n = n - k \iff k = n - r_n$ to change intuition from run-lengths $r_n \in \{0, 1, \ldots, n-1\}$ to scenarios with a supposed change at time $k \in \{1, 2, \ldots, n\}$. The scenario $r_n = n \iff k = 0$ can be included to explicitly state the scenario of no change. We switch notation in $p(y_n|r_n, y_{1:(n-1)}) = p(y_n|y_{1:(n-1)}, k)$ and define a loss

$l_{k,n}$ and cumulative loss $L_{k,n}$ via

$$\begin{aligned} l_{k,n} &:= -\log[p(y_n|y_{1:(n-1)}, k)] \\ L_{k,n} &:= \sum_{s=k}^{n} l_{k,s} = -\sum_{s=k}^{n} \log[p(y_s|y_{1:(s-1)}, k)] = -\log[\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k)] \end{aligned} \tag{3.20}$$

at time n and for the scenario of assumed change at k. For the run-length posterior of length r_n, we want to write $p(r_n = n-k|y_{1:n}) = \tilde{\kappa}_{k,n}$ and its un-normalized form $\kappa_{k,n}$ to obtain the recursive, sequential learning formulation

$$\begin{aligned} \tilde{\kappa}_{k,n} &\propto \kappa_{k,n} = (1-\varphi)\exp(-l_{k,n})\kappa_{k,n-1}, \ \forall k < n \\ \tilde{\kappa}_{k,n} &\propto \kappa_{k,n} = \varphi \sum_{s=1}^{n-1} \exp(-l_{s,n})\kappa_{s,n-1}, \ k = n. \end{aligned} \tag{3.21}$$

Both, (3.19) and (3.21) provide a form to write the BOCPD approach, leaving out the evidence as it reduces to a normalization constant. In parallel we also want to adapt the one-step ahead forecasts from (3.18) to

$$\begin{aligned} \tilde{\kappa}^f_{k,n} &\propto \kappa_{k,n} = (1-\varphi)\kappa_{k,n-1}, \ \forall k < n \\ \tilde{\kappa}^f_{k,n} &\propto \kappa_{k,n} = \varphi \sum_{s=1}^{n-1} \kappa_{s,n-1}, \ k = n. \end{aligned} \tag{3.22}$$

As argued in (Alami et al., 2020), the sum over all current scenarios and its propagation is very challenging to evaluate theoretically yet provides a lot of the power of the BOCPD approach. It combines the likelihood of a change occurring with every scenario into the likelihood of change at the current time step over all scenarios, this way providing the prior for the new scenario $\kappa_{k,k}$. Exact evaluation of this initial weight for each scenario at every time step needs evaluating a combinatorial number of cumulative losses and is fairly intractable for theoretical work. Different from the R-BOCPD procedure in (Alami et al., 2020), we instead want to suggest a simplified initial weight given via

$$\bar{\kappa}_{r,k-1} = \exp(-L_{r,k-1}) = \prod_{s=r}^{k-1} p(y_s|y_{1:(s-1)}, r). \tag{3.23}$$

Adding the Restart Rule

The addition of a (re-)starting time $r \in \mathbb{N}_0$, not to be confused with the run-lengths r_n, hereby allows for the procedure to adapt to the last detected change point via a restart procedure for decision making as introduced in (Alami et al., 2020). $\bar{\kappa}_{r,k-1}$ is a reduced baseline of the initial weight in the second line of (3.21), much simplified without the sum passing down weight from the other scenarios but only from the current null hypothesis scenario of no change since the last restart at r. In essence, it describes the scenario of no change since the last detected change at time $r > 0$ or no change for $r = 0$. In this simplification lies the key for controlling evaluation criteria. With the introduction of the starting time, we also want to change notation of the scenario weights to $\kappa_{r,k,n}$ explicitly stating the initial weight $\bar{\kappa}_{r,k-1}$ via the index r. Accordingly, the additional index in r is required further down to indicate the current baseline explicitly depending on r at the time a respective scenario was initialized. For the adapted scenario weight $\kappa_{r,k,n}$ the indices denote (a) the last detected instance of change, thus the last restart, at time r (b) for the scenario of a new instance of change assumed at time k and (c) the current time n. While the triple indices seem clunky, they are required to convey all relevant information for the adapted scheme. Exact definition of the triple index scenario weights will follow.

Next, we want to also utilize the decision rule introduced in (Alami et al., 2020) via comparing the baseline scenario weight of no change since a restart at time r, so $\kappa_{r,r,n}$ with $r \leq n$, with each respective scenarios weight assuming instances of change at some later time k with $r < k \leq n$ given by $\kappa_{r,k,n}$, resulting in the stopping time

$$N_\kappa = \inf\{n \geq 1 : \max_{r<k\leq n} \kappa_{r,k,n} > \kappa_{r,r,n}\}. \tag{3.24}$$

The idea is straight forward in that for time steps $n < \eta$ with no change present, the weight tends to concentrate on the baseline weight $\kappa_{r,r,n}$. Contrary, when a scenario weight $\kappa_{r,k,n}$ overtakes the baseline weight, it is reasonable to assume for a change to have occurred at time k or in close proximity. To expand on the notion of the restart in the R-BOCPD procedure, we then want to delete all scenarios before time k, set $r = k$ and make it the new baseline for comparison, so again $\max_{r<k\leq n} \kappa_{r,k,n} > \kappa_{r,r,n}$ with the new value in r. In practice, this decision rule is very similar to tracking the MAP estimator over run-lengths posteriors in BOCPD as both mark the most plausible value as the most recent change point. The difference lies in making a decision whenever a more recent scenario overtakes the previously dominant scneario.

Specifying CR-BOCPD Scneario Weights in Tuning

The last difference introduced in (Alami et al., 2020), and this is also what mainly makes their approach interesting for us, lies in dropping the change point prior and introducing a new tuning parameter $\beta_{r,k,n}$ in its place. However, it also partially erases the Bayesian idea from BOCPD leaving only the implicit conjugacy. To keep similar asymptotic behavior to $(1-\varphi) \approx 1$, so small hazard rate or probability of change, we want to choose $\beta_{r,k,n}$ such that $\frac{\beta_{r,k,n}}{\beta_{r,k,n-1}} \to 1$ as $n \to \infty$. Accordingly, it is also this fraction replacing the change point prior. Notice again the triple index indicating that it can depend on all three relevant time points. While they can be required for matching tuning parameter and respective scenario weight, in practice they will likely be simplified to a constant value akin to the constant threshold in CUSUM startegies such as in (3.2).

Taking everything together, the result we want to propose is similar to the R-BOCPD procedure in (Alami et al., 2020), yet with one major difference. R-BOCPD procedure reduces to cross-sectional terms passing down in the last step via $\bar{\kappa}^{\text{R-BOCPD}}_{r,k-1} = \sum_{s=r}^{k-1} l_{s,k-1}$ for its simplification. Our approach simplifies to a longitudinal term, so passing down only from the most plausible scenario via $-\log(\bar{\kappa}_{r,k-1}) = \sum_{s=r}^{k-1} l_{r,s}$. We want to put emphasis on the scenario currently describing the data best and, more important, we want to bring together insights from established results in sequential change point detection via CUSUM strategies with the intuition in BOCPD of change in a latent Bayesian model. The result is the CUSUM restarted Bayesian online change point detection (CR-BOCPD) procedure described via the un-normalized weights

$$\begin{aligned} \kappa_{r,k,n} &= \frac{\beta_{r,k,n}}{\beta_{r,k,n-1}} \exp(-l_{k,n})\kappa_{r,k,n-1}, \ \forall k < n \\ \kappa_{r,k,n} &= \beta_{r,k,n} \exp(-l_{k,k})\bar{\kappa}_{r,n-1}, \ k = n, \end{aligned} \tag{3.25}$$

detection procedure via the stopping time (3.24), the initial conditions $r = 0$, $\kappa_{r,r,0} = 1$ and $\beta_{r,r,0}$=1 and the restart time r replaced whenever the stopping time activates. What is not yet specified, is the exact choice of $\beta_{r,k,n}$ for $k \in \{r+1, r+2, \ldots, n\}$ which is closely tied to the statistical guarantees via notions of false alarm rate and detection delay. In the case of a Bernoulli process generating the data, the authors proved for their R-BOCPD procedure and a choice of $\beta_{r,k,n} = \frac{1}{(n-k+1)}$ and a fixed PFA to attain their specified asymptotic minimax lower bound in detection delay as previously mentioned. However, they point out, that their approach makes use of very specific properties of Bernoulli distributions via concentration inequalities for controlling the cumulative loss.

Picking up on the idea of the one step ahead forecast, we can adapt the BOCPD formulation for this setting in

$$\begin{aligned} \kappa^f_{r,k,n} &= \frac{\beta_{r,k,n}}{\beta_{r,k,n-1}} \kappa_{r,k,n-1}, \ \forall k < n \\ \kappa^f_{r,k,n} &= \beta_{r,k,n}\bar{\kappa}_{r,n-1}, \ k = n. \end{aligned} \tag{3.26}$$

However, these forecasts do not serve as a detection rule or run-length forecast as with BOCPD. Ideally, they provide a way to incorporate a newly initialized scenario at $k = n$ into an aggregated forecast, but for the moment they need to be handled with great care needing much more further investigation.

3.3.1. Constructing the Link

As said, we want to denote our derived approach CUSUM restarted Bayesian online change point detection (CR-BOCPD). We take the Bayesian model for the density functions with an implicit conjugacy and marginalization from BOCPD, we take the idea of simplification of the initial weight, the restart rule and the tuning parameter for dropping the change point prior from R-BOCPD, and we construct each component in a way to resemble a non-*iid* CUSUM procedure with conditional density

functions. Accordingly, the basic idea of the CR-BOCPD is in achieving equivalence of (3.24) to the detection rule of the CUSUM procedure in (3.2) with the decision threshold incorporated in the tuning parameter $\beta_{r,k,n}$.
Set $\beta_{r,r,n} = 1$ for $n \geq 1$. Looking at the decision rule in (3.24) given by $\kappa_{r,k,n} > \kappa_{r,r,n}$, we observe for $k \in \{r+1, r+2, \dots, n\}$ that

$$\begin{aligned}
&\kappa_{r,k,n} > \kappa_{r,r,n} \\
\iff &\frac{\beta_{r,k,n}}{\beta_{r,k,n-1}} \exp(-l_{k,n}) \kappa_{r,k,n-1} > \frac{\beta_{r,r,n}}{\beta_{r,r,n-1}} \exp(-l_{r,n}) \kappa_{r,r,n-1} \\
\iff &\beta_{r,k,n} \exp(-L_{k,n}) \cdot \bar{\kappa}_{r,k-1} > \bar{\kappa}_{r,n} \\
\iff &\beta_{r,k,n} \exp(-L_{k,n}) \cdot \exp(-L_{r,k-1}) > \exp(-L_{r,n}) \\
\iff &\log(\beta_{r,k,n}) - L_{k,n} - L_{r,k-1} > -L_{r,n} \\
\iff &\log(\beta_{r,k,n}) + \log(\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k)) + \log(\prod_{s=r}^{k-1} p(y_s|y_{1:(s-1)}, r)) > \log(\prod_{s=r}^{n} p(y_s|y_{1:(s-1)}, r)) \\
\iff &\log(\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k)) + \log(\prod_{s=r}^{k-1} p(y_s|y_{1:(s-1)}, r)) - \log(\prod_{s=r}^{n} p(y_s|y_{1:(s-1)}, r)) > -\log(\beta_{r,k,n}) \\
\iff &\log[\frac{\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k) \cdot \prod_{s=r}^{k-1} p(y_s|y_{1:(s-1)}, r)}{\prod_{s=r}^{n} p(y_s|y_{1:(s-1)}, r)}] = \log[\frac{\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k)}{\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, r)}] > -\log(\beta_{r,k,n}) \\
\iff &\sum_{s=k}^{n} \log[\frac{p(y_s|y_{1:(s-1)}, k)}{p(y_s|y_{1:(s-1)}, r)}] > -\log(\beta_{r,k,n}).
\end{aligned} \tag{3.27}$$

Setting $S_{r,k,n} = \log(\frac{\prod_{s=k}^{n} p(y_s|y_{1:(s-1)},k)}{\prod_{s=k}^{n} p(y_s|y_{1:(s-1)},r)})$ and plugging this formulation back in the stopping time (3.24), we obtain

$$N_\kappa = \inf\{n \geq 1 : \max_{r<k\leq n} S_{r,k,n} > -\log(\beta_{r,k,n})\} \tag{3.28}$$

resembling the CUSUM procedure with pre-change density function $p(\cdot|y_{1:(n-1)}, r)$ and post-change density function $p(\cdot|y_{1:(n-1)}, k)$ at time n and decision threshold $-\log(\beta_{r,k,n})$. This decision threshold can then be chosen in a curved manner, as in (Alami et al., 2020) so depending on r, k and n, but also constant, either way satisfy $\frac{\beta_{r,k,n}}{\beta_{r,k,n-1}} \to 1$ as $n \to \infty$, i.e let $-\log(\beta_{r,k,n}) = c \sim -\log(\alpha_0) \iff \beta_{r,k,n} = \exp(-c) \sim \alpha_0$ for $\alpha_0 \in (0, 1]$ the constrained in fixed FAR from the previous section.
In practice, this translates to the idea of initializing a new scenario k and weight it down by $\beta_{r,k,n} \sim \alpha_0$. We therefore decide against our current baseline scenario and for a scenario initialized at k when

$$\begin{aligned}
&\kappa_{r,k,n} > \kappa_{r,r,n} \\
\iff &\log[\beta_{r,k,n} \prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k)] > \log[\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, r)] \\
\iff &\alpha_0 \prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k) > \prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, r),
\end{aligned} \tag{3.29}$$

so the later scenario of supposed change at time $k \in \{r+1, r+2, \dots, n\}$ overtaking the baseline scenario for last detected change at $r \in \mathbb{N}_0$ although weighted down by the FAR in $\alpha_0 \in (0, 1]$ for evidence accumulated over the observations since k, so of the supposed post-change regime.

The intuition is still the same as with BOCPD: Can the the latent conjugated model learning only on observations from a supposed post-change regime describe these observations substantially better than the latent conjugated model learning on all observations

Another insight is that this relation between the CUSUM procedure and the CR-BOCPD procedure goes both ways. For appropriate choices of pre- and post-change densities in the CR-BOCPD procedure the theoretical results in (Lai, 1998) as well as the results on window-limited detection hold, and vice versa we may write CUSUM procedures as scenarios weights if we can split the log-likelihood ratio statistics adequately.

Investigating Scenario Initialization

Taking a concrete look at the more general pre- and post-change densities in (Adams and MacKay, 2007) and (Altamirano et al., 2023b) compared to the specific choice of Bernoulli processes in (Alami et al., 2020), we observe that the conditional density functions $p(y_n | y_{1:(n-1)}, s)$ obtained via conjugacy and marginalization of an unknown parameter represent a special case of weighted likelihoods as in (3.8). To be precise, let $\theta \in \Theta$ be that unknown parameter regarding the density function $p(\cdot|\theta)$. Further, we assume the prior distribution $g(\theta)$ on the parameter to obtain the respective posterior distribution $p(\theta | y_{s:(n-1)}) \propto \prod_{i=s}^{n-1} p(y_i|\theta) g(\theta)$. The pre- and post-change density functions are the posterior-predictive densities obtained via $p(y_n | y_{1:(n-1)}, s) = \int_\Theta p(y_n|\theta) p(\theta|y_{s:(n-1)}) \mathrm{d}\theta$ for different choices of s, i.e. $s = r$ for the baseline scenario and $s = k$ for later scenarios of assumed change at $k > r$. The posterior predictive implicitly contains the induced uncertainty in the pre- and post-change parameters as well as reduction on the uncertainty via available observations. In practice, we want to choose $g(\theta)$ such that it reflects a balance of our knowledge about the pre- and post-change parameters. In the assumed pre-change regime $r \leq n < k$, the posterior distribution of the latent parameter θ is taken to adapt to the true pre-change parameter. Specifically, at time point k of assumed change, we compare Bayesian learning on the priors of the latent parameter in (a) $p(\theta|y_{r:(k-1)})$ considering observations $y_{r:(k-1)}$, so no change, opposed to (b) $g(\theta)$ for no observations, so change. Therefore, comparing the pre-change density $p(\cdot|y_{1:(n-1)}, r)$ and post-change density $p(\cdot|y_{1:(n-1)}, k)$ comes down to comparing two weighted likelihoods and, furthermore, two conditional densities of the respective posterior predictive after marginalizing the unknown parameter $\theta \in \Theta$:

$$\begin{aligned} p(y_n|y_{1:(n-1)}, r) &= \int_\Theta p(y_n|\theta)p(\theta|y_{r:(n-1)})\mathrm{d}\theta \\ &\propto \int_\Theta \prod_{s=k}^{n} p(y_s|\theta)p(\theta|y_{r:(k-1)})\mathrm{d}\theta \\ p(y_n|y_{1:(n-1)}, k) &= \int_\Theta p(y_n|\theta)p(\theta|y_{k:(n-1)})\mathrm{d}\theta \\ &\propto \int_\Theta \prod_{s=k}^{n} p(y_s|\theta)g(\theta)\mathrm{d}\theta. \end{aligned} \tag{3.30}$$

The same result can be generalized to $p(\theta|y_{1:(n-1)}, k)$ with scenario initialization $p(\theta|y_{1:(k-1)}, k)$ replacing the general prior in $g(\theta)$ at time step k. In other words, a change of prior for the respective sub-sequences.

The CR-BOCPD procedure adapts an *iid* weighted CUSUM procedure with unknown pre- and post-change parameters, independent observations and known prior distributions corresponding to change at certain points in time to a non-*iid* CUSUM procedure with known conditional pre- and post-change density functions via exploiting closed form solutions of conjugated likelihood-prior pairs and marginalization - as available for linear Gaussian state space models via Kalman filters with adapted initial conditions for scenario initialization at k. In simple terms, the prior and the Bayesian learning is included in the respective known conditional density functions with prior and learning being changed at scenario initialization for assumed instances of change.
To conclude, the results in (Lai, 1998) for non-*iid* CUSUM procedures with conditional density functions can therefore be transferred to the CR-BOCPD procedure to obtain first-order asymptotic optimality regarding the introduced notions as well as a window-limited adaptation.

Proposition 3 ***First-Order Asymptotic Optimality of CR-BOCPD***
For a given rate of false alarm $\mathrm{FAR}(N_\kappa) = \alpha_0$ *and appropriate tuning parameter* $\beta_{r,k,n} = \exp(-c)$ *such that* $c \sim |\log(\alpha_0)|$*, then it holds for the stopping time* N_κ *in (3.24) with (3.25), that*

$$\mathrm{WADD}(N_\kappa) \sim \mathrm{CADD}(N_\kappa) \sim \frac{|\log(\alpha_0)|}{I} \tag{3.31}$$

as $\alpha_0 \to 0$ *for a positive constant* I *as initially introduced. Therefore, the stopping time* N_κ *is first-order asymptotic optimal regarding the introduced notions of minimax optimality.*

The *proof* is in the derivation of the proposition via the previous sections of this chapter via expressing the CR-BOCPD detection rule as a specific case of the non-*iid* CUSUM rule in (3.2).

3.4. Detecting Additive Change in Gaussian Models: State-of-the Art

As initially motivated, in the scope of this chapter of addressing signal noise mis-specification, we are mainly interested in detecting and accounting for abrupt additive terms to the signal process not covered by the model, i.e. outliers or realization of heavy-tailedness of the true signal noise. Taking a linear Gaussian state space model and picking up the initial understanding of an instance of change via a sudden change in the parameters of the data generating process, an additive term on the signal process translates to such an abrupt change in mean of the data generating process not accounted for by the model.
The data generating process is hereby the observation model $Y_n = H_n X_n + \Gamma_n V_n$ with $p(y_n|x_n) \sim n(y_n; H_n x_n, R_n)$ and support $\mathcal{Y} = \mathbb{R}^p$ in the Kalman setting (see (2.1) in the previous chapter for details). Accordingly, the resulting challenge is in sequential change point detection of change in mean in the observation sequence $Y_{1:n}$. Framing this approach in detecting instances of mis-specification has not been done before. However, the general problem statement of detecting additive change in the signal process has been around since pioneering work in (A. S. Willsky and Jones, 1974). A strong line of work on the problem with milestones in (A. S. Willsky, 1976), (Basseville and Benveniste, 1983), (Basseville, Nikiforov, et al., 1993), (Lai and Shan, 1999), (I. V. Nikiforov, 2001), (A. Tartakovsky et al., 2014) and (Brodsky, 2016). Specific terminology in application hereby involves fault detection, failure detection and system integrity monitoring.

In essence, an instance or realization of signal noise mis-specification in heavy tailedness, may lead to a very different signal trajectory then anticipated by the signal model, and thus to a change in the mean as generative parameter of the observations post-change. Accordingly, we are interested in detecting change in mean of the observation sequence $Y_{1:n}$, however, as argued in (A. S. Willsky and Jones, 1974) and supported in the majority of the body of work that followed, we want to simplify the problem to detecting change in mean of the innovation sequence $\gamma_n = y_n - H_n m_n^f$ with $p(\gamma_n|y_{1:(n-1)}) = n(\gamma_n; 0, \Sigma_n)$ and $\Sigma_n = H_n P_n^f H_n^T + R_n$. In the pre-change regime, the innovation is a 0 mean random vector with known covariance matrix Σ_n. The enabling property hereby is in that the forecast mean mapped to observation space in the Kalman setting functions as a BLUE estimator (see (Reich and Cotter, 2015) for additional details) for the well-specified model. Accordingly, the problem at hand transfers nicely to the above described setting of known pre-change and unknown post-change density function with the challenge in detecting instances of change to non-zero mean. While the task at hand gets noticeably more difficult with the Kalman filter adapting to the additive change over time, we first want to briefly recall current popular approaches available for testing differences in the mean of Gaussian sequences as well as the SPRTs and the corresponding CUSUM procedures. The central tool hereby is again the Mahalanobis distance and its property to result in χ^2-distributions for Gaussian random vectors. We hereby mainly follow (Basseville, Nikiforov, et al., 1993) and (A. Tartakovsky et al., 2014) in their construction up to the χ^2-CUSUM procedure.

The central idea of this section is in retracing the steps that leads to the current state-of-the-art practice as a second line of work next to the BOCPD strain developed here. This way we want to point out difference in assumptions and shared ideas to emphasize when the novel approach constructed in the first half of the chapter via conditional SPRTs is competitive.

3.4.1. Testing in Mean of Gaussian Sequences

Proficiency in Fixed Sample Size Tests

As motivated, we are interested in reliably detecting relevant deviations from 0 mean for Gaussian sequences, ideally under appropriate notions of optimality. Different challenges emerge with the wording of the task. We want to start with a Gaussian data RV $Z \sim \mathcal{N}(\mu, \Sigma)$ with support in $\mathcal{Z} = \mathbb{R}^p$. The according challenge is therefore in testing between $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. In order to address reliability and optimiality, we need to go beyond the idea of the usual notion of the uniformly most powerful test. As argued in the original work by Wald (see (Wald, 1943)), the alternative hypothesis is "too rich" in its general statement and additional constraints are required resulting in Wald's idea of the uniformly best constant power test (UBCP). Following (A. Tartakovsky et al., 2014) for the case at hand, Wald proposed to impose a constant power function on a family of surfaces $\mathcal{S} = \{S_d : ||\mu||_2^2 = d^2, c > 0\}$ on the high-dimensional space of $M = \{\mu \in \mathbb{R}^p\}$. For a statistical test Λ, let $\Lambda^* \in C(\alpha_0^*) = \{\Lambda : \mathbb{P}_0(\Lambda \neq 0) \leq \alpha_0^*\}$ the set of test satisfy the prescribed type-I error $\alpha_0^* \in (0, 1)$. We then say for Λ^* to be UBCP on $\mathcal{S}$, if (Wald, 1943)

- for any two μ_1, μ_2 on the same surface $S_d \in \mathcal{S}$, the power function give by $\alpha_1^*(\Lambda, \mu) = \mathbb{P}_\mu(\Lambda = 1)$ has $\alpha_1^*(\Lambda^*, \mu_1) = \alpha_1^*(\Lambda^*, \mu_2)$ and
- for any other test $\Lambda \in C(\alpha_0^*)$ which satisfies the previous condition, it holds that $\alpha_1^*(\Lambda^*, \mu) \geq \alpha_1^*(\Lambda, \mu)$.

Next to the spheres, an intuitive choice of surfaces lies in the ellipsoids $S_b = \{\mu : ||\mu||_{\Sigma^{-1}}^2 = b^2\}$ centered around $\mu_0 = 0$. Recall, $||\mu||_{\Sigma^{-1}}^2 = \mu^T \Sigma^{-1} \mu$ was introduced in the previous chapter as Mahalanobis distance, however, is also frequently referred to as signal-to-noise ratio.

The central enabling result is the theorem by Wald in (Wald, 1943) which proofs for the case of unit covariance with spherical surfaces S_d and $d > 0$, that the test

$$\Lambda^*(\tilde{Z}) = \begin{cases} 1, & \text{if } ||\tilde{Z}||_2 \geq h(\alpha_0^*) \\ 0, & \text{if } ||\tilde{Z}||_2 < h(\alpha_0^*) \end{cases} \tag{3.32}$$

is UBCP for $\tilde{Z} \sim \mathcal{N}(0, \mathbf{1}_{p \times p})$ and $\Lambda^*(\tilde{Z}) \in C(\alpha_0^*)$ and the above set of hypothesis. The proof can then be generalised to arbitrary positive definite covariance matrices via substituting $\tilde{\mu} = \Sigma^{-\frac{1}{2}} \mu$ and invariance properties of Gaussian distributions such that the hypothesis pair remains invariant for resulting ellipsoid surfaces (see (A.

Tartakovsky et al., 2014) for additional details and an adapted theorem). Moreover, the theorem generally also holds for Gaussian linear models such as $Z = H\mu + W$ with noise $W \sim \mathcal{N}(0, \Sigma)$ in that we are then interested in the quantity $||H\mu||^2_{\Sigma^{-1}}$ for our UBCP test, thus arguing for the Mahalonbis distance to be our tool of interest here. Note hereby the direct applicability to our context of interest up to time varying covariance and the required invariant transformation. As a remark, we then no longer test about μ but instead are interested in testing the mapped parameter $H\mu$. Following additional arguments in (A. Tartakovsky et al., 2014) via least favorable distributions invariant under transformation, arguments can be expanded for the introduced test to be minimax. An important enabling feature lies hereby in that the statistic $||Z||^2_{\Sigma^{-1}}$ is distributed according to a $\chi^2(p, a^2)$ (non-central chi-square) distribution with p degrees of freedom, given via the dimension of the RV, and non-centrality parameter $a^2 = ||Z||^2_{\Sigma^{-1}}$, thus providing good evaluation of the power function.

To conclude, utilizing the Mahalonbis distance for testing deviation from 0 mean generally satisfies the UBCP constraints as well as providing a minimax approach under additional considerations such as a given minimal deviation $||\mu||^2_{\Sigma^{-1}} > b^2$ for $b > 0$ under the alternative hypothesis. The χ^2 distribution of the Mahalnobis distance for Gaussian RVs establishes it as a central tool in controlling uncertainties for the task at hand.

Towards Proficiency in χ^2-SPRTs

Taking the step from fixed sample size tests, we want to transfer the ideas to the concept of the sequential probability ratio test to obtain the popular adapted CUSUM procedure. The χ^2-SPRT as taken from (A. Tartakovsky et al., 2014), expands on the previous results. Let $Y_{1:n} \sim_{iid} \mathcal{N}(\mu, \Sigma)$ and parameter space $M = \{\mu \in \mathbb{R}^p\}$. Picking up on the previously derived UBCP test, we are interested in sequentially testing the hypothesis pair $H_0 : \mu = 0$ vs $H_1 : ||\mu||^2_{\Sigma^{-1}} = b^2$ for some minimal deviation $b > 0$. Let $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ be the empirical mean and the sufficient statistic for testing between H_0 and H_1. For $\sqrt{n}\bar{Y}_n \sim \mathcal{N}(\sqrt{n}\mu, \Sigma)$, we observe $n||\bar{Y}_n||^2_{\Sigma^{-1}} \sim \chi^2(p, n||\mu||^2_{\Sigma^{-1}})$. Therefore, the sufficient statistics now takes the place of the Gaussian RV Z from the previous paragraph.

After transformation the initial problem may be reduced to testing the non-centrality parameter a^2 of a non-central χ^2-distribution in $H_0 : a^2 = 0$ vs $H_1 : a^2 = b^2$ with minimal deviation $b > 0$. As given in (A. Tartakovsky et al., 2014), this transformation can hereby be expressed in accumulating the observation sequence $Y_{1:n}$ into the sequence of scaled empirical means $\{i||\bar{Y}_i||^2_{\Sigma^{-1}}\}_{1 \le i \le n}$. As in the first section of this chapter, we are interested in evaluating the log-likelihood ratio statistics

$$S_n = \log(\frac{f_{b^2}(||\bar{Y}_1||^2_{\Sigma^{-1}}, 2||\bar{Y}_2||^2_{\Sigma^{-1}}, \dots, n||\bar{Y}_n||^2_{\Sigma^{-1}})}{f_0(||\bar{Y}_1||^2_{\Sigma^{-1}}, 2||\bar{Y}_2||^2_{\Sigma^{-1}}, \dots, n||\bar{Y}_n||^2_{\Sigma^{-1}})} \tag{3.33}$$

with f_{a^2} the joined density function of the sufficient statistics for non-centrality parameter $a^2 \ge 0$. Direct computation of this ratio is challenging with several obstacles. Yet, key investigations in (Jackson and Bradley, 1961) on the sequential χ^2-test utilized Cox's theorem for the factorization

$$f_{a^2}(||\bar{Y}_1||^2_{\Sigma^{-1}}, 2||\bar{Y}_2||^2_{\Sigma^{-1}}, \ldots, n||\bar{Y}_n||^2_{\Sigma^{-1}}) = \underbrace{f_0(n^2||\bar{Y}_n||^2_{\Sigma^{-1}})}_{\text{df}} \exp(-\frac{na^2}{2})G(\frac{p}{2}, \frac{a^2n^2||\bar{Y}_n||^2_{\Sigma^{-1}}}{4}) \tag{3.34}$$

with G the generalized hyper-geometric function $G(r,s) = \sum_{i=0}^{\infty} \frac{\Gamma(r)s^i}{\Gamma(r+i)i!}$ and $\Gamma(r)$ the usual gamma function. This the leads to the adaptation of the LLR in (3.33) to

$$S_n = -\frac{nb^2}{2} + \log[G(\frac{p}{2}, \frac{b^2n^2||\bar{Y}_n||^2_{\Sigma^{-1}}}{4})] \tag{3.35}$$

via reducing the denominator of the LLR for $a^2 = 0$ as $G(r, 0) = 1 \iff \log(G(r, 0)) = 0$ and the density functions f_0 cancelling out. The resulting closed SPRT is then the stopping time

$$\tilde{T} = \inf\{n \geq 0 : S_n \notin (-c_0, c_1)\} \tag{3.36}$$

with decision rule

$$\tilde{\Lambda}(Y) = \begin{cases} 1, & \text{if } S_{\tilde{T} \geq c_1} \\ 0, & \text{if } S_{\tilde{T} \leq -c_0} \end{cases}, \tag{3.37}$$

the sequential χ^2-test or χ^2-SPRT. Further following results in (A. Tartakovsky et al., 2014), the derived invariant sequential χ^2-test is asymptotically optimal in that it minimizes all positive moments of the stopping time $\tilde{T}$ with $\tilde{\Lambda}$ satisfying α_0^* and α_1^* the respective fixed type-I and type-II error. So, as $\max\{\alpha_0^*, \alpha_1^*\} \to 0$,

$$\begin{aligned} \inf_{\Lambda(\alpha_0^*,\alpha_1^*)} \mathbb{E}_1[T] \sim \mathbb{E}_1[\tilde{T}] \sim \frac{2|\log(\alpha_0^*)|}{b^2}, \\ \inf_{\Lambda(\alpha_0^*,\alpha_1^*)} \mathbb{E}_0[T] \sim \mathbb{E}_0[\tilde{T}] \sim \frac{2|\log(\alpha_1^*)|}{b^2} \end{aligned} \tag{3.38}$$

for $\Lambda(\alpha_0^*, \alpha_1^*)$ tests with the respective errors, and thresholds chosen appropriately. Notice hereby the close similarity of the asymptotic optimal lower bound on stopping time of the SPRT in comparison to Lai's asymptotic lower bound for detection delay, also stopping time, of change point detection procedures (Lai, 1998). The KL-divergence of the null and alternative hypothesis hereby reduces to $\frac{1}{2}(\mu_0 - \mu_1)^T\Sigma^{-1}(\mu_0 - \mu_1) = \frac{1}{2}||\mu_1||^2_{\Sigma^{-1}} = \frac{b^2}{2}$ up to an additional constant and takes the role of the positive constant I in (3.5) as we assume the iid case (see (Duchi, 2007) or the previous chapter for details on the KL-divergence of two Gaussian random variables). As the corresponding CUSUM procedure is constructed via the open-end SPRT, this relation makes intuitive sense.

In a brief simulation study via the relative efficiency comparing expected sample size of the χ^2-SPRT to the fixed sample size UBCP χ^2-test, the authors in (A. Tartakovsky et al., 2014) found for the SPRT to generally perform better with almost twice as good efficiency for fixed small error probabilities compared to the UBCP. Twice as good efficiency hereby refers to the SPRT requiring about half the samples of the FSS test to achieve the same fixed errors.

3.4.2. Detecting Deviation From 0 Mean Gaussian Sequences

Combining the obtained insights on the SPRT for the mean of a Gaussian sequence via empirical means with the previous considerations of the CUSUM procedure as open-end SPRT with type-II error $\alpha_1^* = 1$, then taking the step from the χ^2-SPRT to the χ^2-CUSUM procedure is imminent.
Given the independent sequence $\{Y_n\}_{n\geq 1}$ of Gaussian random vectors with change in mean at $\eta \geq 1$, so

$$Y_n \sim \begin{cases} \mathcal{N}(0, \Sigma), & \text{if } n < \eta \\ \mathcal{N}(\mu, \Sigma), & \text{if } n \geq \eta. \end{cases} \tag{3.39}$$

We hereby assume pre-change mean 0 and the covariance matrix Σ to be known. For the post change mean we assume known Mahalanobis distance $||\mu||^2_{\Sigma^{-1}} = b^2$, thus known KL-information $D_{\mathrm{KL}}(f_0||f_1) = \frac{b^2}{2}$ and recall the asymptotic lower bound on Pollak's CADD and Lorden's WADD for given FAR via $\alpha_0 \to 0$ as in (3.5) via

$$\mathrm{WADD}(T) \sim \mathrm{CADD}(T) \sim \frac{2|\log(\alpha_0)|}{b^2}. \tag{3.40}$$

Obtaining the χ^2-CUSUM Procedure

Following (A. Tartakovsky et al., 2014), we are interested in adapting the CUSUM procedure (3.2) via retracing the steps of transforming the problem from detecting change in Gaussian mean to change of the non-centrality of the corresponding χ^2 sequence of empirical means under invariant transformation where required. The LLR in the regular CUSUM procedure is replaced with the derived χ^2-LLR in (3.35) resulting in

$$\begin{aligned} \bar{Y}_{k:n} &= \frac{1}{n-k+1}\sum_{i=k}^{n} Y_i \\ S^{\chi}_{k:n} &= -(n-k+1)\frac{b^2}{2} + \log[G(\frac{p}{2}, \frac{b^2(n-k+1)^2||\bar{Y}_{k:n}||^2_{\Sigma^{-1}}}{4})] \\ N^{\chi} &= \inf\{n \geq 1 : \max_{1\leq k\leq n} S^{\chi}_{k:n} > c\} \end{aligned} \tag{3.41}$$

and the corresponding window-limited adaptation

$$N^{\chi}_{m(\alpha_0)} = \inf\{n \geq 1 : \max_{(n-m(\alpha_0))\leq k\leq n} S^{\chi}_{k:n} > c\}. \tag{3.42}$$

Note hereby the pre-change density function vanishing via $-\log(1) = 0$ after factorization based on Cox's theorem. However, this still allows for a naive split of pre- and post-change density functions as derived previously for the CR-BOCPD procedure leading to the reformulation

$$N^{\chi} = \inf\{n \geq 1 : \max_{r<k\leq n}[\exp(-c)\exp(-(n-k+1)\frac{b^2}{2})G(\frac{p}{2}, \frac{b^2(n-k+1)^2||\bar{Y}_{k:n}||^2_{\Sigma^{-1}}}{4})] > 1\} \tag{3.43}$$

with $\kappa_{r,k,n}$ on the left hand side of the inequality and $\kappa_{r,r,n}$ on the right hand side.

While this popular χ^2-CUSUM procedure is the direct derivation via the open-end χ^2-SPRT, it is non-recursive in needing computing the sufficient statistics for each scenario within a window. Recursive modifications utilizing repeated closed χ^2-SPRTs have been developed in (I. V. Nikiforov, 2001) and previous works. Considering the goal of obtaining scenario weights for inference for considering change arising via signal noise mis-specification, we want to remain with the derived open-end procedure.

Expanding to χ^2-GLR CUSUM Procedures

Next to the regular CUSUM procedure with the adapted LLR, the GLR CUSUM scheme has been studied with initial work in (A. Willsky and Jones, 1976) and rigorous results in (Lai, 1998) and (Lai and Shan, 1999). The enabling feature of the approach is in exploiting the sequence of sufficient statistics also utilized in the χ^2-SPRT for easy estimation of the MLE of the parameter of the supposed post-change regime. A crucial upside hereby lies in that there is no need for assumption of the KL-information or Mahalanobis distance of the post-change mean. As shown in (A. Tartakovsky et al., 2014), the LLR with the MLE as the post change parameter simply reduces to plugging in the empirical mean for the post change parameter. Picking up at (3.10) and modifying for the χ^2-GLR CUSUM scheme, we obtain

$$\begin{aligned} \max_{\mu} S_{k:n}(\mu) &= \frac{n-k+1}{2}||\bar{Y}_{k:n}||^2_{\Sigma^{-1}} \\ N^{\bar{Y}} &= \inf\{n \geq 1 : \max_{1\leq k\leq n} \frac{n-k+1}{2}||\bar{Y}_{k:n}||^2_{\Sigma^{-1}} > c\} \end{aligned} \tag{3.44}$$

with $S_{k:n}(\mu)$ the parameter dependent LLR taking observation from time k to n akin to (3.9). Further, we obtain the corresponding window-limited adaptation

$$N^{\bar{Y}}_{m(\alpha_0)} = \inf\{n \geq 1 : \max_{(n-m(\alpha_0))\leq k\leq n} \frac{n-k+1}{2}||\bar{Y}_{k:n}||^2_{\Sigma^{-1}} > c\}. \tag{3.45}$$

The 0 mean pre-change density function again vanishes with the above reformulation available in

$$N^{\bar{Y}} = \inf\{n \geq 1 : \max_{r<k\leq n} \exp(-c)\exp(\frac{n-k+1}{2}||\bar{Y}_{k:n}||^2_{\Sigma^{-1}}) > 1\}. \tag{3.46}$$

Proficiency of the Derived χ^2-Detection Strategies

The first-order asymptotic optimality regarding the introduced notions of minimax optimality in detection delay of either approach can be derived via Lai's arguments in (Lai, 1998). However, it was also derived in (I. Nikiforov, 1994) and (I. V. Nikiforov, 1999) with proposing either strategies. Approaching the χ^2-CUSUM procedure with Lai's arguments, the assumption on knowledge about the post-change density function translates to knowledge of the KL-information, magnitude or minimal deviation, so the parameter b^2. Nikiforov's arguments are directly derived from the sequential test of non-centrality and its properties.

A highly interesting case for our purpose is given when there is no such parameter of minimal deviation or KL-divergence of pre- and post-chnage density function is available. We then want to insetad assume a minimal deviation $b^2 > 0$, or equivalently KL-information $\frac{b^2}{2}$, and perform the χ^2-CUSUM procedure given that choice. Let the true post-change mean be such that $||\mu||^2_{\Sigma^{-1}} = a^2 > \frac{b^2}{4}$. In (A. Tartakovsky et al., 2014) it is then shown that the asymptotic relation of the threshold c in a modified χ^2-CUSUM procedure with repeated closed χ^2-SPRTs and Lorden's WADD is given by

$$\mathrm{WADD}(N) \sim \frac{2c}{a^2 - (a-b)^2} \tag{3.47}$$

as $c \to \infty$ or, similarly, for a given rate of false alarm $\mathrm{FAR}(N) = \alpha_0$ with $c \sim |\log(\alpha_0)|$

$$\mathrm{WADD}(N) \leq \min\{\frac{2|\log(\alpha_0)|}{a^2 - (a-b)^2}, \frac{1}{\alpha_0}\}(1 + o(1)) \tag{3.48}$$

as $\alpha_0 \to 0$. The condition $a^2 > \frac{b^2}{4}$ is hereby required to ensures a positive denominator.

The key takeaway is then, that for a well chosen assumption on the minimal deviation, so $b^2 \approx a^2$, it will recover close to first-order asymptotic optimal results regarding worst average detection delay, however, for larger deviations between assumed and true value the asymptotic relation of detection delay and threshold or FAR is prone to blow up. It makes sense that a similar intuition will hold up for the regular open-end SPRT based χ^2-CUSUM procedure and assumed minimal deviation $b^2 > 0$.

This case of unknown KL-information or minimal deviation in (A. Tartakovsky et al., 2014) is treated similar to the case of unknown mean μ in (Basseville, Nikiforov, et al., 1993). Akin to assuming a value for the procedure and running with it, the older work suggests to use the limiting case $a^2 \to 0$ resulting in the approach in (3.44).

Providing an Additional Perspective

For another suitable expression of the problem especially valuable for the linear Gaussian setting at hand, we can also exploit the weighted CUSUM procedure in (3.8) for detecting change in mean of the Gaussian sequence. The idea is hereby in a Gaussian assumption on the additive change $u \sim \mathcal{N}(\bar{u}, U)$. Again, take the same independent sequence $\{Y_n\}_{n\geq 1}$ of Gaussian random vectors with additive change in mean at $\eta \geq 1$ with known Gaussian profile of change u, so

$$Y_n = W_n + \mathbf{1}\{n \geq \eta\}u = \begin{cases} 0 + W_n, & \text{if } n < \eta \\ u + W_n, & \text{if } n \geq \eta \end{cases} \tag{3.49}$$

for $W_n \sim \mathcal{N}(0, \Sigma)$. This then resembles a weighted CUSUM procedure with a prior on the unknown post-change parameter. In practice, we can simply aggregate the parameter uncertainty into a known post-change density function $p_u(y_n) = n(y_n; \bar{u}, \Sigma + U)$ for $n \geq \eta$. Moreover, with the results in (Lai, 1998) we can then again conclude first-order asymptotic optimality for the resulting CUSUM procedure and its window-limited adaptation for appropriate constraints on the FAR and window-size. The arguments hereby are very similar to the CR-BOCPD procedure. In BOCPD and CR-BOCPD each new scenario resets to a specific prior usually leading to a much larger covariance allowing to adapt to the post-change regime. Choosing a probabilistic profile u of the post-change mean in a similar way that it roughly represents the initial prior distribution up to a large covariance matrix U will therefore provide a similar effect to resetting for a sub-sequence of observations. While this intuition is somewhat overly ambitious in the regular Gaussian sequence, it will be more intuitive for detecting change under Bayesian learning as with the Kalman filter.
Note hereby that the model is the same as before, however the change in perspective might suit readers from a more dynamical systems driven background.

3.5. Deriving Adaptive Kalman Filtering Strategies

So far we have spent plenty of work in this chapter on deriving strategies for detecting change (a) via conditional density functions of the observations with a latent conjugated likelihood-prior Bayesian model under considerations of scenario priors and (b) detecting deviation from 0 mean in centered Gaussian sequence. A main takeaway lies in that each provides different approaches and may be applicable given different assumptions. Detecting abrupt jumps or more general additive change in the signal process given the Kalman setting, e.g. resulting from misspecification in heavy-tailedness of the signal noise, under ongoing estimation of the signal via the Kalman filter causes additional challenges to emerge. Recall, a central focus for the work at hand is in inference on the latent signal under threat of change.

Briefly Recollecting the Problem History

The main problem statement and pioneering work was done in (A. S. Willsky and Jones, 1974) and (A. Willsky and Jones, 1976). The foundation of their approach was in deriving equations for determining the dynamic profile of additive terms in the signal process with the Kalman filter adapting to them. The central, enabling feature hereby is in the linear structure of the signal and observation equations. Given knowledge of the dynamical profile, they then implemented a GLR scheme to estimate location and magnitude of the most likely instance of change to test against the null hypothesis of no change. As the initial approach needed tracking a new dynamic profile for each new observation, they suggested the aforementioned window-limited approach for computational feasibility. The more general work in (Lai, 1998) was followed by specifications in (Lai and Shan, 1999) expanding on these previous by Willsky and Jones with results via thorough statistical analysis of the introduced notions of asymptotic optimality regarding choice of decision threshold and window size. Expansive analysis in the monograph (Basseville, Nikiforov, et al., 1993) collected previous work from a more dynamical systems theoretic driven perspective with thorough work on properties of the dynamical profile. For our work, we want to lean more towards results in (Lai and Shan, 1999).

To formalise, recall the Kalman setting in the previous chapter via (2.1). As initially motivated, the main objects of interest are the independent innovations γ_n given via $p(y_n - H_n m_n^f) = p(\gamma_n) = n(\gamma_n; 0, \Sigma_n)$ with $\Sigma_n = H_n P_n^f H_n^T + R_n$. At an unknown yet non-random time η , an additive term u_n emerges in the signal equation, so

$$X_n = A_n X_{n-1} + C_n W_n + \mathbf{1}\{n \geq \eta\} u_n. \tag{3.50}$$

The innovations are still independent, however, they have mean $\mathbb{E}[\gamma_n] = \mathbf{1}\{n \geq \eta\}\mu_n \in \mathbb{R}^p$ after the abrupt jump of the signal process with $\mu_n \neq 0$ for all $n \geq \eta$. Further, the mean sequence $\mathbf{1}\{n \geq \eta\}\mu_n$ is a linear transformations of $\mathbf{1}\{n \geq \eta\}u_n$. Taking the case of a constant additive term $u_n = u \in \mathbb{R}^d$ for all n which is frequently associated with actuator failure in engineering, we can describe the dynamic profile of the change in mean of the innovation sequence via $\mu_n = \rho(n, \eta)u$ for $n \geq \eta$ with recursive evaluation of $\rho(n, \eta)$. Let $n \geq k$, then

$$\begin{aligned}
&\rho(k,k) = 0, \quad S(k,k) = 0, \quad F(k,k) = 0\\
&S(n+1,k) = A_n S(n,k) + \mathbf{1}_{d\times d}\\
&F(n,k) = A_{n-1}F(n-1,k) + K_n \rho(n,k)\\
&\rho(n,k) = H_n[S(n,k) - A_{n-1}F(n-1,k)]
\end{aligned} \tag{3.51}$$

as given in (Lai and Shan, 1999). The equations are adapted from the original set of equations in (A. S. Willsky and Jones, 1974) which were based on the more specific choice

$$X_n = A_n X_{n-1} + C_n W_n + \mathbf{1}\{n = \eta\} u_n, \tag{3.52}$$

thus not needing the identity matrix in the second equation of (3.51).
Again, the idea is in splitting the effect of the additive change from the modelled signal process assuming no change via exploiting linearity. The sequence $S(n, k)$ hereby carries the forward propagation and accumulation of the additive term u_n in the signal process. $F(n, k)$ describes the Kalman filter adapting to the additional additive term over time via the Kalman gain and is closely related to the notion of the relative gap in (Alami et al., 2020) for their R-BOCPD procedure. Lastly,

$\rho(n,k)$ captures the difference between the propagated, accumulated additive term and what the Kalman filter has accounted for thus far mapped to the observation space to grasp its current influence on an observation. It then makes intuitive sense to employ a GLR-CUSUM procedure combining estimation and testing via the stopping time

$$N_\rho = \inf\{n \geq 1 : \max_{1\leq k\leq n} \max_{u\in\mathbb{R}^d} \sum_{i=k}^{n} \log(\frac{\exp(-\frac{1}{2}||\gamma_i - \rho(i,k)\theta||^2_{\Sigma_i^{-1}}}{\exp(-\frac{1}{2}||\gamma_i||^2_{\Sigma_i^{-1}}}) > c\}. \tag{3.53}$$

A key insight we want to take along hereby is that for large temporal distances $n-k+1$, the procedure can be approximated via replacing $\rho(i,k)$ with its steady state version ρ_∞ derived in (Lai and Shan, 1999) under usual assumptions for steady state convergence of the system components and asymptotic stability of the Kalman filter. Thus under usual assumption we can also take the post-change innovation mean to be steady up to the dynamic profile.

3.5.1. Constructing Inference Schemes for Filtering Under Change

While the derived set of equations offer a versatile tool and are interesting for a large variety of other approaches, we do not want to rely on them. With a keen eye on robustness to observation noise mis-specification via the derived diffusion score matching approach in the previous chapter, the choice of the IMQ-kernel, but also any other non-linear kernel, makes a split such as in (3.51) highly difficult if not impossible to evaluate. This shows explicitly in the divergence term of the mean update in (2.58). Additionally, we aim for strategies that employ change point detection while allowing for adjusted inference on the latent signal accounting for additional uncertainty induced by potential change mis-specification of the signal noise, i.e. additive change.

Looking at the long list of examples from application in the literature, e.g. in (Basseville, Nikiforov, et al., 1993) and more recent in (A. Tartakovsky et al., 2014), the most frequent approach for all the relevant reasons derived in the previous sections is the χ^2-GLR CUSUM procedure in (3.44) simply monitoring the innovation sequence for non-zero mean. In short, under very little assumptions and with the sole aim to detect abrupt jumps in the signal process, so substantial deviations from zero mean of the innovations, monitoring the empirical innovation mean after invariant transformations $\tilde{\gamma}_i = \Sigma_i^{-\frac{1}{2}}\gamma_i$ accounting for the individual time-varying covariance matrices provides most desired properties in reliability and feasibility:

$$\begin{aligned} N^{\bar{\gamma}}_{m(\alpha_0)} &= \inf\{n \geq 1 : \max_{(n-m(\alpha))\leq k\leq n} \frac{n-k+1}{2}||\bar{\gamma}_{k:n}||_2^2 > c\} \\ &= \inf\{n \geq 1 : \max_{(n-m(\alpha))\leq k\leq n} \exp(-c)\exp(\frac{1}{2(n-k+1)^2}||\sum_{i=k}^{n} \Sigma_i^{-\frac{1}{2}}\gamma_i||_2^2) > 1\} \end{aligned} \tag{3.54}$$

with $\bar{\gamma}_{k:n} = \frac{1}{n-k+1}\sum_{i=k}^{n}\Sigma_i^{-\frac{1}{2}}\gamma_i$ and appropriately chosen $c \sim |\log(\alpha_0)|$, i.e. via Monte Carlo simulation of the pre-change regime, and $m(\alpha_0)$ such that (3.11) holds. The second line again enables transfer to the CR-BOCPD intuition.

Accounting for Change During Filtering

Similar to the pioneering work in (A. S. Willsky and Jones, 1974), our aim reaches beyond solely detecting change but aiming for an adaption to the filtering routine. The overarching goal is in accounting for mis-specification of heavy tailedness in signal noise alongside Kalman filtering. Willsky and Jones based their approach on a direct estimate or covariance incrementation. Exploiting the workflow of the GLR procedure in (3.53) based around estimating change location and jump magnitude to determine a post-change density, the obtained parameter $\hat{\theta}_{\hat{\eta}}$ is then utilized to restart the filtering procedure at time $\hat{\eta}$ with explicitly including the additive term $\hat{\theta}$ in the signal equation at time $\hat{\eta}$. Their approach in covariance incrementation from a more practical side is in artificially increasing the signal covariance matrix at $\hat{\eta}$ for the restart, thus allowing the Kalman filter to better adapt to the detected jump on its own. Two major problems arise for either approach given large detection delay. Restarting the filter too far back may be computationally expensive and is especially in real-time applications likely not feasible. More important however, until we can make the decision that a signal jump occurred, we act under complete ignorance of potential jumps for estimation and forecasting of the signal with their approach.

We propose utilizing scenario weighted sums such as in (Adams and MacKay, 2007) with the CR-BOCPD approach inspired by (Alami et al., 2020) to incorporate signal estimation and forecasting under uncertainty about abrupt jumps alongside decision making. The key is in developing the equation in (3.15) with the notion of the scenario weights in (3.21) and (3.22) for exact Bayesian computation or the simplified scenario weights in (3.25) and (3.26) to exploit powerful properties of CUSUM strategies. The central equations for the context of Gaussian mixture model Kalman filtering are then given by

$$\begin{aligned} p_\kappa(x_n|y_{1:(n-1)}) &= \sum_{s=0}^{n} \tilde{\kappa}^f_{s,n} p(x_n|y_{1:(n-1)}, s) \\ p_\kappa(x_n|y_{1:n}) &= \sum_{s=0}^{n} \tilde{\kappa}_{s,n} p(x_n|y_{1:n}, s) \end{aligned} \tag{3.55}$$

with $\tilde{\kappa}_{k,n} = \frac{\kappa_{k,n}}{\sum_{s=r}^{n} \kappa_{s,n}}$ the normalized scenario weights for probabilistic change points given knowledge of the change point prior and

$$
\begin{aligned}
p_{\kappa}(x_n|y_{1:(n-1)}) &= \sum_{s=r}^{n} \tilde{\kappa}^{f}_{r,s,n} p(x_n|y_{1:(n-1)}, s) \\
p_{\kappa}(x_n|y_{1:n}) &= \sum_{s=r}^{n} \tilde{\kappa}_{r,s,n} p(x_n|y_{1:n}, s)
\end{aligned}
\tag{3.56}
$$

with $\tilde{\kappa}_{r,k,n} = \frac{\kappa_{k,n}}{\sum_{s=r}^{n} \kappa_{r,s,n}}$ the normalized scenario weight for deterministic yet unknown change points or probabilistic change points and no access to a change point prior. Recall, $\tilde{\kappa}^{f}_{k,n}$ and $\tilde{\kappa}^{f}_{r,k,n}$ are hereby the respective scenario forecast weights. While $\tilde{\kappa}^{f}_{k,n}$ is exact given the assumptions, needs to be treated with sufficient care $\tilde{\kappa}^{f}_{r,k,n}$. Note, the idea of Gaussian mixture model filters is not new (see (Anderson and Moore, 2012) and (Reich and Cotter, 2015) for details), however, deriving their weights terms via change point detection is.

The idea is in obtaining a Gaussian mixture model of individual Kalman filters for each currently considered scenario weighted via a notion of scenario plausibility. Other then in the regular model predictive approach of BOCPD, our interest lies explicitly in estimating and forecasting the latent model, the signal process, and not in uncertainty quantification of the next observation. As the concept of Gaussian mixture model is well established, we can then make use of their known properties such as moments and confidence regions.

Specifying Scenario Initialization and Scenario Weight Computation

It is a difficult task and much beyond the scope of this work to define a novel notion of optimality combining estimation and forecasting as well as detection delay for shared inference, however, for now we want to argue for proficiency of the proposed approaches via optimality in their respective aspects filtering and detection. The proposed scenario weighted Kalman filter needs specifying two central aspects. On the one hand, scenario initialization needs defining, so how a new scenario $p(x_k|y_{1:k-1}, k)$ is constructed. On the other hand, the computation of the scenario weights needs deciding given assumptions. For probabilistic change points with access to the change point prior we can employ the Bayesian computation of scenario forecasts and posteriors. For deterministic change points or no access to a reliable change point prior we need to consider additional assumptions for deciding between the conditional CR-BOCPD procedure or adaptation of χ^2-based procedures to the CR-BOCPD framework.

To formalise, we want to pick up on the usual signal process adaptation similar to the original problem statement in (A. S. Willsky and Jones, 1974). Let

$$
X_n = A_n X_{n-1} + \Gamma_n W_n + \delta_{\eta,n} u_n \tag{3.57}
$$

be the signal process with unknown additive term $u_n \neq 0$. For probabilistic change points, we want to assume $\delta_{\eta,n} = \pi_n$ with π_n non-negative discrete random variable as previously introduced. As before, we want to assume a geometrical distribution on the relative frequency of change points from prior knowledge if not specified otherwise. Deterministic change points in abrupt jumps of the signal process are obtained via setting $\delta_{\eta,k} = \mathbf{1}\{n = \eta\}$ with η unknown. Either may be adapted appropriately for multiple instances of change.

Starting with the question on scenario initialization, we want to propose three courses of action. The first is based on the idea in (Adams and MacKay, 2007) and forgetting about past, supposedly pre-change observations to allow for adaptation only to supposed post-change observations. Looking at $p(x_k|y_{1:k-1}, k)$, we can reduce it to $p(x_k|k)$ and simply propagate our initial prior $p(x_0)$ via the signal process to point in time $k > 0$. Depending on the system specifications however, this may lead to the covariance term blowing up way beyond what is desired. Instead, we want to suggest that only the reduction in covariance from past observations is targeted and it either replaced or inflated. The idea hereby lies in that the mean estimate is reliable up to the change point and only then additional uncertainty is induced. We can therefore maintain part of the information before a change up to an introduction of additional uncertainty. For replacing, the initial covariance matrix P_0 and its forecast may be suitable candidates. The idea of inflating covariance resembles that of covariance incrementation in (A. S. Willsky and Jones, 1974) and was also picked up in (Lai and Shan, 1999) via employing a scalar factor $\xi > 1$ to obtain $p(x_k|y_{1:k-1}, k) = n(x_k; m_k^f, \xi^2 P_k^f)$. A major difficulty regarding this second approach lies in tuning of either the choice of replacement or scaling factor and is somewhat arbitrary. Regardless, given the theoretical results, either will result in first-order asymptotically optimal detection delay with the CR-BOCPD procedure. The third approach is arguably the most useful yet also the most restrictive in assumption. Picking up on the idea in (3.49), we may assume a specific Gaussian distribution of the additive term in $u_n \sim \mathcal{N}(\bar{u}_n, U_n),\ n \geq 1$. The new scenario can then be expressed via $p(x_k|y_{1:k-1}, k) = n(x_k; m_k^f + \bar{u}_k, P_k^f + U_n)$. Whenever we have sufficient information about the nature of the additive term via large mean $\bar{u}_n$ and comparatively small covariance matrix U_n, e.g. via previous experiments or application context, this last case is most useful. Otherwise, so for small mean and relatively large covariance, it recovers the previous case of inflating the covariance matrix.

Considering the nature of the Kalman filter and, moreover, Bayesian learning, we mainly exploit its properties in balancing uncertainties. We may therefore reduce the sketched cases to the simple question whether we have specific knowledge about the nature of the change we can employ for scenario initialization, i.e. in approach three, or just increasing uncertainty to artificially allow for better adaptation to a potential post-change regime, so similar to (A. S. Willsky and Jones, 1974) but in an online fashion.

With the idea of parallel implementation of several Kalman filter algorithms for the respective scenarios, we need adapting notation. Let hereby $p(x_n|y_{1:(n-1)}, k) = n(x_n; m_{k,n}^f, P_{k,n}^f)$ be the forecast distribution of the Kalman filter at time n initialized at k and $p(x_n|y_{1:n}, k) = n(x_n; m_{k,n}^a, P_{k,n}^a)$ the respective analysis distribution. Accordingly, the respective scenario Kalman filters are therefore started via $p(x_k|y_{1:k-1}, k) = n(x_k; m_{k,k}^f, P_{k,k}^f)$. For the first course of action, set $m_{k,k}^f = \prod_{i=1}^k A_i m_0$ and $P_{k,k}^f = \psi_k(\psi_{k-1}(\cdots \psi_1(P_0)))$ with $\psi_i(P) = A_i P A_i^T + Q_i$ and m_0, P_0 parameters of the initial prior $p(x_0|0) = n(x_0; m_0, P_0)$. The second approach changes $P_{k,k}^f$ either by replacement or scaling while maintaining $m^f_{k,k} = m^f_{0,k}$. The third action initializes via parameters $m^f_{k,k} = m^f_{0,k} + \bar{u}_k$ and $P^f_{k,k} = P^f_{0,k} + U_n$. The parameters $m^f_{0,k}$ and $P^f_{0,k}$ are hereby the parameters of the forecast with the initial prior distribution at time k and represent the parameters of the no-change scenario, so $p(x_k|y_{1:(k-1)})$. For the restart procedure, we want to replace $m^f_{0,k}$ and $P^f_{0,k}$ with the parameters of the current baseline scenario, so $m^f_{r,k}$ and $P^f_{r,k}$ with initial $r = 0$.

Looking at computation of the scenario weights, we can be a lot more specific on assumptions. Starting with assuming reliable knowledge about a change point prior as introduced in (Adams and MacKay, 2007), the equations in (3.21) and (3.22) provide required scenario weight posteriors and forecasts. We do neither have a decision rule nor investigations on usual measures of reliability available, however, given the exact Bayesian computation the approach is easily argued to be proficient up to error in the priors and variability in an appropriate model predictive. This is further supported by strong experimental results such as in (Van den Burg and Williams, 2020).

Whenever we cannot assume such a change point prior, we want to utilize computation of the scenario weights via the conditional CR-BOCPD procedure in (3.25) and (3.26) for scenario posteriors and forecasts. For a given false alarm rate α_0, we obtain first-order asymptotic optimality for both probabilistic and deterministic change points via results in (Lai, 1998), also for window-limited adaptations with widow-size $m(\alpha_0)$ chosen such that (3.11) holds. Recall, the key constraint hereby lies in that we need to provide conditional pre- and post-change density functions, i.e. via marginalization of the unknown parameter subject to change, the signal, to obtain $p(y_n|y_{1:(n-1)}, k)$.

Next to the conditional CR-BOCPD, we may choose to use the χ^2-CUSUM or GLR procedure for detection given unreliable pre- and post-change density functions although this approach needs plenty more work. Hereby we utilize the transfer in (3.27) in the opposite direction via splitting the LLR. Exploiting the factorization in Cox's theorem (3.34) and retracing the steps in (3.27), we can assign $\kappa_{r,r,n} = 1$ and $\kappa_{r,k,n} = \beta_{r,k,n} \exp(\frac{(n-k+1)b^2}{2}) G(\frac{p}{2}, \frac{(n-k+1)^2 b^2 ||\bar{\gamma}_{k,n}||_2^2}{4})$ with $G(r, s)$ the generalized hypergeometric function as introduced previously, p dimension of the observation space, b^2 a minimal squared euclidean distance of the post-change mean from 0 pre-change mean of the innovations and $\bar{\gamma}_{k,n} = \frac{1}{n-k+1} \sum_{s=k}^{n} \Sigma_s^{-\frac{1}{2}} \gamma_s$, the mean over standardized innovation terms from k to n.

Similarly, when we do not want assuming such a minimal distance b^2, we may exploit the χ^2-GLR procedure instead via changing $\kappa_{r,k,n} = \beta_{r,k,n} \exp(\frac{(n-k+1)}{2} ||\bar{\gamma}_{k,n}||_2^2)$ and maintaining $\kappa_{r,r,n} = 1$. Recall the derivation in (3.43), (3.46) and (3.54). These two approaches need plenty more work, yet may offer interesting perspectives driven a lot more form statistical theory. As they do not need knowledge of a model predictive and instead focus on detecting departure from 0 mean of the innovation sequence, they offer a reliable and highly proficient way to compute scenario weights resilient to tuning choices of scenario initialization. In the scope of this work, we will only briefly investigate them in experiments for providing effective weights for

Gaussian mixture model filters for non-linear approximation.

Merging Insights into Results

Given choice of scenario initialization and computation of scenario weights from assumptions, implementation via the regular Kalman scheme of forecast and analysis then only needs aggregation at the respective time steps to obtain the corresponding Gaussian mixture for forecast and analysis via (3.55) for probabilistic change points and available change point prior and (3.56) otherwise. A general algorithm for the Change Scenario GMM- Kalman filter is given by (2) without specifications. Two more concrete algorithms, the BOCPD GMM-Kalman filter (3) for probabilistic change points and the CR-BOCPD GMM-Kalman filter (4) for deterministic change points, are provided in the appendix.

At this point, we want to circle back to the initial motivation of this chapter and the broader scope of the work at hand. Regarding change point detection in dynamical systems, the introduced χ^2-based procedures provide a state-of-the-art with the χ^2-CUSUM procedure providing desirable detection properties for a known minimal deviation of the post-change innovation mean sequence. Furthermore, the χ^2-GLR procedure as given in (3.54) works without specifying such a minimal deviation for detecting departure from 0 mean of the innovation sequence while maintaining desired properties. Both have their main challenge in efficient online implementation resulting from need for computing the sufficient statistics, even for window-limited adaptations. The recursive detection procedures based on the closed χ^2-SPRT introduced in (I. V. Nikiforov, 1999) and (I. V. Nikiforov, 2001) overcome this issue to some extent. Yet, the central aspect of this work is in inference under presence of change in additive terms of the signal process and the induced uncertainty as well as doing so under robustness to observation noise mis-specification. The mean of standardized innovations as sufficient statistic employed in the χ^2-schemes is highly susceptible to outliers and realizations of mis-specification in heavy-tailedness. While adaptation for more robust sufficient statistics in Huber's M-estimators (see (Huber, 2004) for details) may provide an interesting route for future research, it is not the direction we are interested in.

The BOCPD approach as introduced in (Adams and MacKay, 2007) as well as the robust adaptation central to this work in (Altamirano et al., 2023b) provided new and promising insights. Given their assumptions, the Kalman filter equations with their inherent conjugacy provide somewhat of a special case of the model predictive with an additional dynamical system of the latent parameter. The type of change in BOCPD shares a desirable resemblance to the concept of Huber contamination as utilized to resemble observation noise mis-specification in heavy-tailedness in the previous chapter. In ideal circumstances, we assume knowledge of the change point prior as well as a Gaussian profile of the additive terms. We can then very accurately specify computation of scenario weights and scenario initialization to ob-

Algorithm 2 The Scenario Kalman Filter

```
Input:
  • Kalman filter requirements (initial condition $p(x_0|k = 0) \sim n(x_0; m_0, P_0)$, signal model, observation model)
  • scenario initial condition $p(y_k|y_{1:(k-1)}, k)$
  • scenario weight computation and detection criteria
Output:
  • scenario/run-length posterior or last detected change point $r$
  • aggregated signal forecast $p_\kappa(x_n|y_{1:(n-1)})$ at time $n$ and
  • aggregated signal state $p_\kappa(x_n|y_{1:n})$ at time $n$
for n=1,2,... do
  initialize new scenario forecast for $k := n$ and forecast step for all previous scenarios
  initialize new scenario weight forecast for $k := n$, forecast scenario all previous scenario weights via (3.22) or (3.26) and normalize
  aggregate signal forecast via (3.55) or (3.56)
  receive observation: $y_n$
  initialize new scenario weight for $k := n$, update all previous scenario weights via (3.21) or (3.25) and normalize
  if given, check restart/detection rule
  analysis step for all scenarios
  aggregate signal state estimate via (3.55) or (3.56)
end for
```

tain reliable estimation and forecasts via scenario aggregation. Under much less ideal circumstances and likely muc more common in practice, change points may either be non-random or random but we can not assume a sufficiently reliable change point prior. Further, we have little to no knowledge about structure or profile of the additive terms. The developed conditional CR-BOCPD procedure with restart rule equivalent to the detection rule in CUSUM strategies provides desirable properties in reliability, also for the window-limited modification. We can aggregate scenario forecasts and estimations and make use of the robust inference in the previous chapter merging several desired properties.

The two sketched situations are highly interesting as they describe two ends of the spectrum on knowledge relevant for the challenge at hand. We can account for probabilistic or deterministic change as well as very specific Gaussian additive terms up to general additive terms with little to no knowledge available.

3.6. Experiments: Latent Change and Non-Linear Simulations

We want to stay with the set-up from the previous chapter, however, reduced to only the target tracking and Lorenz-63 examples. For the CR-BOCPD GMM-Kalman filter we choose a constant threshold via Wald's approximation for open-end SPRTs. Following (Wald, 1992), we choose

$$c \approx -\log(\alpha_0), \tag{3.58}$$

wit tightness of the approximation is hereby governed by overshoot which can generally be assumed to be reasonable small and providing the desired scaling for asymptotic properties. Again, more precise results may be achieved via tuning with Monte Carlo Simulations of the ARL under the null hypothesis. Additionally, we want to employ a window-size $m(\alpha) = \log(\alpha_0)^2$ which satisfies the requirements in (3.11). Choosing a rate of false alarm of $\alpha_0 = 0.05$ then results in tuning parameter $\beta_{r,k,n} = \exp[\log(\alpha_0)] \approx 0.05$ and window size $m \approx 9$, so $k \in \{n-m, n-m+1, \dots, n\}$, with the baseline as additional scenario at r. Similarly, we want to prune the BOCPD GMM-Kalman filter to only consider the 10 scenarios/run-lengths with the highest posterior probability plus the newly initialized scenario.

The experiments are only evaluated superficially regarding inference of the underlying dynamical system. There is no thorough empirical analysis regarding frequent measures of accuracy and error of change point detection via F1-score or a Hausdorff metric in relation to signal-to-noise ratio. Again, the central reason hereby lies in that we aim for reliable inference in filtering, i.e. signal estimation in the analysis step, although change points as realizations of heavy-tailedness of the signal noise may be present.

All simulations were done in R version 4.2.2.

3.6.1. Experiment A: Target Tracking Latent Gaussian Change

The general setup is the same as in the target tracking example of the previous chapter apart from the contamination of the observations. We choose the popular target tracking task with signal space $\mathcal{X} = \mathbb{R}^4$ containing x-position, x-velocity, y-position as well as y-velocity, and observation space $\mathcal{Y} = \mathbb{R}^2$ containing the measured x-position and y-position. We design the signal model discrete and focus for this experiment on probabilistic change via additional signal noise terms with large covariance at random times in the unobserved velocity dimensions. Given the usual Kalman setting with random additive terms, we have

$$\begin{aligned} X_{n+1} &= A_n X_n + Q_n^{\frac{1}{2}} W_n + \pi_n u_n \\ Y_n &= H_n X_n + R_n^{\frac{1}{2}} V_n \end{aligned} \tag{3.59}$$

with W_n and V_n standard Gaussian noise in the respective dimensions, $\pi_n \sim_{iid} \text{Ber}(\varphi = 0.05)$ and $u_n \sim_{iid} \mathcal{N}(0, U)$ with $U = 100\begin{pmatrix} 0&0&0&0 \\ 0&1&0&0 \\ 0&0&0&0 \\ 0&0&0&1 \end{pmatrix}$. As usual for these kind of models, yet much more simplified, we choose $A_n = \begin{pmatrix} 1&1&0&0 \\ 0&1&0&0 \\ 0&0&1&1 \\ 0&0&0&1 \end{pmatrix}$, $H_n = \begin{pmatrix} 1&0&0&0 \\ 0&0&1&0 \end{pmatrix}$ and signal noise covariance $Q_n = 0.1 \cdot \begin{pmatrix} 1&0.5&0.5&0 \\ 0.5&1&0&0 \\ 0.5&0&1&0.1 \\ 0&0&0.1&1 \end{pmatrix}$. The initial signal is chosen as $X_0 = (0, 1, 0, 0)^T$. The observations are generated via additional standard Gaussian observation noise V_n and$R_n = \begin{pmatrix} 0.1&0.01 \\ 0.01&0.1 \end{pmatrix}$. We simulate X_n for n in $[0, 2]$ with step-size $\mathrm{d}n = 0.01$ resulting in 200 positions with $n_{out} = 1$, so 200 observations. Next to the regular Kalman filter, we implement a BOCPD GMM-Kalman filter and a CR-BOCPD GMM-Kalman filter with the corresponding scenario initialization assuming knowledge of the term u_n and geometric change point prior matching π_n or threshold as previously stated. Our main interest lies in the analysis mean of all three methods for the unobserved velocity dimensions. As can be seen in figure (3.1), both proposed methods reliable detect and adapt to instances of change with the regular Kalman filter lacking behind for prolonged time periods. While the BOCPD based scheme hereby tends to be a little faster in adaptation to a new post-change regime, it is also much more uncertain and frequently needs time to stabilize after a jump. The novel CR-BOCPD scheme tends to be a little slower, however the difference in delay is almost negligible. Yet, after detection of an instances of change, the restart mechanism throwing away past scenarios helps stabilizing with an additional correction after detection showing in a second smaller jump. For the given trajectory in figure (3.1), changes happen at time points $\eta = (6, 20, 46, 64, 87, 91, 102, 129, 139, 149, 155, 188)$. The restart of the CR-BOCPD procedures happen at $\hat{\eta} = (0, 7, 8, 21, 22, 47, 48, 65, \ldots, 150, 151, 156, 157, 189, 190)$. It is hereby common practice in change point literature to combine change points in close proximity into a single change point, however, for our sake it is a very interesting observation that the CR-BOCPD scheme shows this behavior of adjusting the

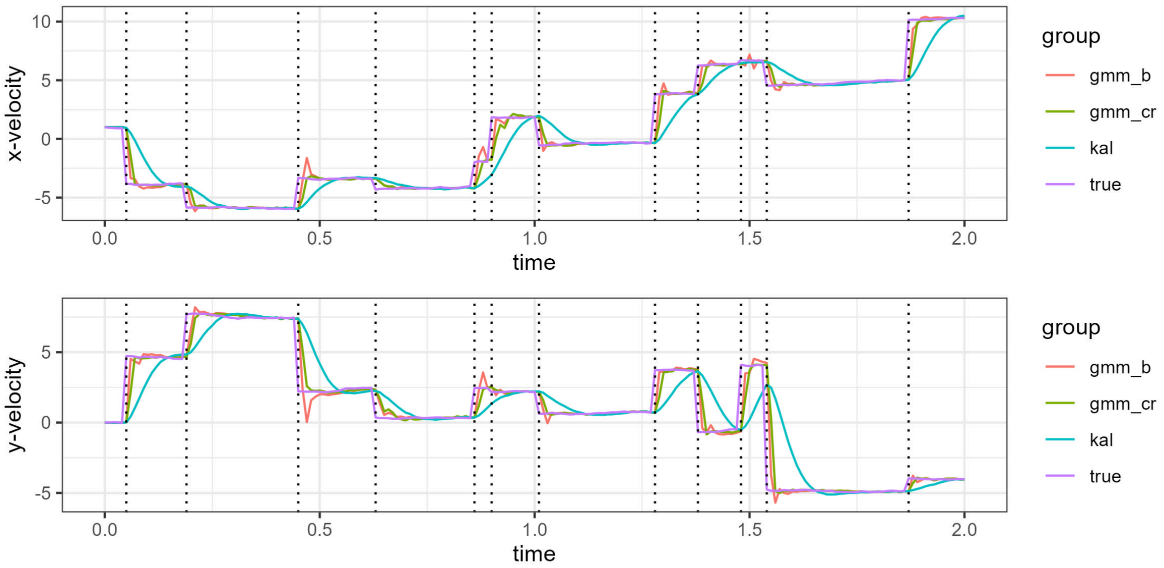


Figure 3.1.: True trajectory in x and y-velocity (true), analysis mean velocities for the regular KF (kal), the BOCPD GMM-KF (gmm_b) and CR-BOCPD GMM-KF (gmm_cr). Dotted lines indicate instances of change.

analysis estimates in two steps. For the BOCPD scheme, we can hereby observe a switching back-and-forth behavior of the MAP estimator on the scenario posterior, i.e. the MAP-sequence $(6, 7, 8, 7)$ while adjusting to the change point at $\eta = 6$.
Taking the results on the analysis mean of the latent velocities, we can observe corresponding impact on estimating the position. Figure (3.2) clearly shows the impact of the slow adjustment of the regular Kalman filter to the true position in line with the slow adjustment in the latent velocities. The reason hereby is likely straight forward in overconfidence of estimates via small analysis covariance. This is neither surprising nor shocking as the regular Kalman filter is not designed to account for the type of behavior produced by the jumps. The BOCPD and CR-BOCPD schemes on the other hand specifically account for this behavior via their covariance increments, however, they nicely showcase the price in large analysis covariance and the resulting increased reliance on observations with almost over-fitting like behavior for short time periods after detected instances of change.
The experiments support the theory in construction and derived properties of the schemes providing good estimates of both observed and latent signal variables. While the time-averaged MSE of the mean of BOCPD scheme tends to be slightly lower than the CR-BOCPD mean, both are a power of 10 smaller than the regular Kalman filter mean. While there is only a single trajectory presented here, all insights also applied for repeated experiments.

Additional graphs of the combined and individual signal dimensions as well as differences from the true trajectory are provided via figure (A.21) to (A.30) in the appendix.

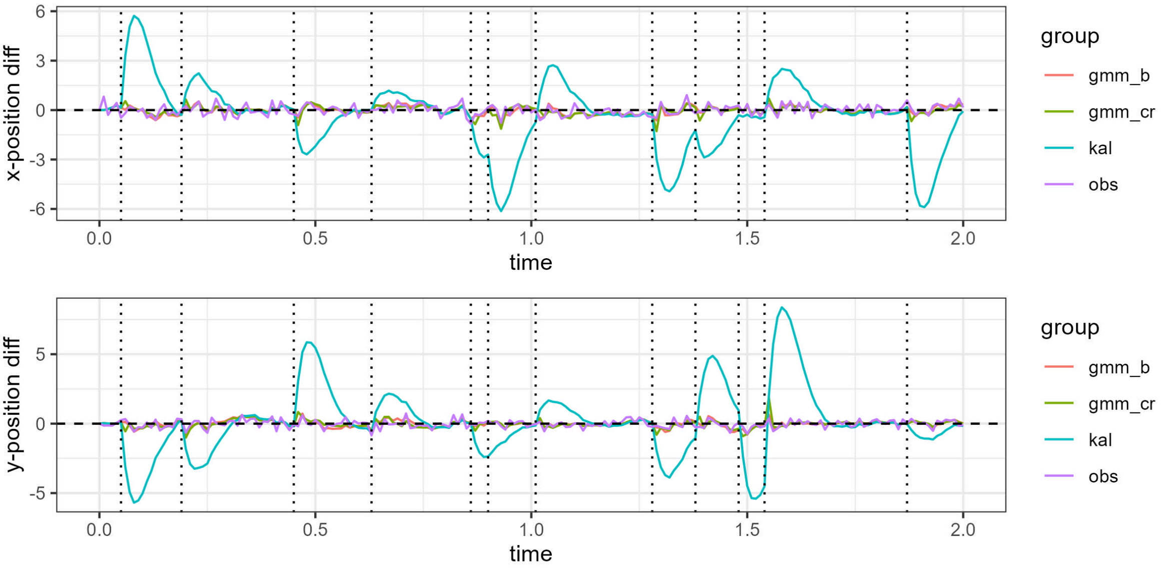


Figure 3.2.: Difference of analysis means and true x and y-position for the observations (obs), the regular KF (kal), the BOCPD GMM-KF (gmm_b) and CR-BOCPD GMM-KF (gmm_cr). Dotted lines indicate instances of change.

3.6.2. Experiment B: The Lorenz-63 Test

While not explicitly discussed thus far, the idea of re-adjusting signal estimation to jumps via covariance increments provides an inherent potential to account for non-linearity and chaotic behavior. We can hereby substitute separate Kalman filters for each scenario with ensemble Kalman filters with perturbed observations. For the CR-BOCPD scheme, new scenarios are then initialized by drawing a new ensemble from the current baseline with appropriate covariance adjustment. In practice this leads to a GMM-EnKF or simply an EnKF with multiple sub-ensemble with their corresponding scenario weights. For the window-limited version, the oldest ensemble apart from the baseline is replaced with a new ensemble before the forecast step.
As in the previous chapter, the Lorenz-63 model provides a very suitable candidate for investigation in non-linear and chaotic properties. With the diffusion score matching Kalman filter showing highly interesting behavior regarding non-linearity via the adaptive, observation dependent Kalman gain, we want to include it for comparison to the CR-BOCPD GMM-ENKF schemes.

Again, we take the simulation mainly from (Reich and Cotter, 2015). Recall, let z be the signal variable, we then have the vector field f given by

$$f(z) := \begin{pmatrix} \sigma(z_2 - z_1) \\ z_1(\rho - z_3) - z_2 \\ z_1 z_2 - \beta z_3 \end{pmatrix} \tag{3.60}$$

with parameters $\sigma = 10$, $\rho = 28$ and $\beta = \frac{8}{3}$. We chose a step-size of $\delta t = 0.001$ and the initial value $z_0 = (-0.587, -0563, 16.870)^T$. To implement a forward Euler scheme as numerical approximation, we include a non-autonomous forcing term g_n that essentially comes down to a tent map iteration. We set $a = (\delta t)^{-\frac{1}{2}}$ and the initial forcing term as $g_0 = (a(2^{-\frac{1}{2}} - \frac{1}{2}), a(3^{-\frac{1}{2}} - \frac{1}{2}), a(5^{-\frac{1}{2}} - \frac{1}{2})$ with the entry-wise recursive definition

$$g_{n+1,i} = \begin{cases} 1.99999 g_{n,i} + \frac{a}{2} & \text{if } g_{n,i} < 0 \\ -1.99999 g_{n,i} + \frac{a}{2} & \text{otherwise.} \end{cases} \tag{3.61}$$

The signal is the propagated via

$$z_{n+1} = z_n + \delta t(f(z_n) + g_n) \tag{3.62}$$

over a window $[0, 10]$ for investigating performance without mis-specification. For the $\mathcal{D}_w$-EnKF we use the default choice of tuning parameter in $\beta = 1$. Observations were generated at $t_{out} = 100$ via

$$Y_n = H_n z_n + \sqrt{2} V_n \tag{3.63}$$

with V_n standard Gaussian noise and a stricter observation map $H_n = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$ only allowing noisy observation of the first component. We chose an ensemble size of $M = 5$ to mimic application for expansive forward models.

We implement two CR-BOCPD schemes with the first as in the previous experiment based on the CUSUM procedure for conditional probabilities and the second based on the χ^2-GLR procedure. Both threshold and window-size are as previously described via Wald's approximation and the poly-logarithm. At initialization of a new scenario at time $k > r$, the empirical covariance $\hat{P}_{r,k}$ of the current baseline scenario inflated by a factor $\lambda = 5$ and the empirical mean $\hat{m}_{r,k}$ and are taken to draw a new ensemble $\{x_{k,k}^{a,(l)}\}_{l \in \{1,2,\dots,M\}}$, so $X_{k,k}^l \sim_{iid} \mathcal{N}(m_{k,k}, P_{k,k})$ with $m_{k,k} = \hat{m}_{r,k}$ and $P_{k,k} = \lambda \hat{P}_{r,k}$. This new ensemble is then propagated for the next forecast, but not considered for the previous analysis aggregation.

As can be seen in figure (3.3), all proposed schemes provide valuable information about the true trajectory from noisy observations of the first component. Again, the good performance of the EnKF with diffusion score matching analysis step somewhat stands out especially in competing with the CR-BOCPD GMM-EnKFs with larger ensemble size via the multiple scenarios. Both of the GMM-EnKFs show the expected behavior in re-adjusting via the covariance increments and accounting to non-linerity this way. The vertical dotted lines figure (3.3) show precisely at which time points the CR-BOCPD GMM-EnKF based on the conditional evidence has its restart rule activated, so its previous baseline model overtaken by a new one with incremented signal covariance. Accordingly, it is these points where the regular EnKF is potentially thrown of from the true signal via interaction of non-linearity and observation error.

Implementing abrupt jumps in the Lorenz-63 system is difficult as simply adding

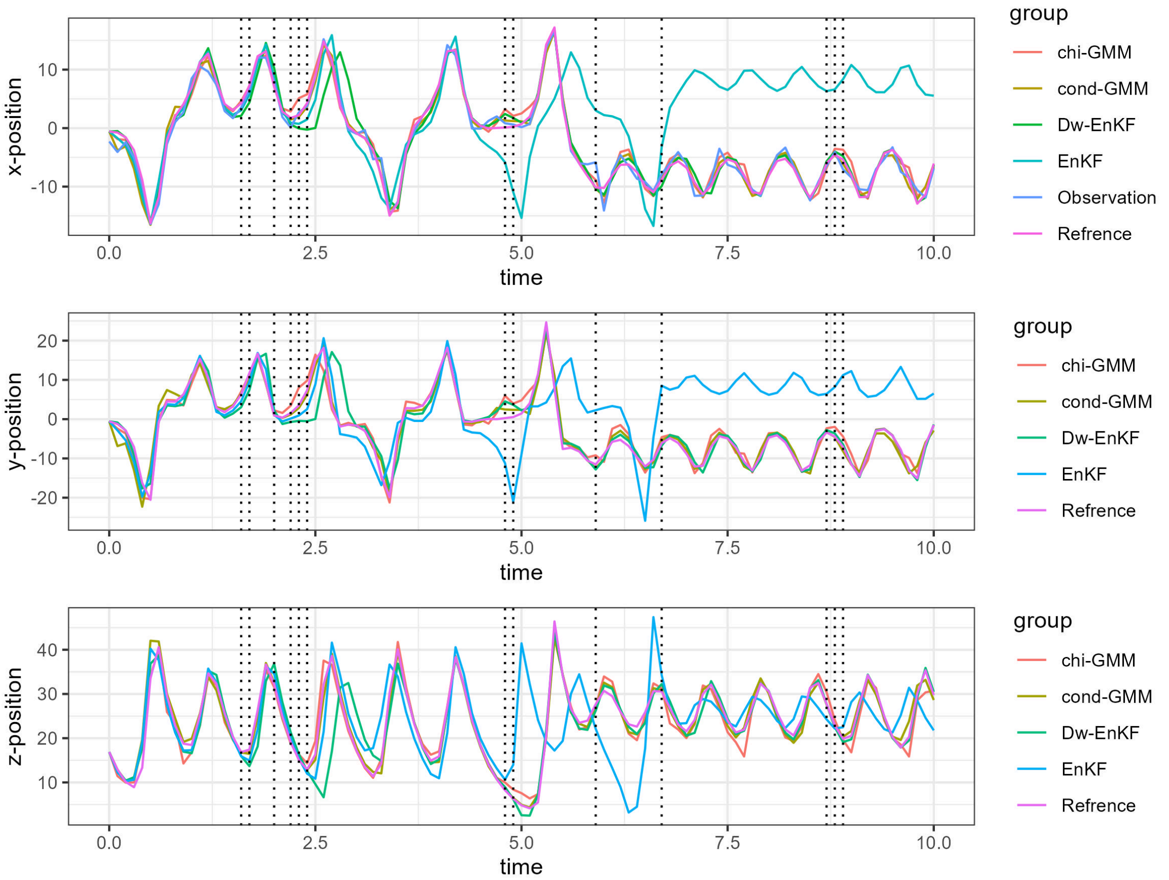


Figure 3.3.: Side-by-side graph comparison of the simulated signal (Refrence), generated observations (Observation), the analysis mean of the regular EnKF (EnKF), the default $\mathcal{D}_w$-EnKF (Dw-EnKF), the CR-BOCPD GMM-EnKF with conditional probabilities (cond-GMM) and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion (chi-GMM). Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

large random perturbations to it might degenerate the system and make it insignificant in some way. Instead we want to use deterministic resets of the system at random times to still include jumps of the systems without loosing the desired non-linear behaviour.
We start by simulating a reference trajectory as previously described for the time interval $[0, 40]$. Next we draw five time points over the whole number of steps of the simulation with a binomial distribution with probability $p = 0.15$. We combine the resulting sub-sequences of the reference trajectory up to the sampled time points back-to-back to obtain a new trajectory that resets after these random time points to the initial conditions. Again, we only observe the first component and add noise to obtain the observations as described.

We observe the expected result in figure (3.4) in that the regular EnKF is thrown off shortly after the first jump back to the initial conditions. Moreover, the same applies to the $\mathcal{D}_w$-EnKF, however, while the regular the EnKF stays out of balance for major parts of the trajectories, the EnKF with diffusion score matching analy-

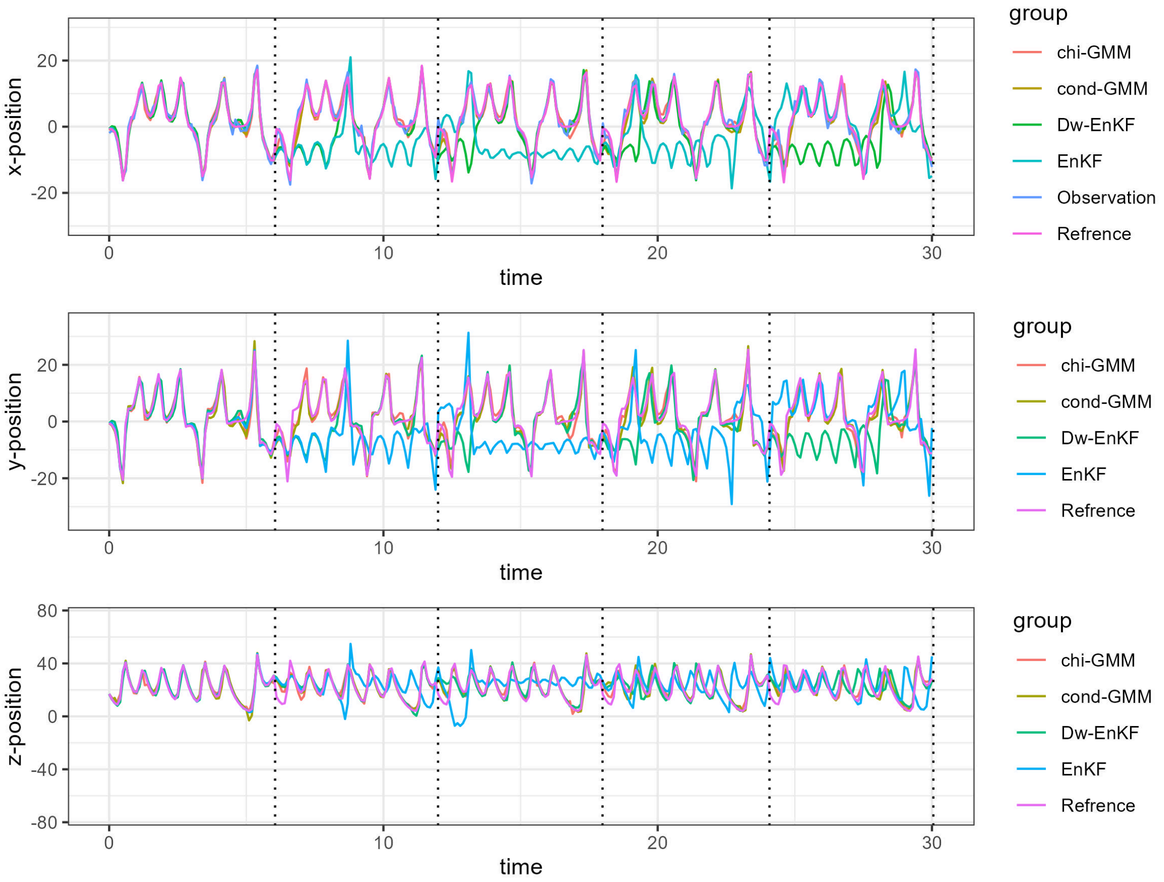


Figure 3.4.: Side-by-side graph comparison of the simulated signal with jumps (Refrence), generated observations (Observations), the analysis mean of the regular EnKF (EnKF), the default $\mathcal{D}_w$-EnKF (Dw-EnKF), the CR-BOCPD GMM-EnKF with conditional probabilities (cond-GMM) and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion (chi-GMM). Dotted lines indicate instances of change via a reset of the system.

sis step manages to always catch on again to the true trajectories after short time periods. Both CR-BOCPD GMM-EnKF show the desire performance in reliably tracking the true trajectories also for the latent variables.

Again, the results presented are only examples with the described behavior observed for a variety of different set-ups. The time averaged MSE of the $\mathcal{D}_w$-EnKF is generally less than half the MSE of the regular EnKF with the proposed CR-BOCPD GMM-EnKFs again improving by an order of 10 lower. Still, conclusions should be drawn with great care. The presented trajectory emphasize the ability of the $\mathcal{D}_w$-EnKF with perturbed observations to recover contrary to the regular EnKF. The two ensemble filters designed in the scope of this chapter perform according to their design, however, given the additional cost in the multiple ensemble for each scenario, the result needs to be expected.
Additional graphs for the individual dimensions are provided in figure (A.31) to (A.36) in the appendix

3.7. Discussion and Conclusion

The presented results are most of all proof-of-concept with main contribution in connecting the intuitions in Bayesian online change point detection via conditional model predictives obtained from cheap conjugacy and marginals, or evidence, to Lai's results on reliability of CUSUM strategies for conditional pre- and post-change density functions and exploiting the resulting synergy in the context of Bayesian filtering. We hereby want to explicitly point out again, that the conditional model predictive or evidence $p(y_n|y_{1:(n-1)})$ is generally taken to be the best indicator of performance of a filtering system in practice with no knowledge of a ground truth available. It is therefore very intuitive for the proposed CR-BOCPD schemes to utilize this quantity.

Regarding the central aim of adjusting estimation for signal noise mis-specification in heavy tailedness, the proposed approaches provided good performance with desired improvement. However, other than with the previous chapter this needs much more careful evaluating in the context of additional cost, i.e. the additional filters for respective scenario. While this cost can be controlled to some extent with the via window-size without loss in reliability, it is still essentially solving the initial problem via a more complex model - there is no free lunch. As mentioned, employing Gaussian mixture models in filtering is not a novel approach. In (Anderson and Moore, 2012) GMMs are discussed in their property to approximate any given distribution given sufficient components and the resulting possibilities regarding non-linear filtering. There, they aggregate a number of extended Kalman filters with weights adjusted via posterior residuals of the observation. The pioneering line of work in that regard can be traced back to (Sorenson and Alspach, 1971), (Alspach and Sorenson, 1972) and (Tam and Moore, 1977). An extension to particle filters was done in (Kotecha and Djuric, 2003) with an introduction of Gaussian mixture transform particle filters in (Reich and Cotter, 2015). Accordingly, the contribution of the work at hand is by no means in proposing aggregating a system of filters via Gaussian mixture models to better account for non-linearity. However, the idea in utilizing open-end sequential probability ratio tests of the model predictive, the conditional observation evidence, seems to not having been done thus far. The closest relative to the proposed CR-BOCPD GMM-KFs is likely the switching Kalman filter in (Murphy, 1998), however, it is also fairly more restrictive in explicitly assuming knowledge of a discrete hidden Markov model and its switching probabilities. The presented idea in combining testing goodness-of-fit and variance increment is much more general and aims to appeal to common data assimilation practice.

A major challenge of the presented approach is in tuning the covariance increment factor denoted λ in the experiment on the Lorenz-63 system and ξ referring to (Lai and Shan, 1999). Whenever we have little to no knowledge about the nature of change u_n and we do not want to propagate the system from the initial conditions, this becomes a necessary yet challenging task. Similar to tuning the generalised Bayesian learning rate in the previous chapter we may want to utilize access to an assumed model. Via simulating from that model we can then choose the scaling factor λ such that the resulting BOCPD GMM-KF performs satisfactory in some evaluation metric in a Monte Carlo fashion.

From here onwards, the presented results open pathways to several directions for additional investigation. In the context of sequential learning, and more specifically aggregating expert knowledge, the scenario weights offer plenty of room of improvement. Next to considering results in (Saad and Blanchard, 2021) for faster learning rates for expert advice, the current constraint on rate of false alarms via a constant threshold needs addressing. Switching to a curvilinear threshold while aiming to maintain some notion of optimality opens up the possibility to instead focus on a given probability of false alarm, so $\mathbb{P}_0[T < \infty] < 1$. The topic has first, insightful results in (Borovkov, 1999), proving first-order asymptotic optimality of detection delay for a double logarithmic threshold and very large change point location $\eta \gg 0$. Borovkov states that it appears to be a difficult problem, however, suggests logarithmic or even linear boundaries for change point locations η not too large.
The work in (Alami et al., 2020) provides such a curvilinear threshold achieving first-order asymptotic optimality in their specified sense for a given fixed PFA with their choice of tuning parameter in$\beta_{r,k,n} = \frac{1}{n-k+1}$. As argued in the work at hand, the tuning parameter in the CR-BOCPD procedure and the threshold in the CUSUM procedure share the relation $-\log(\beta_{r,k,n}) = c(r,k,n)$ with curvilinear threshold $c(r,k,n)$. Their choice of $\beta_{r,k,n} = \frac{1}{n-k+1}$ therefore then recovers a logarithmic threshold in $c(r,k,n) = \log(n-k+1)$ similar to what is suggested in (Borovkov, 1999). Intuitively it makes sense to relate the curvilinear threshold to controlling the tail behavior of the LLR sequence. For the Gaussian case as we are interested in, some more work is required. The current intermediate results following similar arguments as in (Alami et al., 2020) suggest a similar logarithmic curvilinear threshold. The main task hereby is in controlling the tail behavior of post-change product density functions $\prod_{s=k}^{n} p(y_s|y_{1:(s-1)}, k)$ via sub-exponential bounds for (non-central) χ^2-distributions as in (M. Ghosh, 2021). Controlling the probability of false alarm then takes a union bound argument over $n-k+1$ events, hence arguing for a curvilinear threshold of logarithmic order from the tail bounds. These first, superficial considerations encourage empirical investigation of a choice of $\beta_{r,k,n} = \frac{1}{n-k+1}$ in the context at hand with algorithm (4).

On a broader scope, connecting the field of sequential hypothesis testing and particle or ensemble filters opens several interesting new perspectives. To propose one such, employing the idea or the recursive χ^2-GLR algorithm based on closed SPRTs in (I. V. Nikiforov, 2001) and (A. Tartakovsky et al., 2014), we may aim for more adaptive replacement routines of single particles or ensemble members by evaluating their individual trajectory innovations in departure from mean 0 and only then replacing them with an offspring of a well-performing sibling. See algorithm (5) in the appendix for an outline of the approach.

As mentioned with the experiment on the Lorenz-63 model, the proposed approaches were not designed with non-linear, chaotic systems as main object of interest, however, the intuition transfer easily with the discussed change points in abrupt jumps easily transferring to concepts of phase-transition, regime switching, Levy-processes or simply low stability of chaotic systems. The main components regarding non-linear filtering discussed here in ensemble Kalman filters and Gaussian mixture models are already popular in that context. Adding ideas from sequential learning via SPRTs detecting model mis-fit is a promising addition and may be understood as a statisticians way to address challenges in non-linearity.

4. Double Noise Mis-Specification: Robust Inference Under Abrupt Change

4.1. Understanding the Challenge

While each of the previous two chapters respectively provided valuable insights based on theoretical foundation, the shared problem of inference under abrupt change via sudden jumps of the signal process as well as observation outliers, each potentially arising from mis-specification of heavy tailedness of the true noise distributions, is intuitively more difficult.
Recalling the original linear Gaussian setting of the celebrated Kalman filter in (2.1), we want to add the investigated modifications of the previous chapters to obtain the combined challenge of robust inference of the signal process under additive change and contaminated observations:

$$\begin{aligned} X_n &= A_n X_{n-1} + C_n W_n + \delta_{\eta,n} u_n \\ Y_n &= H_n X_n + \Gamma_n V_n^{\varepsilon}. \end{aligned} \tag{4.1}$$

As discussed in chapter (3), the additive term $\delta_{\eta,n}u_n$ hereby expresses condition and magnitude of respective additive changes and V_n^{ε} refers to contaminated observation noise as introduced in chapter (2).

Focusing on the proposed approaches for quantifying and targeting each mis-specification individually, there seems to be no immediate conflict between the two of them. The diffusion score matching Kalamn filter via robust Bayesian inverse inference in the analysis step can easily be combined with covariance increments in the forecast step and the Gaussian mixture model meta-structure based on the individual scenario evidences. Further, a superficial look at the developed theory seems to support the idea of intertwining both approaches as their individual assumptions do not interfere. The diffusion score matching analysis step still produces conditional Gaussian posteriors as required for the results on reliability of the change point detection schemes and the scenario initialization of the change point detection scheme produce Gaussian priors required for Gaussian diffusion score matching posterior.
Problems arise in the conflicting heuristics of each approach via re-scaling signal covariance.

4.1.1. Balancing a Trade-Off in Noise Covariance Scaling

A central motivation for the work at hand were the results on robust change point detection in (Altamirano et al., 2023b). The approach seems to support easy adaptation to the sequential Gaussian linear setting, however, this is somewhat misleading. The latent model they assume has no inherent stochastic sequential structure, but assumes a constant pre- and post-change regime of the latent component. The introduction of the signal process transforms the problem into a much more challenging one. More precise, it is the increase of uncertainty in the forecast step via the stochastic signal process that requires the popular dynamical balancing of signal state uncertainty and observation uncertainty via the Kalman gain that is resulting in problematic and difficult interactions for robust change point inference in linear Gaussian systems with the proposed methods.
Assuming that we do not have access to strong knowledge about the nature of an additive change in magnitude and direction under small uncertainty as will be usually the case, we need relying on the idea of covariance increments at scenario initialization to create forecasts that allow for better adaptation to a post-change regime then a current best model and compare observation evidence - the core idea of CR-BOCPD. Hereby, we explicitly utilize the property of the Kalman filter to balance uncertainties and by inflating uncertainty in the signal state of a forecast, we explicitly put emphasis and allow for more trust in the observations. The robust analysis step via the diffusion score matching posterior works similar, yet the other way around - it evaluates a notion of plausibility based on the forecast, the Mahalanobis distance in the IMQ-kernel, to counteract over-fitting. If the forecast is artificially inflated at scenario initialization, this notion of plausibility of an observation to determine outliers can not work effectively. While the proposed methods do not interfere with each other in theory, they somewhat work against each other in practice. All approaches employing covariance increments for scenario initialization will necessarily run into this issue.

This problem becomes much more severe, when telling apart outliers and change points is especially challenging due to very similar signatures on observations. In that case both instances of mis-specification in heavy tailedness cover and disguise each other and produce a problem highly prone to over-fitting on supposed change. An outlier that appears to be a plausible observation under a newly initialized scenario will not be corrected in the analysis step of this specific scenario filter. If it then happens to produce sufficient evidence, one may falsely account for a jump. Additionally, after a recent decision for change, uncertainty in the signal state is especially high with the filter needing to stabilize. This again somewhat weakens the workings of the robust posterior for a certain time period in that outliers may then be considered plausible given the high signal uncertainty in the early post-change regime.

Relying on Uncertainties for Intractable Epsiodes

So we understand why and in which constellation the problem is likely to be inherently difficult given the proposed methods regarding signal state estimation. Yet, there is a strong argument that although given the challenges, the approach combining the conditional CR-BOCPD GMM-Kalman filter with robust diffusion score matching posteriors will likely perform reasonable in appropriate evaluation as one of its strong points lies in producing well quantified uncertainties due to considering both types of mis-specification contrary to each previous method individually or the regular Kalman filter. This point needs additional emphasizing. The introduced setting of the double noise mis-specification makes exact estimation of the signal state close to impossible at times with change and outliers in quick succession and proximity, however, good uncertainty quantification, so indicating whenever we have a good idea about the signal state or when we do not, is then even more crucial. Accordingly, the experiments will slightly shift in focus and include an increased focus on evaluation metrics considering estimation confidence.

4.2. Experiments: Latent Estimation and Non-Linear Simulations

For this last set of experiments we again want to employ the target tracking example and the Lorenz-63 test with features to resemble double noise mis-specification in heavy tailedness. Tuning is hereby generally done as introduced in the respective sections of the previous chapters.
New to this chapter's experiments will be the mentioned more explicit focus on evaluation of uncertainty in signal estimates via median negative log-likelihood of the respective posterior distributions. At each time step we compute $-\log[p(x_n^t|y_{1:n})]$ for each investigated method with x_n^t the true latent signal at time n. In practice, this can be seen as a form of information criterion also accounting for uncertainty about the signal state. Some methods, such as the regular Kalman filter, occasionally produce $p(x_n^t|y_{1:n}) \approx 0$ in numerical evaluation due to completely out of tune uncertainties. Therefore, we do not want to simply take the sum over the negative log-likelihoods, but instead resort to the median for a notion of average performance of a method. Accordingly our main tool of evaluation for experiments in this section will lie in $\text{IC}(p) = \widehat{\text{med}}_N(\{-\log[p(x_n^t|y_{1:n})]\}_{1\leq n\leq N})$.

All simulations were done in R version 4.2.2.

4.2.1. Experiment A: Target Tracking Under Double Mis-Specification

The general setup is the combination of the target tracking examples of the previous two chapters including contamination of the observations and probabilistic additive jumps of the signal. We choose the popular target tracking task with signal space

$\mathcal{X} = \mathbb{R}^4$ containing x-position, x-velocity, y-position as well as y-velocity, and observation space $\mathcal{Y} = \mathbb{R}^2$ containing the measured x-position, as well as y-position. We design the signal model discrete. Given the usual Kalman setting but with the additional terms, we have

$$\begin{aligned} X_{n+1} &= A_n X_n + Q_n^{\frac{1}{2}} W_n + \pi_n u_n \\ Y_n &= H_n X_n + V_n^{\varepsilon} \end{aligned} \tag{4.2}$$

with W_n standard Gaussian noise and V_n^{ε} contaminated noise in the respective dimensions, $\pi_n \sim_{iid} \mathrm{Ber}(\varphi)$ and $u_n \sim_{iid} \mathcal{N}(0, U)$ with $U = \tilde{u}\begin{pmatrix} 0&0&0&0 \\ 0&1&0&0 \\ 0&0&0&0 \\ 0&0&0&1 \end{pmatrix}$. The contaminated observations are generated via $V_n^{\varepsilon} \sim \mathcal{N}(0, R_n) + \varepsilon\mathcal{N}(0, \tilde{c}R_n)$ with $R_n = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}$ and $\varepsilon \in [0, 1]$.
We are interested in investigating performance for different choices of probabilities ε and φ as well as magnitudes $\tilde{c}$ and $\tilde{u}$.

As usual for these kind of models, yet much more simplified, we choose $A_n = \begin{pmatrix} 1&1&0&0 \\ 0&1&0&0 \\ 0&0&1&1 \\ 0&0&0&1 \end{pmatrix}$, $H_n = \begin{pmatrix} 1&0&0&0 \\ 0&0&1&0 \end{pmatrix}$ and signal noise covariance $Q_n = 0.1 \cdot \begin{pmatrix} 1 & 0.5 & 0.5 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.5 & 0 & 1 & 0.1 \\ 0 & 0 & 0.1 & 1 \end{pmatrix}$. The initial signal is chosen as $X_0 = (0, 1, 0, 0)^T$.
We simulate X_n for n in $[0, 2]$ with step-size $\mathrm{d}n = 0.01$ resulting in 200 positions with $n_{out} = 1$, so 200 observations. Next to the regular Kalman filter, we implement the $\mathcal{D}_w$-Kalman filter with default learning rate $\beta = 1$ and a CR-BOCPD GMM-Kalman filter with the corresponding scenario initialization, threshold and window size as in the previous section. Additionally we implement a combination of both, the robust CR-BOCPD GMM-Kalman filter, implementing diffusion score matching posteriors in the CR-BOCPD scheme.

Starting with a more qualitative investigation, on the central methods of the previous chapters, we set $\varphi = 0.05$, $\varepsilon = 0.15$ and $\tilde{u} = \tilde{c} = 100$.

We observe the expected results in figure (4.1). The CR-BOCPD GMM-KF suspects change for a large number of outliers and is generally very unstable. The $\mathcal{D}_w$-Kalman filter is generally stable, however, very slow in adapting to instances of change instead suspecting outliers. The regular Kalman filter performs surprisingly well in that it also lacks behind instances of change, however, this way somewhat balancing out outliers. This is especially interesting when considering Kalman's interpretation of the Kalman filter as a minimum squared error estimator under severely relaxed assumptions. Figure (4.2) showing the difference in mean from the true position further supports these results.

As expected, the combined scheme performs noticeably better. As can be seen in figure (4.3) and figure (4.4), it still occasionally over-fits to outliers, however, in much less severe scale and frequency compared to the regular CR-BOCPD GMM-KF counterpart. Similarly, it also quickly corrects itself afterwards. Looking at the information score, the combined method can be assumed to provide reasonably tuned quantification of its uncertainties, especially considering the general difficulty of the problem as discussed. For repeated experiments, similar insights with

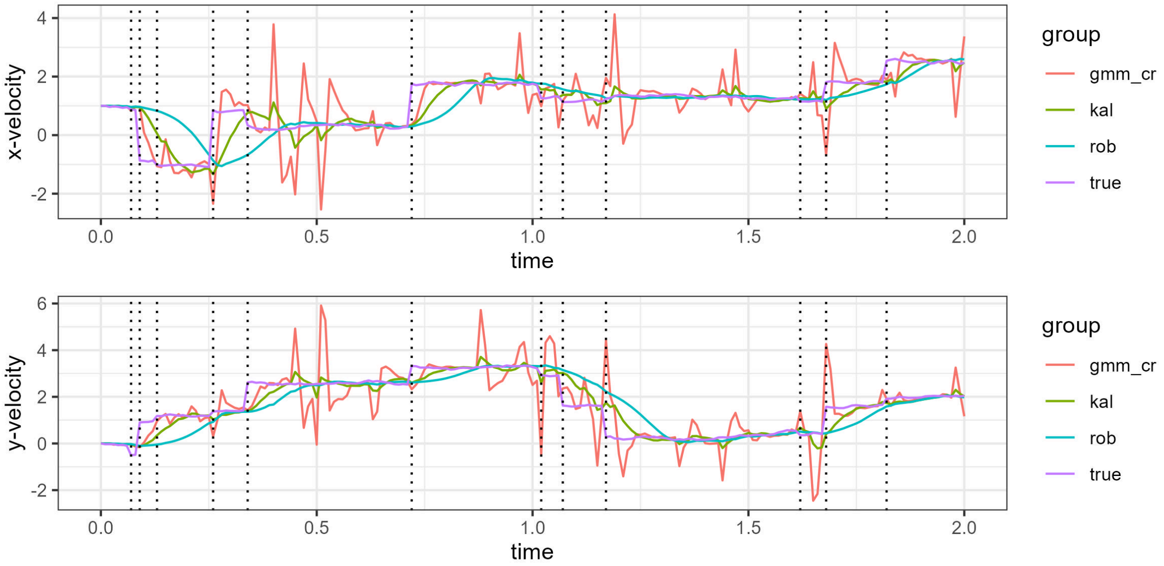


Figure 4.1.: True trajectory in x and y-velocity (true) and analysis mean velocities for the regular KF (kal), $\mathcal{D}_w$-KF (rob) and CR-BOCPD GMM-KF (gmm_cr). Dotted lines indicate instances of change.

similar scales of the information score could be observed.

Evaluation	KF	$\mathcal{D}_w$-KF	GMM-KF	$\mathcal{D}_w$-GMM-KF
MSE	0.57	5.79	1.09	0.27
IC	17.07	13.17	3.47	-1.57

Table 4.1.: Mean squared error and information criterion score of each method for the trajectories of the provided graphs in figure (4.1), (4.2), (4.3) and (4.4).

Table (4.1) provides a very curious insight in that the regular Kalman filter provides a surprisingly low MSE with a large score in information criterion somewhat aligning with its second interpretation. Otherwise, results are as expected with the combined scheme showing strong performance both in MSE and in information criterion score. However, considering the additional resources it requires this result needs putting into perspective. One might consider employing other evaluation measures such as the Akaike information criterion or the Bayesian information criterion to better account for the additional parameters in the more complicated schemes. However, both these measures will necessarily run into the numerical issue described that requires use of the median of negative log-likelihoods in the first place. In other words, due to the strongly misjudged uncertainties, i.e. of the Kalman filter, evaluating the AIC or BIC for these methods will blow up to an extend that no number of parameters in the more sophisticated schemes can catch up.
Additional graphs of full side-by side comparisons as well as the individual dimensions and differences from the true signal are provided via figure (A.37) to (A.56) in the appendix.

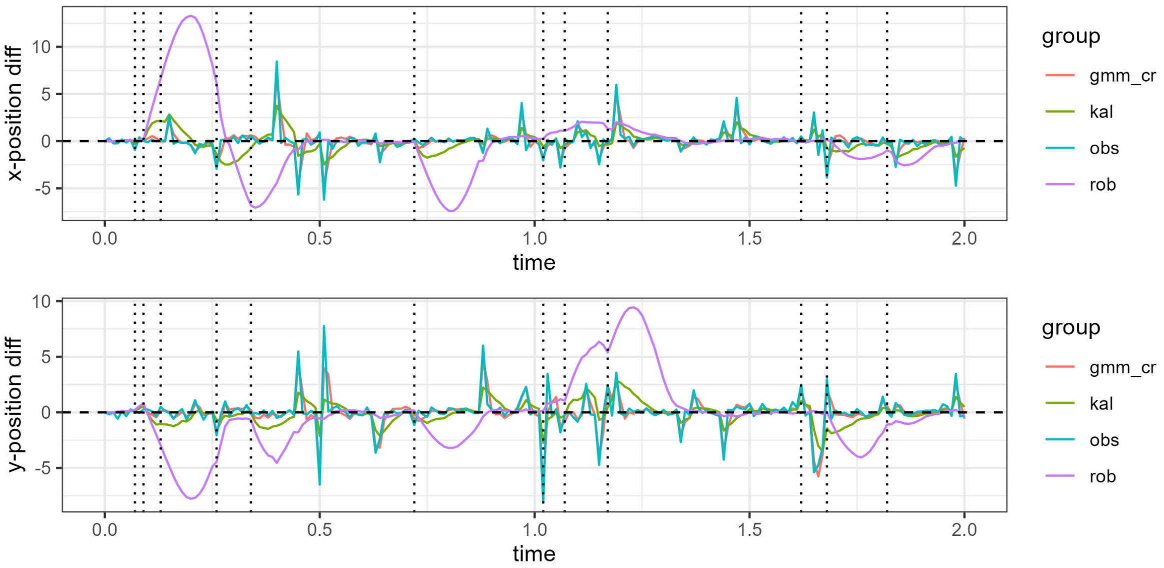


Figure 4.2.: Difference of analysis means and true x and y-positions for the observation (obs), the regular KF (kal), $\mathcal{D}_w$-KF (rob) and CR-BOCPD GMM-KF (gmm_cr). Dotted lines indicate instances of change.

A main interest of this section lies in investigating performance of the proposed method related to severeness of the individual mis-specifications. Fixing the covariance factor of the respective events mimicking mis-specification in heavy tailedness at $\tilde{u} = 100$ and $\tilde{c} = 100$, we want to scale the individual magnitudes φ and ε and investigate the impact in the information score for both the regular Kalman filter and the combined method. We hereby scale each from 0 to 0.5 in steps of 0.05 and take the median over the information criterion scores of 100 Monte Carlo Simulations for each pairing on the resulting grit.

The results in figure (4.5) are not necessarily surprising, however, they indicate at discussed aspects. Further, the scales of the information criterion scores for the individual methods support arguments that the scale of evaluation metrics of single experiments, e.g. as provided in the tables, generally translate for repeated simulations.
To go more into detail, for the two specific methods in the $\mathcal{D}_{\sqsupseteq}$-Kalman filter and the CR-BOCPD GMM-Kalman Filter, we can clearly see the color gradient in the expected direction. The $\mathcal{D}_w$-KF performs seemingly independent of the contamination probability ε in that the color gradient scales directly with the change point probability φ and the other way around for the CR-BOCPD GMM-KF. For the regular Kalman filter and the combined scheme, the color gradient runs more diagonal from low probabilities to high probabilities. Interestingly, both seem to scale better with φ than with ε, however, this may likely also result from the choice of the corresponding covariance terms relative to model dynamics, so the covariance factor impacting observed position values compared to the covariance factor for change distorting the latent velocity values.
Similar color gradients can be observed also in MSE and for scaling covariance with fixed probabilities. The corresponding plots are provided in the appendix via figures

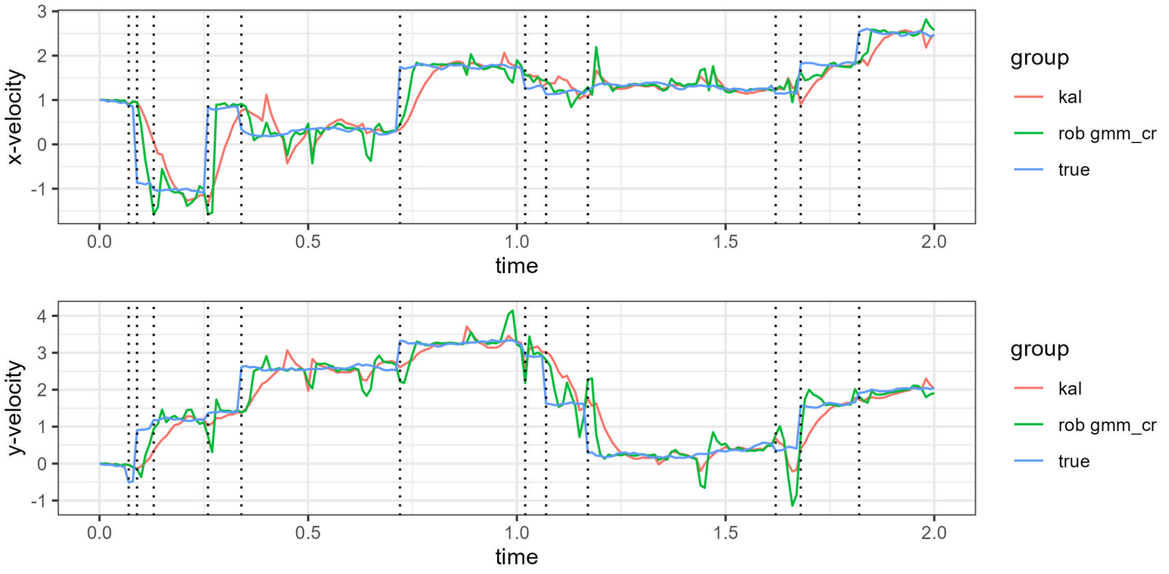


Figure 4.3.: True trajectory in x and y-velocity (true), analysis mean velocities for the regular KF (kal) and the combined robust CR-BOCPD GMM-KF (rob gmm_cr). Dotted lines indicate instances of change.

(A.57), (A.59) and (A.58).

4.2.2. Experiment B: The Lorenz-63 Test

For the last experiment, we want to investigate transfer to non-linear dynamical systems. Hereby the problem again is much more challenging as even small differences in estimation may lead to severe error due to the chaotic nature, as previously discussed. Again, we switch to ensemble based implementations.
As before, we take the simulation mainly from (Reich and Cotter, 2015). Recall, let z be the signal variable, we then have the vector field f given by

$$f(z) := \begin{pmatrix} \sigma(z_2 - z_1) \\ z_1(\rho - z_3) - z_2 \\ z_1 z_2 - \beta z_3 \end{pmatrix} \tag{4.3}$$

with parameters $\sigma = 10$, $\rho = 28$ and $\beta = \frac{8}{3}$. We chose a step-size of $\delta t = 0.001$ and the initial value $z_0 = (-0.587, -0563, 16.870)^T$. To implement a forward Euler scheme as numerical approximation, we include a non-autonomous forcing term g_n that essentially comes down to a tent map iteration. We set $a = (\delta t)^{-\frac{1}{2}}$ and the initial forcing term as $g_0 = (a(2^{-\frac{1}{2}} - \frac{1}{2}), a(3^{-\frac{1}{2}} - \frac{1}{2}), a(5^{-\frac{1}{2}} - \frac{1}{2}))$ with the entry-wise recursive definition

$$g_{n+1,i} = \begin{cases} 1.99999 g_{n,i} + \frac{a}{2} & \text{if } g_{n,i} < 0 \\ -1.99999 g_{n,i} + \frac{a}{2} & \text{otherwise.} \end{cases} \tag{4.4}$$

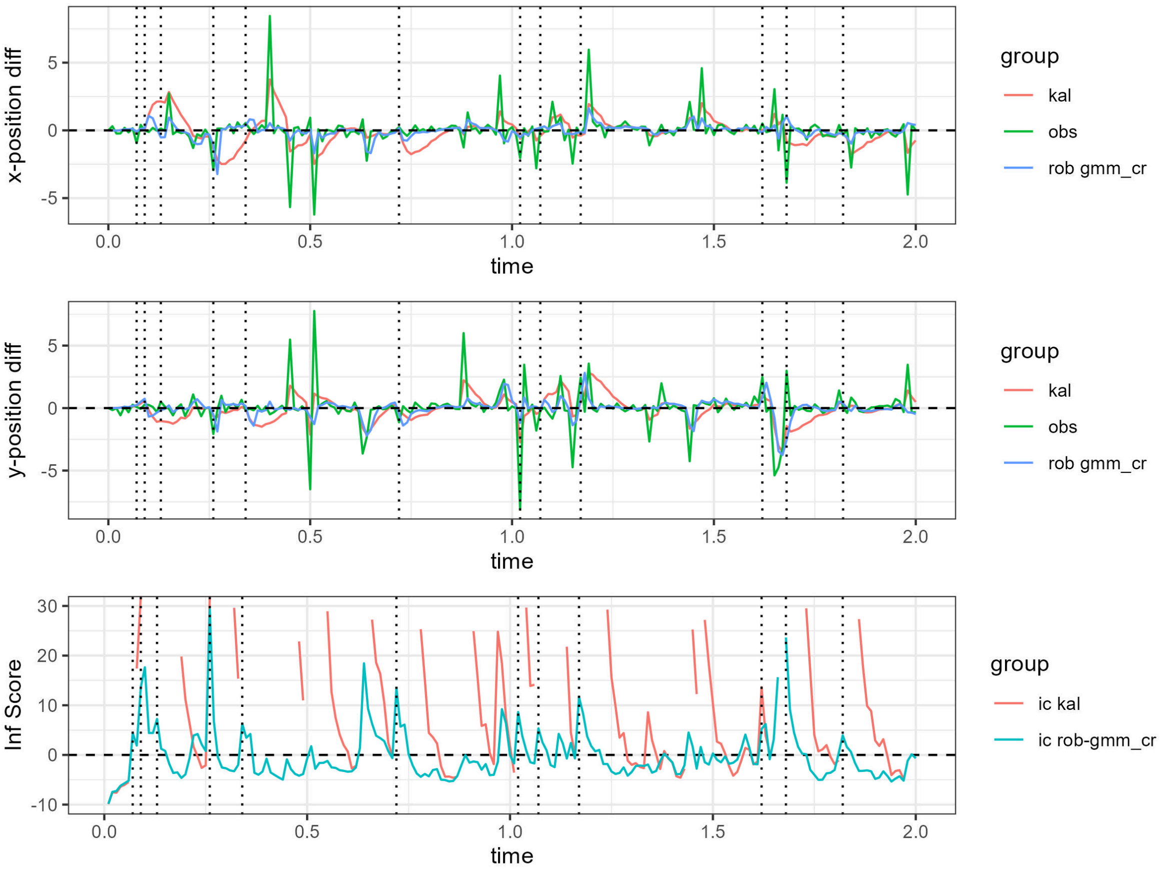


Figure 4.4.: Difference of analysis means and true x and y-positions (top and middle) for the observations (obs), the regular KF (kal) and the combined robust CR-BOCPD GMM-KF (rob gmm_cr). Information criterion score over time (bottom) for the regular KF (ic kal) and robust CR-BOCPD GMM-KF (ic rob-gmm_cr). Dotted lines indicate instances of change Missing sections of the graph result from values outside the range of the graph.

The signal is the propagated via

$$z_{n+1} = z_n + \delta t(f(z_n) + g_n) \tag{4.5}$$

over a window $[0, 10]$ for investigating performance without mis-specifications. Observations were generated at $t_{out} = 100$ via

$$Y_n = H_n z_n + \sqrt{2} V_n \tag{4.6}$$

with a strict observation map $H_n = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$ only allowing noisy observation of the first component via V_n standard Gaussian noise. We chose an ensemble size of $M = 5$ to mimic application for expansive forward models.

We implement four schemes in total with the first three taken from previous experiment in the regular perturbed EnKF, the $\mathcal{D}_w$-EnKF and the CR-BOCPD GMM-EnKF based on the CUSUM rule for conditional density functions. The $\mathcal{D}_w$-EnKF uses the default choice of tuning parameter in $\beta = 1$. Threshold and window-size in

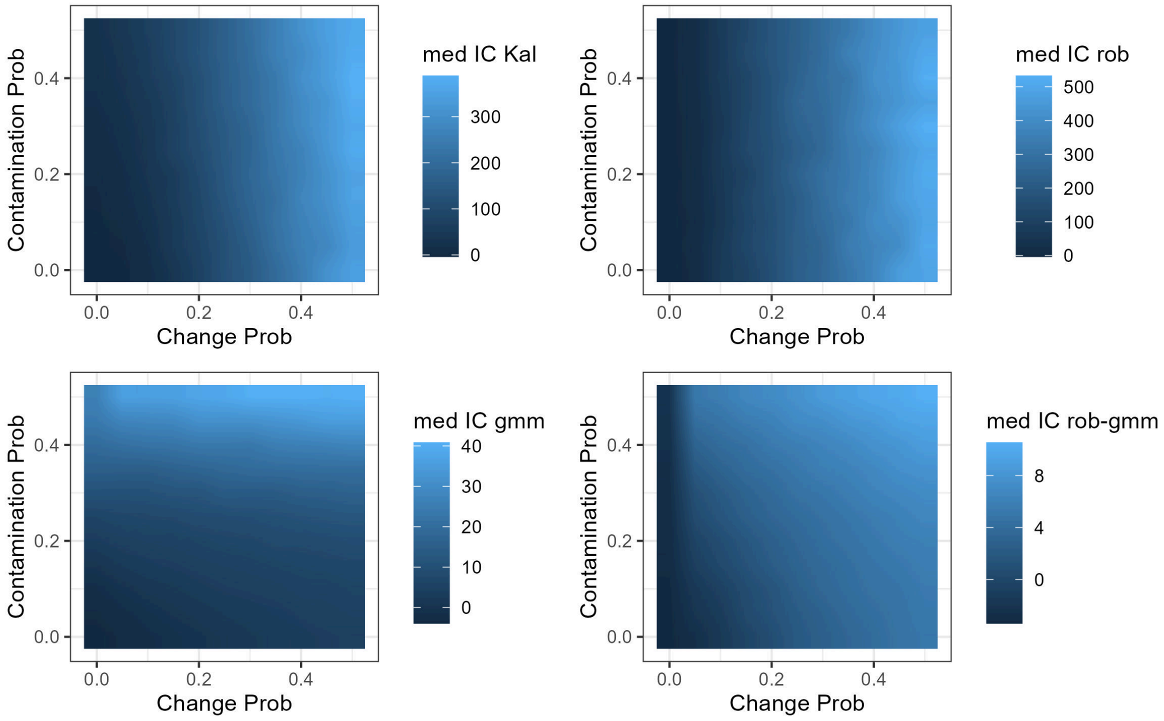


Figure 4.5.: Tile plot with interpolation of change point probability φ and contamination probability ε regarding median information criterion score over repeated Monte Carlo simulations for fixed covaraince factors $\tilde{u}$ and $\tilde{c}$ and each individual method: The regular KF (top-left), the $\mathcal{D}_{\sqsupseteq}$-KF (top-right), the CR-BOCPD GMM-KF (bottom-left) and the robust CR-BOCPD GMM-KF (bottom-right).

CR-BOCPD scheme are chosen as previously described via Wald's approximation and the poly-logarithm. At initialization of a new scenario at time $k > r$, the empirical covariance $\hat{P}_{r,k}$ of the current baseline scenario inflated by a factor $\lambda = 6.66$ and the empirical mean $\hat{m}_{r,k}$ are taken to draw a new ensemble $\{x_{k,k}^{a,(l)}\}_{l \in \{1,2,\dots,M\}}$, so $X_{k,k}^{l} \sim_{iid} \mathcal{N}(m_{k,k}, P_{k,k})$ with $m_{k,k} = \hat{m}_{r,k}$ and $P_{k,k} = \lambda \hat{P}_{r,k}$. This new ensemble is then propagated for the next forecast, but not considered for the previous analysis aggregation.

Evaluation	EnKF	$\mathcal{D}_w$-EnKF	GMM-EnKF	$\mathcal{D}_w$-GMM-EnKF
MSE	103.9	10.7	2.5	11.8
IC	-	41.7	6.7	7.9

Table 4.2.: Mean squared error and information criterion score of each method for the trajectories of the provided graphs in figure (4.6).

Both the graphs in figure (4.6) and the corresponding numerical evaluation criteria in table (4.2) indicate, that for the Lorenz-63 model with no mis-specifications the CR-BOCPD GMM-EnKF provides both the best mean estimate as well as uncertainty quantification. While the regular EnKF with perturbed observations is

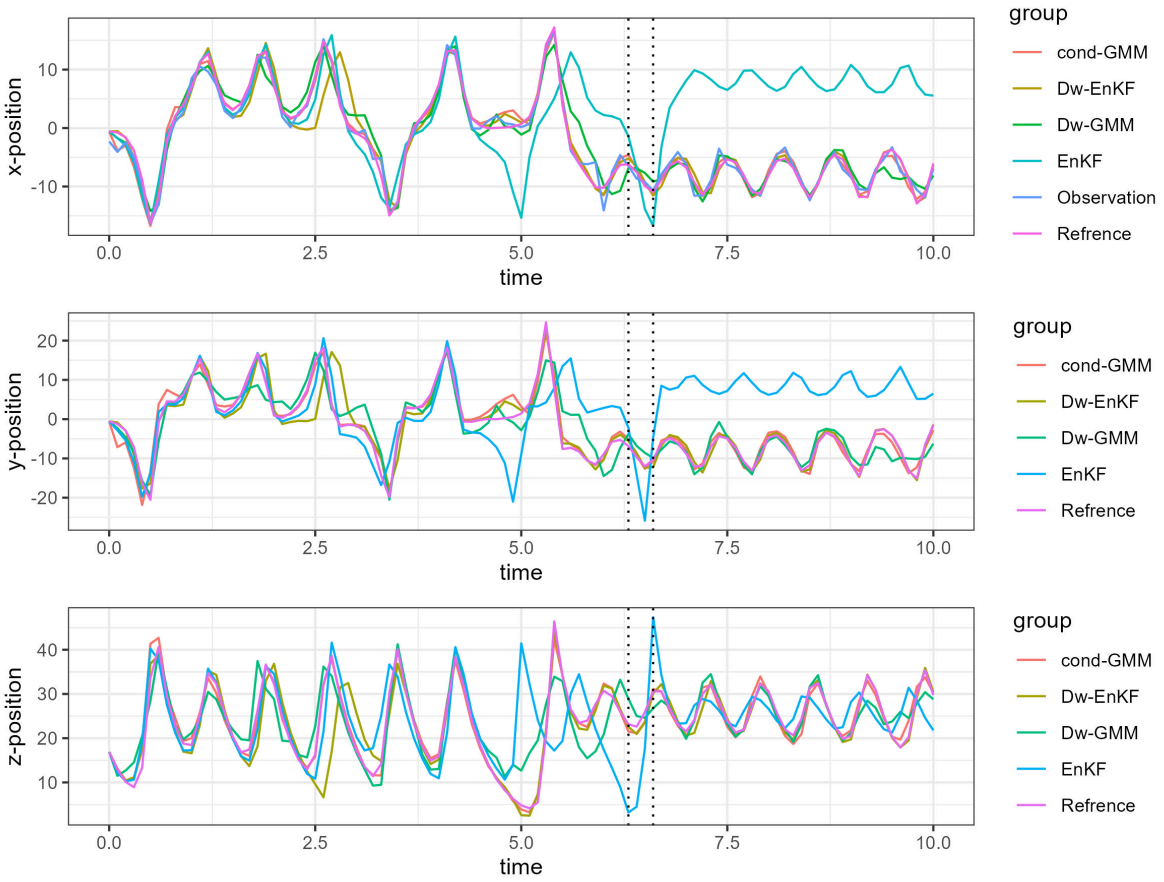


Figure 4.6.: Side-by-side graph comparison of the simulated signal (Reference), generated observations (Observation), the analysis mean of the regular EnKF (EnKF), the default $\mathcal{D}_w$-EnKF (Dw-EnKF), the CR-BOCPD GMM-EnKF (cond-GMM) and the robust CR-BOCPD GMM-EnKF (Dw-GMM). Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

thrown off and blew up even in median of the negative log-likelihoods, the $\mathcal{D}_w$-EnKF and the combined scheme provided reasonable performance. Both are off at times, however, manage to recover again. The combined scheme hereby seems to be somewhat "over-engineered" for the well-specified non-linear case.

We implement abrupt jumps in the Lorenz-63 system as before via deterministic resets of the system at random times keeping the desired non-linear behaviour. We start by simulating a reference trajectory as previously described for the interval $[0, 40]$. Next we draw five time points over the whole number of steps of the simulation with a binomial distribution with probability $p = 0.2$. We combine subsequences of the reference trajectory up to the sampled time points back-to-back to obtain a new trajectory that resets after these random time points to the initial conditions Different from the previous experiment, we add contaminated noise to the observation of the first component via replacing V_n with $V_n^\varepsilon \sim \mathcal{N}(0, 1) + \varepsilon\mathcal{N}(0, 100)$ with $\varepsilon = 0.2$.

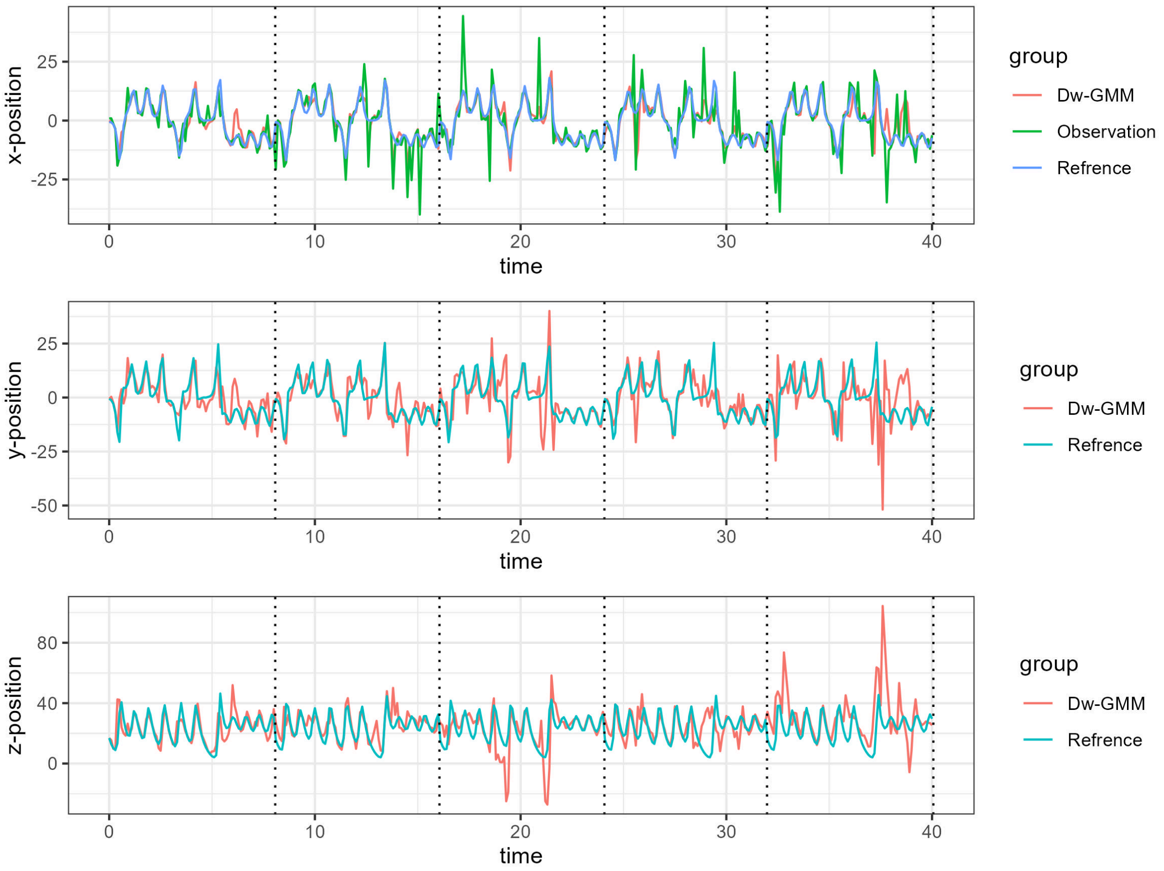


Figure 4.7.: Side-by-side graph comparison of the simulated signal with jumps (Reference) and contaminated observations (Observations) as well as the analysis mean of the robust CR-BOCPD GMM-EnKF (Dw-GMM). Dotted lines indicate instances of change via a restart of the system.

The results in figure (4.7) and table (4.3) align with results in the previous experiment up to considerations of non-linearity. The frequent outliers throw of the mean estimate of the CR-BOCPD GMM-EnKF yet with still reasonable uncertainties indicated by the IC score. The $\mathcal{D}_w$-EnKF performs has a reasonable performance as could be expected given its previous performances. The change in the occasional reset is a lot less severe compared to the target tracking experiment and thus does not play into its main weakness as much. The combined scheme can be considered to perform best with reasonable mean estimates and the best IC score. Accordingly, only its trajectory is shown with the graphs otherwise overly cluttered.

While comparison regarding number of model parameters was already considered with the previous experiment and the take-away still holds, we face another, much more significant discussion here in the number of ensemble members. The GMM-EnKF schemes employ a much larger ensembles by a factor of the CR-BOCPD window-size in practice. For the concrete experiment this means that while EnKF and the $\mathcal{D}_w$-EnKF work on the given ensemble size of $M = 5$, the GMM-EnKF schemes use up to $m(\alpha) \cdot M = 55$ ensemble members with non-uniform weights.

Evaluation	EnKF	$\mathcal{D}_w$-EnKF	GMM-EnKF	$\mathcal{D}_w$-GMM-EnKF
MSE	109.5	53.3	490.4	71.8
IC	-	137.8	57.0	10.1

Table 4.3.: Mean squared error and information criterion score of each method for the trajectories of the provided graphs in figure (4.7).

This point has been discussed in the previous section, however, the results of this experiments emphasize the proficiency of the $\mathcal{D}_w$-EnKF compared to the GMM-EnKFs given contamination. Accordingly we want to repeat the experiment with an adapted number of 55 ensemble members for the EnKF and the $\mathcal{D}_w$-EnKF. The experimental set up is hereby the exact same otherwise.

Evaluation	EnKF	$\mathcal{D}_w$-EnKF	GMM-EnKF	$\mathcal{D}_w$-GMM-EnKF
MSE	39.7	1.6	3.6	9.1
IC	178.5	2.4	8.5	6.1

Table 4.4.: Mean squared error and information criterion score of each method for the trajectories of the provided graphs in figure (4.8).

For the experiment with no mis-specification as presented in figure (4.8) and table (4.4), we observe are rapid increase in performance of the regular perturbed EnKF and the $\mathcal{D}_w$-EnKF. Moreover, the $\mathcal{D}_w$-EnKF provides a very strong performance both in MSE and IC score overtaking both GMM-EnKFs.

Evaluation	EnKF	$\mathcal{D}_w$-EnKF	GMM-EnKF	$\mathcal{D}_w$-GMM-EnKF
MSE	122.3	17.6	261.7	83.4
IC	-	3.6	67.5	12.6

Table 4.5.: Mean squared error and information criterion score of each method for the trajectories of the provided graphs in figure (4.9).

The results with contaminated observations and restarts of the system presented in figure (4.9) and table (4.5) further support these insights. While the adapted number of ensemble members are not sufficient for the perturbed EnKF to overcome the mis-specifications, it strongly improves the performance of the $\mathcal{D}_w$-EnKF to a degree that it noticeably stands out even from the combined method. Moreover, the results further emphasize a strong proficiency and scaling of the $\mathcal{D}_w$-EnKF regarding non-linear models. We hereby have to keep in mind that its weak point in slow adaptation to strong change as was observed in the target tracking experiment does no come into play as much here with the small number of resets.

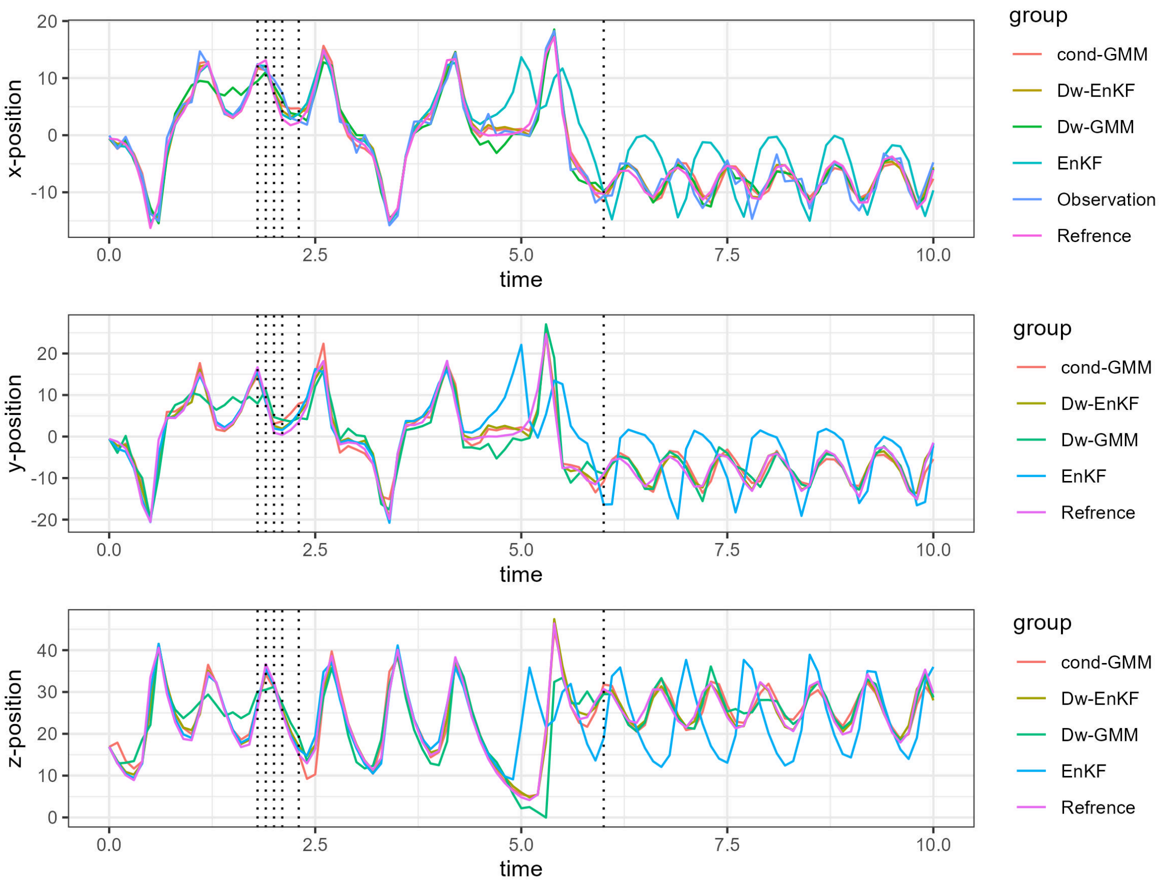


Figure 4.8.: Side-by-side graph comparison of the simulated signal (Reference), generated observations (Observation), the analysis mean with equalized number of ensemble members of the regular EnKF (EnKF), the default $\mathcal{D}_w$-EnKF (Dw-EnKF), the CR-BOCPD GMM-EnKF (cond-GMM) and the robust CR-BOCPD GMM-EnKF (Dw-GMM). Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

For the applications in mind, every ensemble member is costly, yet the GMM-EnKFs necessarily need a fairly large number of them to sufficiently approximate each scenario. The results of this experiment indicate that for non-linear dynamical systems with strong contamination and comparably weak or low magnitude of change, the diffusion score matching ensemble Kalman filter with perturbed observations provides a very strong approximation outperforming the GMM-EnKF given you can afford the number of ensemble members. If you can not afford the number of ensemble members required for the GMM-EnKF variants, the $\mathcal{D}_w$-EnKF still provides reasonable performance.

Again it has to be said that the results have not be drawn over a large number of Monte Carlo simulations, however, the results repeated for different runs and can be taken to apply in similar scale. Additional graphs including the individual dimensions are provided via figure (A.60) to (A.71) in the appendix.

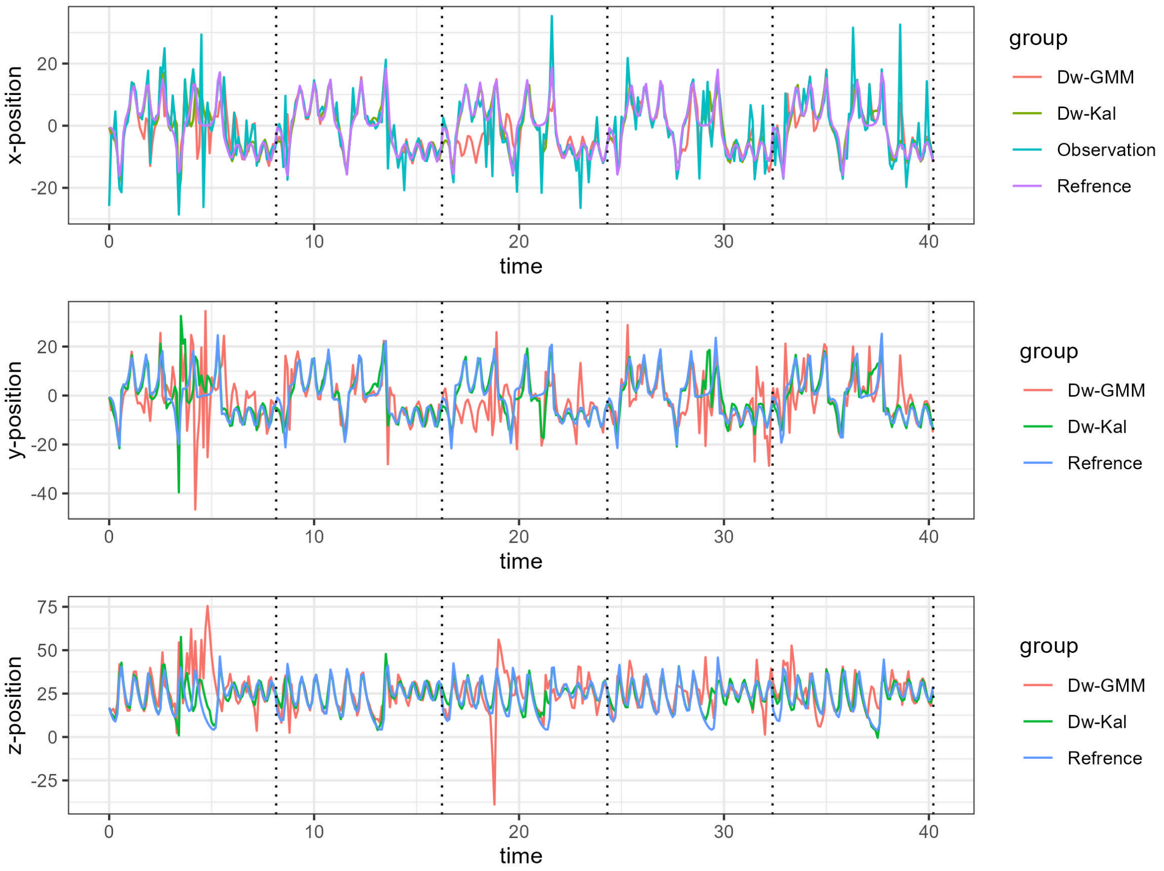


Figure 4.9.: Side-by-side graph comparison of the simulated signal with jumps (Reference) and contaminated observations (Observation) as well as the analysis mean with equalized number of ensemble members of the $\mathcal{D}_w$-EnKF (Dw-Kal) and robust CR-BOCPD GMM-EnKF (Dw-GMM). Dotted lines indicate instances of change via a restart of the system.

4.3. Discussion and Conclusion

This chapter provided the combination of several individual results and in a way represents the peak of considerations on mis-specification in Bayesian filtering. Yet it does not include any new theory instead analysing interaction of the previously derived methods.

As discussed, the problem at hand of double mis-specification of the noise terms in heavy-tailedness is highly difficult. For the linear Gaussian case the presented approach of the CR-BOCPD GMM-KF with diffusion score matching posteriors provided the best results in good uncertainty quantification - as is intended by design. Considering the theoretical and computational machinery in play for this scheme, this result is highly desirable. An interesting observation hereby was mainly with the good performance of the mean estimate of the regular Kalman filter in mean squared error.

Broadening the horizon via considering non-linear, chaotic dynamical systems, the robust CR-BOCPD GMM-EnKF seems somewhat over-engineered requiring a large number of ensemble members for the individual scenarios. Scaling the $\mathcal{D}_w$-EnKF to a comparable number of ensemble members, it provides a very strong performance in proficiency both of mean estimate for the signal state and uncertainty quantification measured via an adapted information criterion.

The key take away of the work at hand lies in these two insights. For the linear Gaussian setting with double noise mis-specification the designed combined scheme provides a reasonable solution. Thus far nothing comparable is present in the literature and neither in related topics such as robust change point detection in linear Gaussian systems. The result is a starting point for the debate on mis-specification, not an exhaustive conclusion. In the severely more complicated non-linear case, approximation with reasonable uncertainty quantification is all one may hope for. The $\mathcal{D}_w$-EnKF provides a highly promising result in that case.

The presented results are ground for further investigation into a large variety of directions. Each chapter has its own, individual details, pits and stepping stones providing both novel insights and questions for adjacent fields.

5. Concluding Remarks

> [One] relies on assumptions about the distribution of errors.[...] This is inherently difficult,
> because it requires statements [...] about things we truly do not know.
>
> *Morzfeld and Reich*

The obtained results of the work at hand have already been discussed for appropriate context at the end of each individual chapter 2, 3 and 4. The main outputs are three propositions, two algorithms and eight simulation experiments. Hereby, we want to consider the propositions and corresponding algorithm on the diffusion score matching Kalman filter in conjugacy and robustness of the posterior most rigorous while other results can be taken to be preliminary needing additional investigation in future work with addressing details. Especially the result in chapter **??** require considering adapted notions of optimality regarding fixed probability of false alarm and more in-depth development of arguments.

Regardless, next to theoretical results the work at hand has value and contribution in opening a discourse for perspectives from limitations and challenges in the practice of statistical modelling. Circling back to the initial context Bayesian forecasting and data assimilation, opening the discussion on model mis-specification in the noise terms is crucial. It needs developing novel methods such as done here to address these necessary hurdles. The proposed strategies in this work try to balance feasibility in practice via numerical complexity and required assumptions with theoretical foundation. Considering the format of the work at hand it then puts what was achived into perspective.

Leaving the context of mis-specification for the much broader context of data assimilation under reduced assumptions explicitly including non-linear systems, we want to highlight the two contributions we deem most relevant for future directions. The very strong performance of the diffusion score matching ensemble Kalman filter with perturbed observations was somewhat surprising, however, different intuitions of its inner workings are arising. Likely the main contributor to its proficiency is the combination of the adapted Kalman gain and the novel divergence term on the level of the individual ensemble member via a push-and-pull type interaction similar to a repulsion counteracting over-fitting and spacing out available particles.

It implicitly incorporates the current practice of artificially incrementing covariance in high-dimensional and costly EnKF applications into its mechanics.
We want to place the second spotlight on the novel idea of exploiting established frequentist practices in sequential hypothesis testing for model fit in the Bayesian dominated field of stochastic filtering. The proposed schemes in the scope of detecting additive changes are just one such exploration of synergies, yet likely (too) costly for some filtering applications. An other example, the proposed intertwining of ensemble or particle methods and sequential testing as briefly discussed at the end of chapter 3 and sketched in algorithm (5) in the appendix makes intuitive sense. It provides a new perspective to the idea of resampling via enabling evaluation of single particle evolutions and their trajectories also when importance weights might be unreliable, e.g. under mis-specification. As said, it can be taken to be a frequentist statisticians way of approaching challenges arising with non-linearity via accepting or rejecting a current model.

Where the work at hand lacks detail, it is rich in diversity of the combined methods actively intersecting different fields. It is a departure towards new perspectives instead of a pushing of current practice with in its known limitations. Plenty of research has to follow, yet different research communities have access to contribute.

Bibliography

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Agamennoni, G., Nieto, J. I., & Nebot, E. M. (2011). An outlier-robust kalman filter, In *2011 ieee international conference on robotics and automation*. IEEE.

Agudelo-España, D., Gomez-Gonzalez, S., Bauer, S., Schölkopf, B., & Peters, J. (2020). Bayesian online prediction of change points, In *Conference on uncertainty in artificial intelligence*. PMLR.

Alami, R., Maillard, O., & Féraud, R. (2020). Restarted bayesian online changepoint detector achieves optimal detection delay, In *International conference on machine learning*. PMLR.

Alspach, D., & Sorenson, H. (1972). Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, *17*(4), 439–448.

Altamirano, M., Briol, F.-X., & Knoblauch, J. (2023a). Robust and conjugate gaussian process regression. *arXiv preprint arXiv:2311.00463*.

Altamirano, M., Briol, F.-X., & Knoblauch, J. (2023b). Robust and scalable bayesian online changepoint detection. *arXiv preprint arXiv:2302.04759*.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Et al. (2023). Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, *38*(1), 120–139.

Anderson, B. D., & Moore, J. B. (2012). *Optimal filtering*. Courier Corporation.

Bai, M., Sun, C., & Zhang, Y. (2022). A robust generalized t distribution-based kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, *58*(5), 4771–4781.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., & Mackey, L. (2019). Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, *32*.

Basseville, M., & Benveniste, A. (1983). Sequential detection of abrupt changes in spectral characteristics of digital signals. *IEEE Transactions on Information Theory*, *29*(5), 709–724.

Basseville, M., Nikiforov, I. V. Et al. (1993). *Detection of abrupt changes: Theory and application* (Vol. 104). prentice Hall Englewood Cliffs.

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55.

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(5), 1103–1130.

Borovkov, A. A. (1999). Asymptotically optimal solutions in the change-point problem. *Theory of Probability & Its Applications*, *43*(4), 539–561.

Boustati, A., Akyildiz, O. D., Damoulas, T., & Johansen, A. (2020). Generalised bayesian filtering via sequential monte carlo. *Advances in neural information processing systems*, *33*, 418–429.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211–243.

Brodsky, B. (2016). *Change-point analysis in nonstationary stochastic models*. CRC Press.

Carrara, N., & Ernst, J. (2020). On the estimation of mutual information, In *Proceedings*. MDPI AG.

Chang, G. (2014). Robust kalman filtering based on mahalanobis distance as outlier judging criterion. *Journal of Geodesy*, *88*(4), 391–401.

Das, S., Kilic, C., Watson, R., & Gross, J. (2021). A comparison of robust kalman filters for improving wheel-inertial odometry in planetary rovers, In *Proceedings of the 34th international technical meeting of the satellite division of the institute of navigation (ion gnss+ 2021)*.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1–18.

Duchi, J. (2007). Derivations for linear algebra and optimization. *Berkeley, California*, *3*(1), 2325–5870.

Evensen, G. Et al. (2009). *Data assimilation: The ensemble kalman filter* (Vol. 2). Springer.

Evensen, G., Vossepoel, F. C., & van Leeuwen, P. J. (2022). *Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem*. Springer Nature.

Fearnhead, P., & Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(4), 589–605.

Frazier, D. T., Kohn, R., Drovandi, C., & Gunawan, D. (2023). Reliable bayesian inference in misspecified models. *arXiv preprint arXiv:2302.06031*.

Gan, D., & Liu, Z. (2020). On the stability of kalman filter with random coefficients. *IFAC-PapersOnLine*, *53*(2), 2397–2402.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

Ghosh, A., & Basu, A. (2016). Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, *68*, 413–437.

Ghosh, M. (2021). Exponential tail bounds for chisquared random variables. *Journal of Statistical Theory and Practice*, *15*(2), 35.

Girshick, M. A., & Rubin, H. (1952). A bayes approach to a quality control model. *The Annals of mathematical statistics*, *23*(1), 114–125.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations*. JHU press.

Gooijer, J. G., & Hyndman, R. J. (2005). *25 years of iff time series forecasting: A selective review*. Tinbergen Institute.

Grünwald, P. (2012). The safe bayesian: Learning the learning rate via the mixability gap, In *International conference on algorithmic learning theory*. Springer.

Harrison, P. J., & Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *38*(3), 205–228.

Huber, P. J. (2004). *Robust statistics* (Vol. 523). John Wiley & Sons.

Husiniga, W. (2021). Lecture notes in statistics. University of Potsdam.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Hyvärinen, A., & Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*(4).

Jackson, J. E., & Bradley, R. A. (1961). Sequential χ2-and t2-tests. *The Annals of Mathematical Statistics*, 1063–1077.

Jewson, J., Smith, J. Q., & Holmes, C. (2018). Principles of bayesian inference using general divergence criteria. *Entropy*, *20*(6), 442.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Knoblauch, J., & Damoulas, T. (2018). Spatio-temporal bayesian on-line changepoint detection with model selection, In *International conference on machine learning*. PMLR.

Kokko, H. (2005). Useful ways of being wrong. *Journal of evolutionary biology*, *18*(5), 1155–1157.

Kotecha, J. H., & Djuric, P. M. (2003). Gaussian sum particle filtering. *IEEE Transactions on signal processing*, *51*(10), 2602–2612.

Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information theory*, *44*(7), 2917–2929.

Lai, T. L., & Shan, J. Z. (1999). Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, *44*(5), 952–966.

Li, H., Medina, D., Vilà-Valls, J., & Closas, P. (2020). Robust variational-based kalman filter for outlier rejection with correlated measurements. *IEEE Transactions on Signal Processing*, *69*, 357–369.

Liao, J., & Berg, A. (2018). Sharpening jensen's inequality. *The American Statistician*.

Liu, S., Kanamori, T., & Williams, D. J. (2022). Estimating density models with truncation boundaries using score matching. *The Journal of Machine Learning Research*, *23*(1), 8448–8485.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, 1897–1908.

Lucas Jr, R. E. (1976). Econometric policy evaluation: A critique, In *Carnegie-rochester conference series on public policy*. North-Holland.

Lyddon, S. P., Holmes, C., & Walker, S. (2019). General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, *106*(2), 465–478.

Mahalanobis, P. C. (2018). On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, *80*, S1–S7.

Maillard, O.-A. (2019). Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds, In *Algorithmic learning theory*. PMLR.

Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 997–1022.

Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2023). Generalized bayesian inference for discrete intractable likelihood. *Journal of the American Statistical Association*, 1–11.

Morzfeld, M., & Reich, S. (2018). Data assimilation: Mathematics for merging models and data.

Murphy, K. P. (1998). Switching kalman filters.

Namoano, B., Starr, A., Emmanouilidis, C., & Cristobal, R. C. (2019). Online change detection techniques in time series: An overview, In *2019 ieee international conference on prognostics and health management (icphm)*. IEEE.

Nikiforov, I. V. (1999). Quadratic tests for detection of abrupt changes in multivariate signals. *IEEE transactions on signal processing*, *47*(9), 2534–2538.

Nikiforov, I. V. (2001). A simple change detection scheme. *Signal Processing*, *81*(1), 149–172.

Nikiforov, I. (1994). On the first order optimality of the discord detection algorithm in the vector case., (1), 87–105.

Niu, Y. S., Hao, N., & Zhang, H. (2016). Multiple change-point detection: A selective overview. *Statistical Science*, 611–623.

Pacchiardi, L. (2021). Generalizing bayesian inference. http://www.lorenzopacchiardi.me/blog/2021/generalizedBayes/

Pacchiardi, L., & Dutta, R. (2021). Generalized bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, *42*(3/4), 523–527.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1/2), 100–115.

Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of Statistics*, 206–227.

Reich, S., & Cotter, C. (2015). *Probabilistic forecasting and bayesian data assimilation*. Cambridge University Press.

Ritchie, S. (2020). *Science fictions: Exposing fraud, bias, negligence and hype in science*. Random House.

Saad, E. M., & Blanchard, G. (2021). Fast rates for prediction with limited expert advice. *Advances in Neural Information Processing Systems*, *34*, 23582–23591.

Słupiński, M. (2023). Exponential families, conjugate priors and kalman filters. Computational Intelligence Research Group, Institute of Computer Science, University of Wroclaw. https://ii.uni.wroc.pl/~lipinski/ADM2023/MSl/ADM__lecture_3.pdf

Solo, V. (1996). Stability of the kalman filter with stochastic time-varying parameters, In *Proceedings of 35th ieee conference on decision and control*. IEEE.

Sorenson, H. W., & Alspach, D. L. (1971). Recursive bayesian estimation using gaussian sums. *Automatica*, *7*(4), 465–479.

Stannat, W. (2023). Lecture notes in stochastic filtering. Technical University Berlin.

Tam, P., & Moore, J. (1977). A gaussian sum approach to phase and frequency estimation. *IEEE Transactions on Communications*, *25*(9), 935–942.

Tang, H., Han, H., Zhang, S., & Feng, W. (2024). A generalized t-distribution-based kernel adaptive filtering algorithm. *IEEE Transactions on Circuits and Systems II: Express Briefs*.

Tartakovsky, A. G. (2009). Asymptotic optimality in bayesian changepoint detection problems under global false alarm probability constraint. *Theory of Probability & Its Applications*, *53*(3), 443–466.

Tartakovsky, A., Nikiforov, I., & Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press.

Umarov, S., Tsallis, C., & Steinberg, S. (2008). On aq-central limit theorem consistent with nonextensive statistical mechanics. *Milan journal of mathematics*, *76*(1), 307–328.

Van den Burg, G. J., & Williams, C. K. (2020). An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*(3), 426–482.

Wald, A. (1947). Foundations of a general theory of sequential decision functions. *Econometrica, Journal of the Econometric Society*, 279–313.

Wald, A. (1992). Sequential tests of statistical hypotheses, In *Breakthroughs in statistics: Foundations and basic theory*. Springer.

Wang, H., Li, H., Fang, J., & Wang, H. (2018). Robust gaussian kalman filter with outlier detection. *IEEE Signal Processing Letters*, *25*(8), 1236–1240.

Willsky, A. S. (1976). A survey of design methods for failure detection in dynamic systems. *Automatica*, *12*(6), 601–611.

Willsky, A. S., & Jones, H. L. (1974). A generalized likelihood ratio approach to state estimation in linear systems subjects to abrupt changes, In *1974 ieee conference on decision and control including the 13th symposium on adaptive processes*. IEEE.

Willsky, A., & Jones, H. (1976). A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic control*, *21*(1), 108–112.

Wu, P.-S., & Martin, R. (2023). A comparison of learning rate selection methods in generalized bayesian inference. *Bayesian Analysis*, *18*(1), 105–132.

Xie, L., Zou, S., Xie, Y., & Veeravalli, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, *2*(2), 494–514.

Zellner, A. (1988). Optimal information processing and bayes's theorem. *The American Statistician*, *42*(4), 278–280.

Zhang, M., Key, O., Hayes, P., Barber, D., Paige, B., & Briol, F.-X. (2022). Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*.

Zhen-Wei, Z., & Hai-Tao, F. (2013). L2-stability of discrete-time kalman filter with random coefficients under incorrect covariance. *Acta Automatica Sinica*, *39*(1), 43–52.

Zhu, X., Soh, Y. C., & Xie, L. (2002). Design and analysis of discrete-time robust kalman filters. *Automatica*, *38*(6), 1069–1077.

Appendix

A. Additional Material

A.1. Chapter 2

A.1.1. Experiment A: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

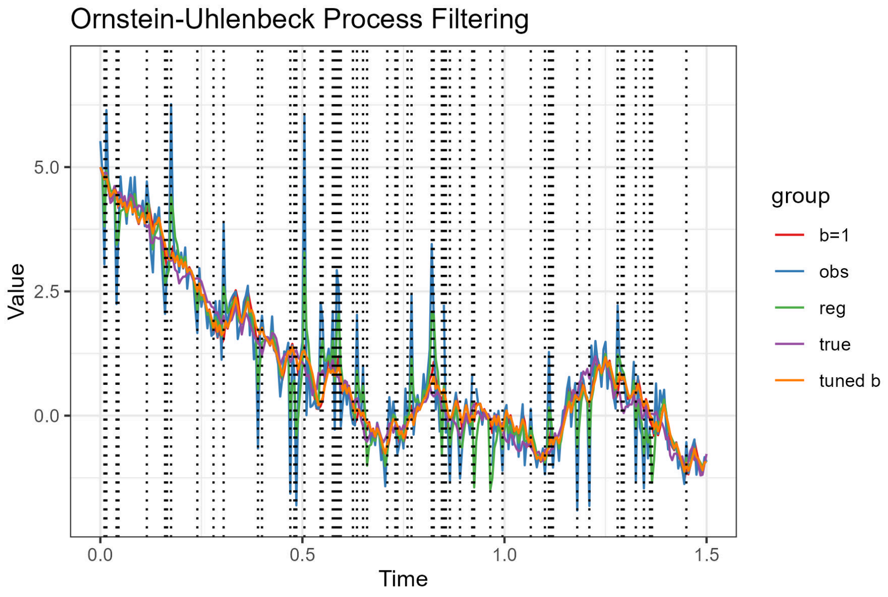


Figure A.1.: Graph of the simulated signal and contaminated observations as well as the analysis/posterior mean estimates of the Kalman filter and the $\mathcal{D}_w$-Kalman filter for $\beta = 1$ and tuned β^*. Dotted lines signal instances of observation contamination.

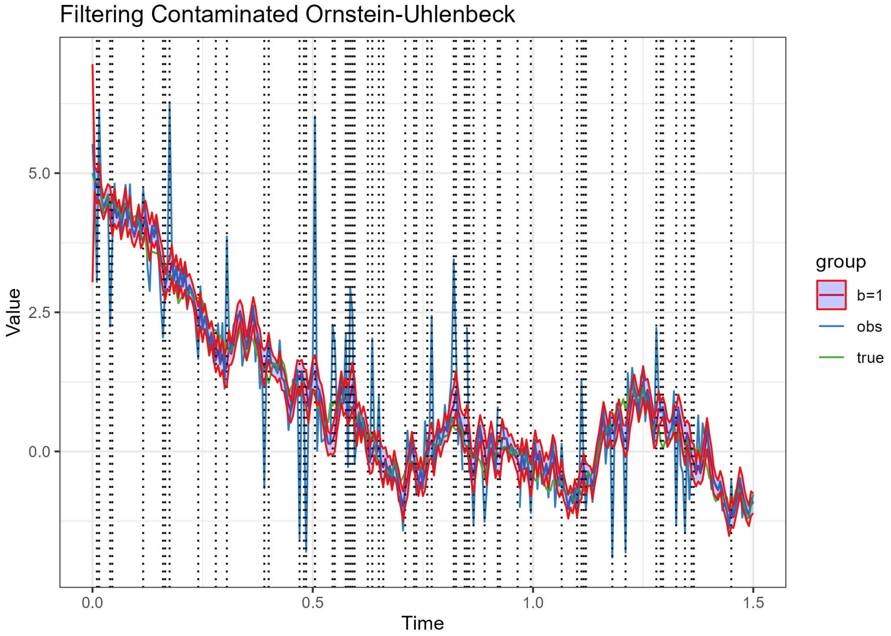


Figure A.2.: Graph of the simulated signal, contaminated observations and the tuned $\mathcal{D}_w$-Kalman filter with 95%-CI. Dotted lines signal instances of observation contamination.

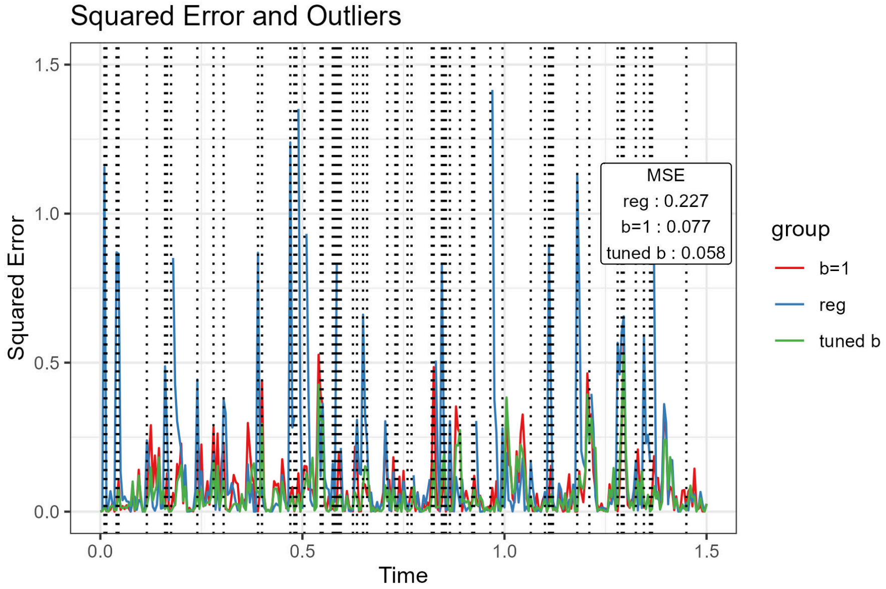


Figure A.3.: Graph of the squared error for each mean estimate with the MSE over all time steps. Dotted lines signal instances of observation contamination.

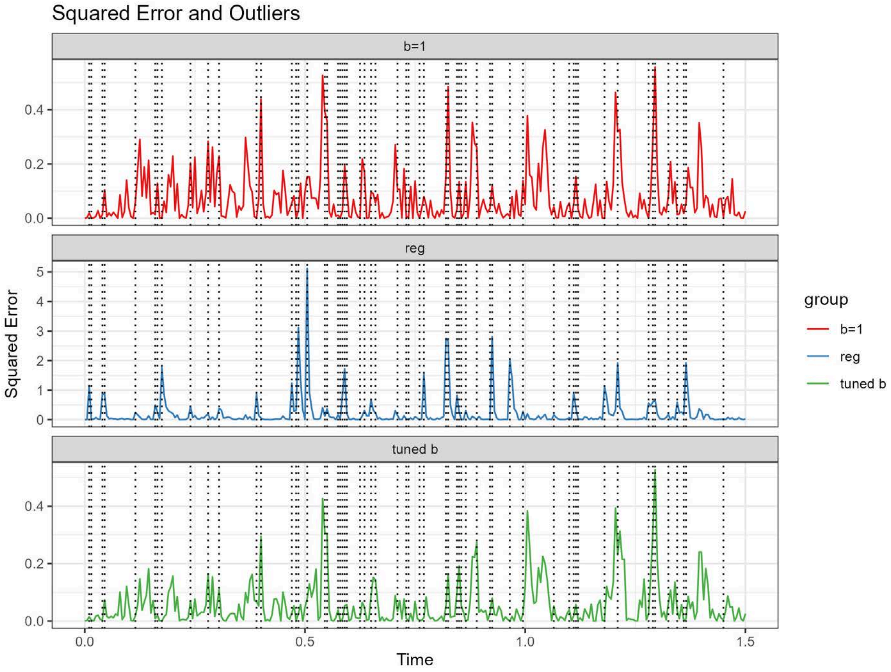


Figure A.4.: Separate graphs of the squared error for each mean estimate respectively. Note the difference in scale. Dotted lines signal instances of observation contamination.

A.1.2. Experiment B: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

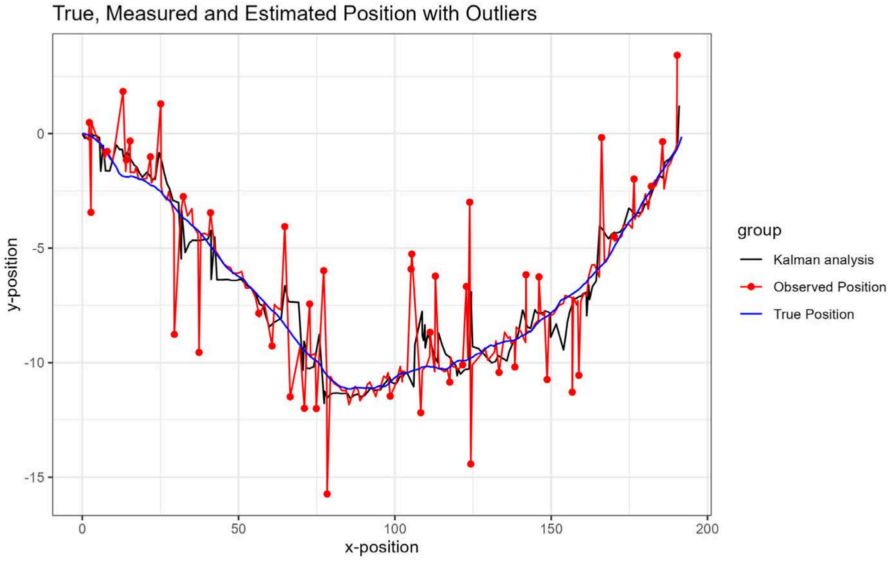


Figure A.5.: Simulated object trajectory, contaminated observed trajectory, analysis and forecast mean of the regular Kalman filter.

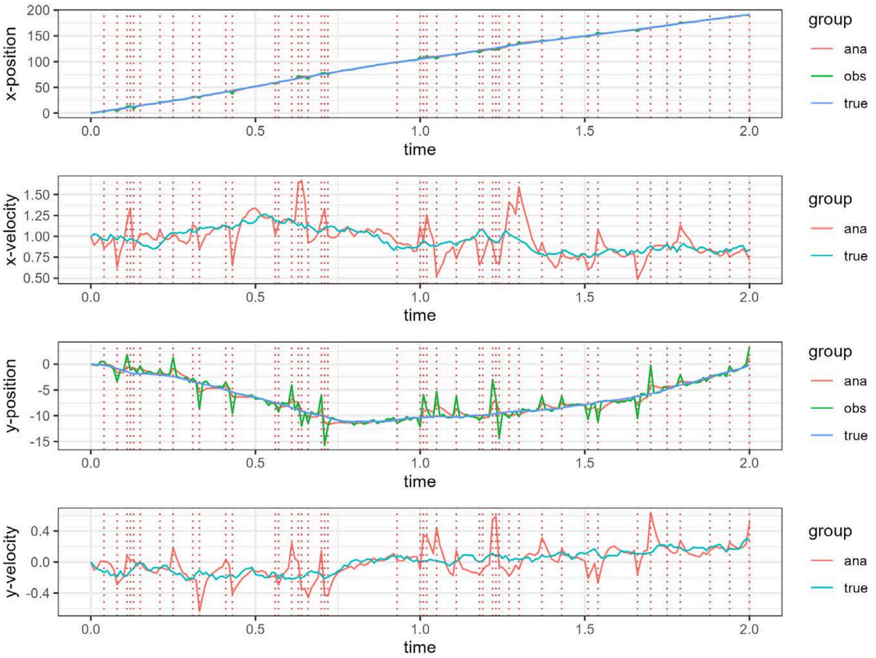


Figure A.6.: Side-by-side graph comparison of the simulated signal, contaminated observations, analysis and forecast mean of the regular Kalman filter for each dimension.

A.1.3. Experiment C: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

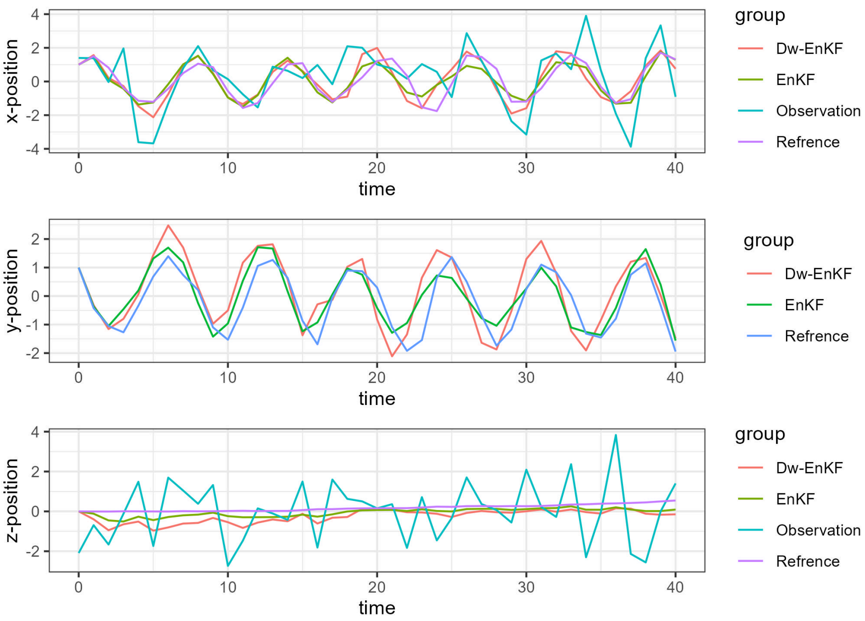


Figure A.7.: Side-by-side graph comparison of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations.

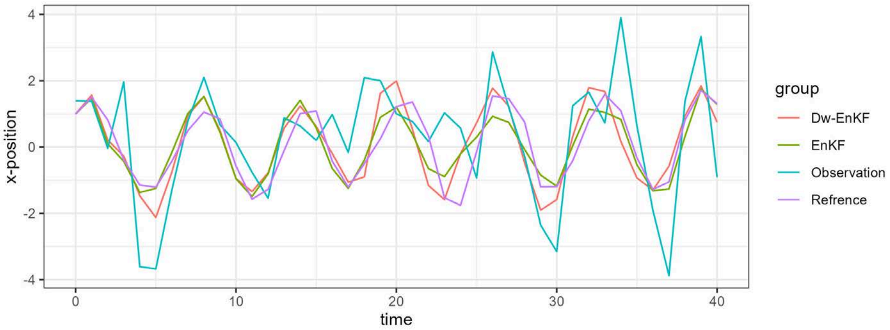


Figure A.8.: First dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations.

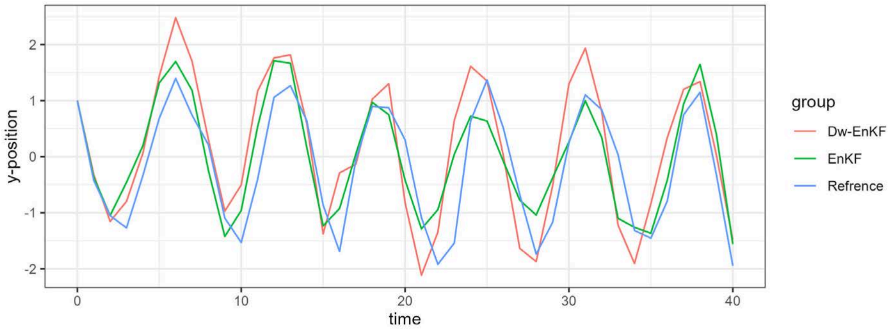


Figure A.9.: Second dimension of the simulated signal, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations. This dimension was not observed directly.

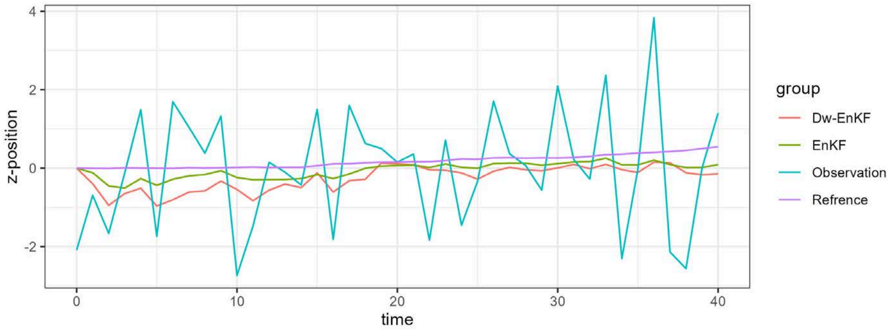


Figure A.10.: Third dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations.

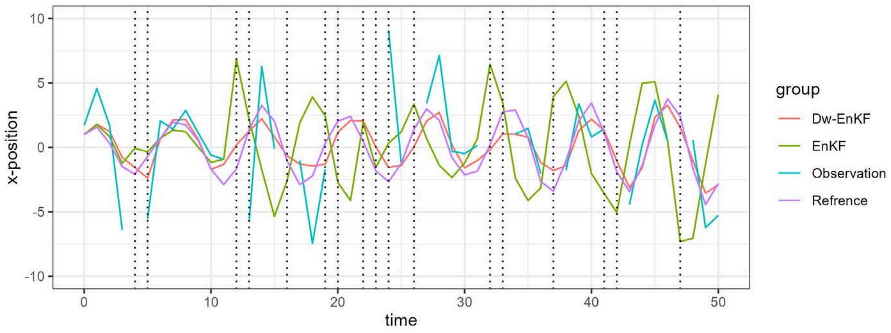


Figure A.11.: First dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for contaminated observations.

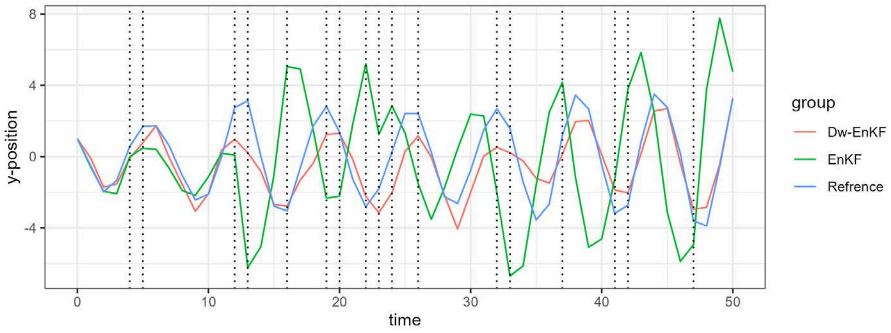


Figure A.12.: Second dimension of the simulated signal, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for contaminated observations. This dimension was not observed directly.

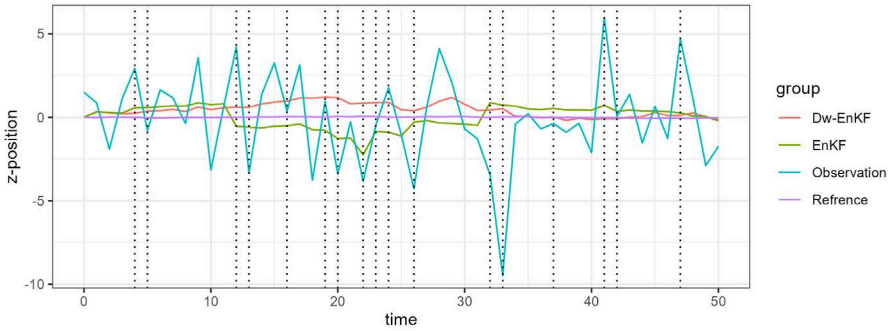


Figure A.13.: Third dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for contaminated observations.

A.1.4. Experiment D: Additional Graphs

The model, simulation and parameter choices are described in the respective section

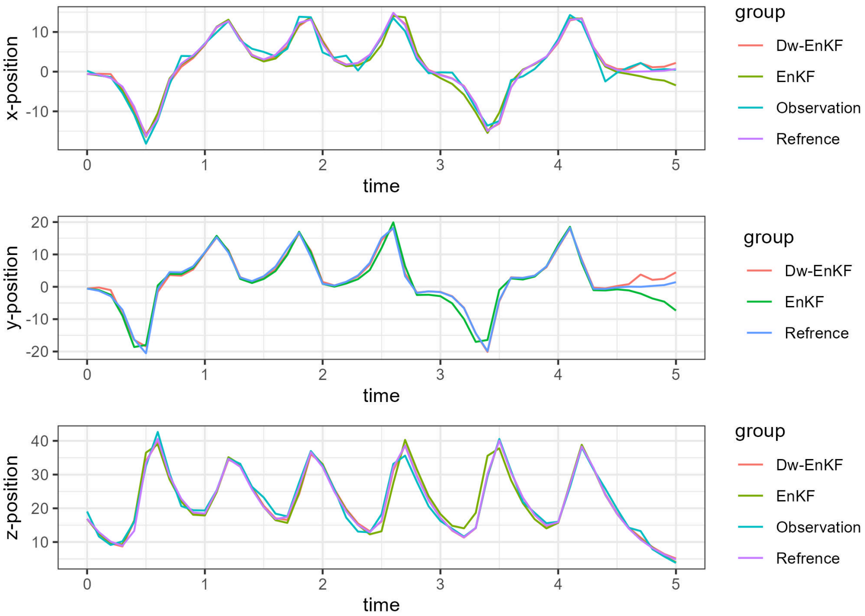


Figure A.14.: Side-by-side graph comparison of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations.

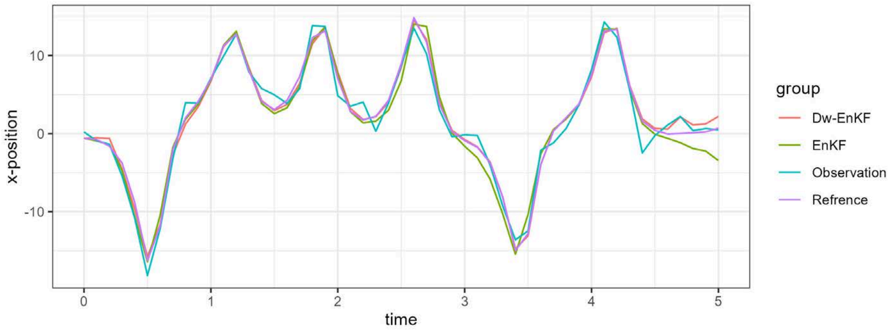


Figure A.15.: First dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations.

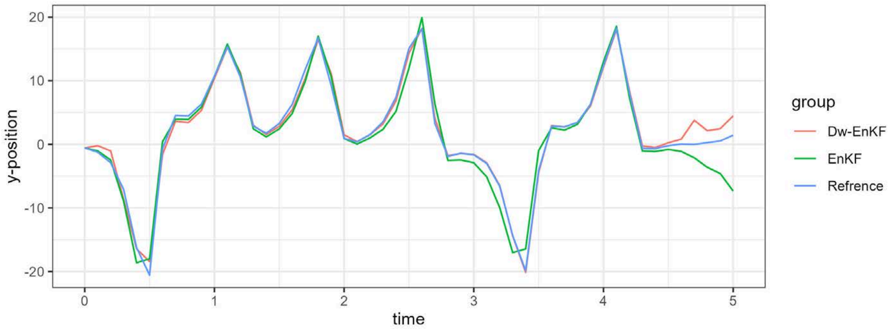


Figure A.16.: Second dimension of the simulated signal, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations. This dimension was not observed directly.

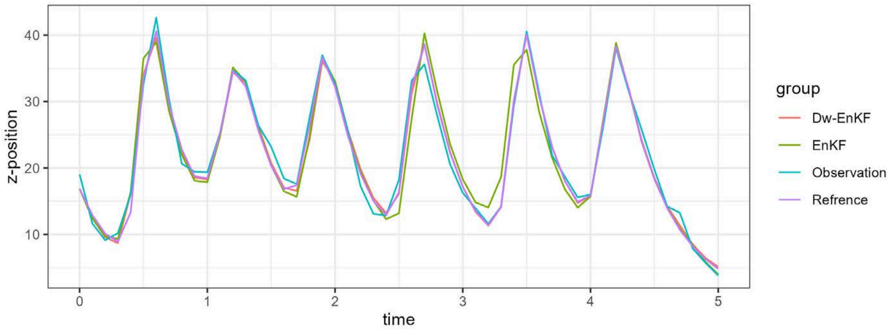


Figure A.17.: Third dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for uncontaminated observations.

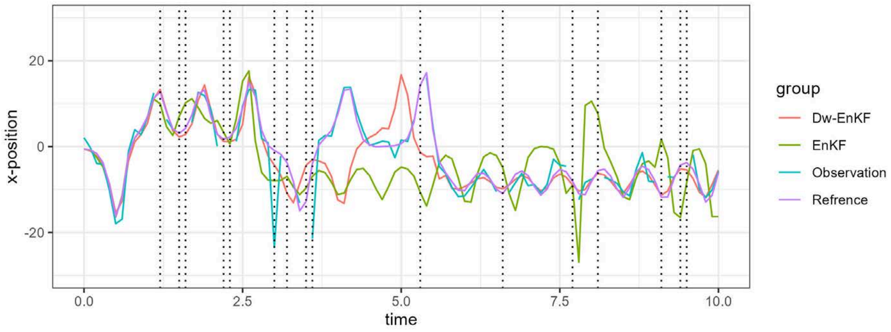


Figure A.18.: First dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for contaminated observations.

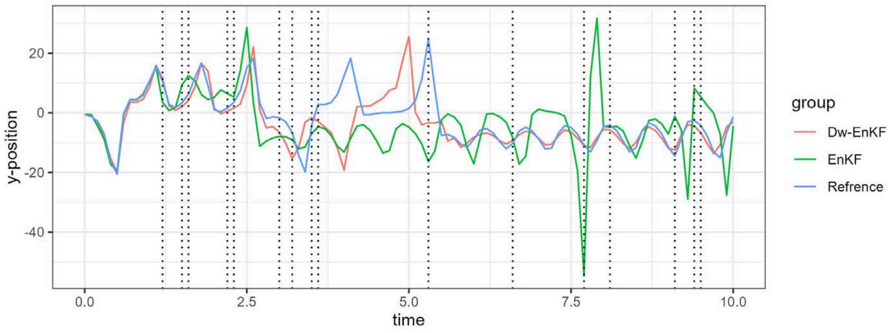


Figure A.19.: Second dimension of the simulated signal, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for contaminated observations. This dimension was not observed directly.

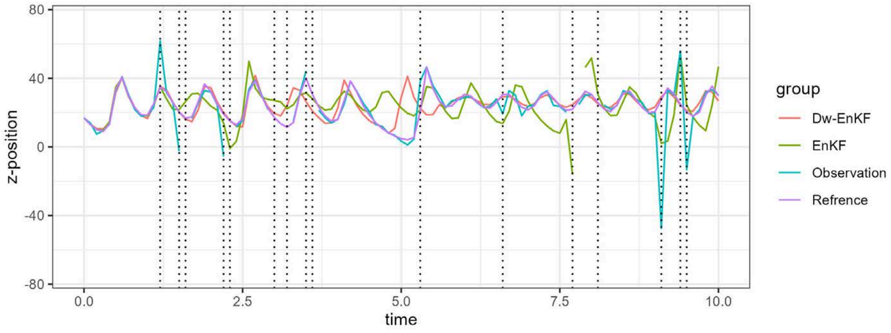


Figure A.20.: Third dimension of the simulated signal, generated observations, the KF analysis/posterior mean estimates and the tuned $\mathcal{D}_w$-KF analysis/posterior for contaminated observations.

A.2. Chapter 3

A.2.1. Algorithms

A.2.2. Experiment A: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

Algorithm 3 The BOCPD GMM-Kalman Filter

Input:

- change point prior $\varphi_n \in (0, 1)$ (for a geometric hazard function)
- scenario initial condition $p(y_k | y_{1:(k-1)}, k)$
- Kalman filter requirements (initial condition $p(x_0 | k = 0) \sim n(x_0; m_0, P_0)$, signal model, observation model)
- initial weight $\kappa_{0,0} = 1$

Output:

- scenario posteriors $\tilde{\kappa}_{k,n}$
- aggregated signal forecast $p_\kappa(x_n | y_{1:(n-1)})$ at time n and
- aggregated signal state $p_\kappa(x_n | y_{1:n})$ at time n

```
for n=1,2,... do
    initialize new scenario forecast: $p(x_k | y_{1:k-1}, k)$ for $k := n$
    initialize new scenario weight forecast: $\kappa^f_{k,n} = \varphi_n \sum_{s=1}^{n-1} \kappa_{s,n-1}$ for $k := n$
    for k=0,1,2,...,n-1 do
        forecast step for all previous scenarios: $p(x_n | y_{1:n-1}, k)$
    end for
    for k=0,1,...,n-1 do
        forecast scenario weights: $\kappa^f_{k,n} = (1 - \varphi_n)\kappa_{k,n-1}$
    end for
    for k=0,1,...,n do
        normalize forecast weights: $\tilde{\kappa}^f_{k,n} = \frac{\kappa^f_{k,n}}{\sum_{s=0}^{n} \kappa^f_{s,n}}$
    end for
    aggregate signal forecast: $p_\kappa(x_n | y_{1:(n-1)}) = \sum_{s=0}^{n} \tilde{\kappa}^f_{s,n} p(x_n | y_{1:(n-1)}, s)$
    receive observation: $y_n$
    initialize new scenario weight: $\kappa_{k,n} = \varphi_n \sum_{s=0}^{n-1} p(y_n | y_{1:(n-1)}, s)\kappa_{s,n-1}$
    for k=0,1,...,n-1 do
        update scenario weights: $\kappa_{k,n} = (1 - \varphi_n)p(y_n | y_{1:(n-1)}, k)\kappa_{k,n-1}$
    end for
    for k=0,1,...,n do
        normalize weights: $\tilde{\kappa}_{k,n} = \frac{\kappa_{k,n}}{\sum_{s=1}^{n} \kappa_{s,n}}$
    end for
    for k=1,2,...,n do
        analysis step for all scenarios: $p(x_n | y_{1:n}, k)$
    end for
    aggregate signal state estimate: $p_\kappa(x_n | y_{1:n}) = \sum_{s=0}^{n} \tilde{\kappa}_{s,n} p(x_n | y_{1:n}, s)$
end for
```

Algorithm 4 The CR-BOCPD GMM-Kalman Filter

```
    Input:
        • scenario initial condition $p(y_k|y_{1_(k-1)}, k)$
        • Kalman filter requirements (initial condition $p(x_0|k = 0) \sim n(x_0; m_0, P_0)$, signal model, observation model)
        • initial restart $r = 0$
        • initial tuning parameter $\beta_{r,0,0} := 1$
        • tuning parameter computation $\beta_{r,k,n} := \exp(-c) \sim \alpha_0$ for $c \sim |\log(\alpha_0)|$(for false alarm rate $\alpha_0$)
        • initial weight $\kappa_{r,0,0} = 1$
    Output:
        • scenario posterior $\tilde{\kappa}_{r,k,n}$
        • last detected instance of change $r$
        • aggregated signal forecast $p_\kappa(x_n|y_{1:(n-1)})$ at time $n$ and
        • aggregated signal state $p_\kappa(x_n|y_{1:n})$ at time $n$
for n=1,2,... do
    initialize new scenario forecast: $p(x_k|y_{1:k-1}, k)$ for $k := n$
    initialize new scenario weight forecast: $\kappa^f_{r,k,n} = \beta_{r,k,n} \prod_{s=r}^{n-1} p(y_s|y_{1:(s-1)}, r)$ for $k := n$
    for k=r,r+1,...,n-1 do
        forecast step for all previous scenarios: $p(x_n|y_{1:n-1}, k)$
    end for
    for k=r,r+1,...,n-1 do
        forecast scenario weights: $\kappa^f_{r,k,n} = \frac{\beta_{r,k,n}}{\beta_{r,k,n-1}} \kappa_{r,k,n-1}$
    end for
    for k=r,r+1,...,n do
        normalize forecast weights: $\tilde{\kappa}^f_{r,k,n} = \frac{\kappa^f_{r,k,n}}{\sum_{s=0}^{n} \kappa^f_{r,s,n}}$
    end for
    aggregate signal forecast: $p_\kappa(x_n|y_{1:(n-1)}) = \sum_{s=0}^{n} \tilde{\kappa}^f_{r,s,n} p(x_n|y_{1:(n-1)}, s)$
    receive observation: $y_n$
    initialize new scenario weight: $\kappa_{r,k,n} = \beta_{r,k,n} p(y_k|y_{1:(k-1)}, k) \prod_{s=r}^{n-1} p(y_s|y_{1:(s-1)}, r)$ for $k := n$
    for k=r,r+1,...,n-1 do
        update scenario weights: $\kappa_{r,k,n} = \frac{\beta_{r,k,n}}{\beta_{r,k,n}} p(y_n|y_{1:(n-1)}, k) \kappa_{r,k,n-1}$
    end for
    check restart rule $r = \max\{k \geq r : \kappa_{r,k,n} \geq \kappa_{r,r,n}\}$
    for k=r,r+1,...,n do
        normalize weights: $\tilde{\kappa}_{r,k,n} = \frac{\kappa_{r,k,n}}{\sum_{s=r}^{n} \kappa_{r,s,n}}$
    end for
    for k=r,r+1,...,n do
        analysis step for all scenarios: $p(x_n|y_{1:n}, k)$
    end for
    aggregate signal state estimate: $p_\kappa(x_n|y_{1:n}) = \sum_{s=r}^{n} \tilde{\kappa}_{r,s,n} p(x_n|y_{1:n}, s)$
end for
```

Algorithm 5 A Recursive χ^2-Ensemble Filter

```
Input:
  • Kalman filter requirements (initial condition $p(x_0|k = 0) \sim n(x_0; m_0, P_0)$, signal model, observation model)
  • ensemble size $M > 0$
  • minimal deviation of interest $b^2 > 0$ and threshold $c > 0$
Output:
  • Gaussian approx. of signal forecast $\hat{p}(x_n|y_{1:(n-1)})$ at time $n$ and
  • Gaussian approx. signal state $\hat{p}(x_n|y_{1:n})$ at time $n$
Start:
  • draw starting ensemble $\{x_0^{a,(l)}\}_{1\leq l\leq M}$,
  • initialize test statistics $\{\hat{S}_0^{(l)}\}_{1\leq l\leq M}$,
  • innovation residual $\{\hat{\Gamma}_0^{(l)}\}_{1\leq l\leq M}$,
  • counter $\{\tilde{n}_0^{(l)}\}_{1\leq l\leq M}$ and set to $0$
for n=1,2,... do
    forecast propagation of all ensemble members: $\{x_(n-1)^{a,(l)}\}_{1\leq l\leq M} \rightarrow \{x_n^{f,(l)}\}_{1\leq l\leq M}$
    estimate empirical forecast mean $\hat{m}_n^f$ and covariance $\hat{P}_n^f$ for $\hat{p}(x_n|y_{1:(n-1)}) = n(x_n; \hat{m}_n^f, \hat{P}_n^f)$
    receive observation: $y_n$
    for l=0,1,...,M do
        recursive update of counter, innovation residual and test statistics via
            $$\begin{aligned} \tilde{n}_n^{(l)} &= \mathbf{1}\{\hat{S}_{n-1}^{(l)}\}\tilde{n}_{n-1}^{(l)} + 1 \\ \hat{\Gamma}_n^{(l)} &= \mathbf{1}\{\hat{S}_{n-1}^{(l)}\}\hat{\Gamma}_{n-1}^{(l)} + \Sigma_n^{-\frac{1}{2}}(H_n x_n^{f,(l)} - y_n) \\ \hat{S}_n^{(l)} &= -\tilde{n}_n^{(l)}\frac{d^2}{n} + d(\hat{\Gamma}_n^{(l)})^T\hat{\Gamma}_n^{(l)} \end{aligned} \tag{A.1}$$
    end for
    eliminate all ensemble members with $\hat{S}_n^{(l)} > c$
    resample from remaining ensemble and duplicate the corresponding test statistics
    analysis step for all ensemble members: $\{x_n^{f,(l)}\}_{1\leq l\leq M} \rightarrow \{x_n^{a,(l)}\}_{1\leq l\leq M}$
    estimate empirical analysis mean $\hat{m}_n^a$ and covariance $\hat{P}_n^a$ for $\hat{p}(x_n|y_{1:n}) = n(x_n; \hat{m}_n^a, \hat{P}_n^a)$
end for
```

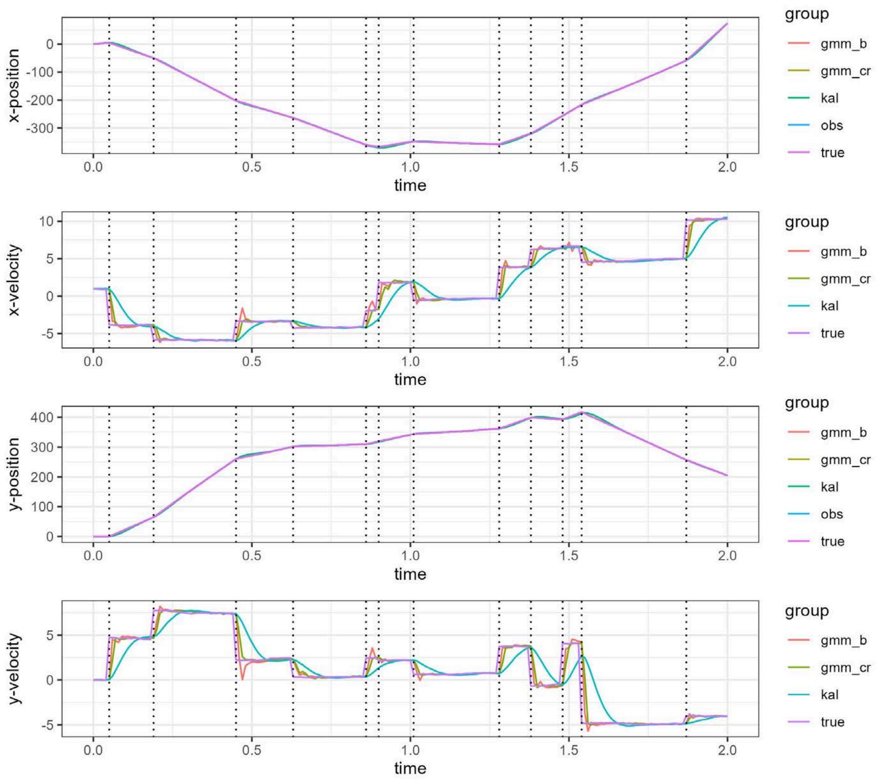


Figure A.21.: Side-by-side of analysis means and true trajectory for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

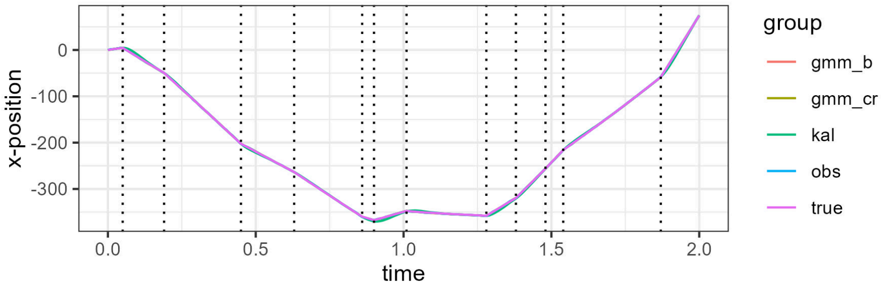


Figure A.22.: Analysis means and x-position for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

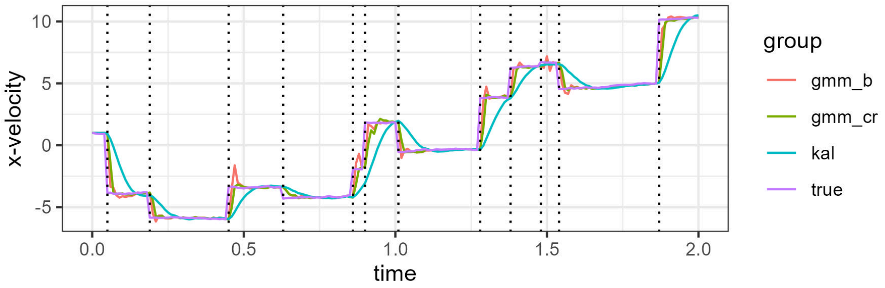


Figure A.23.: Analysis means and x-velocity for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

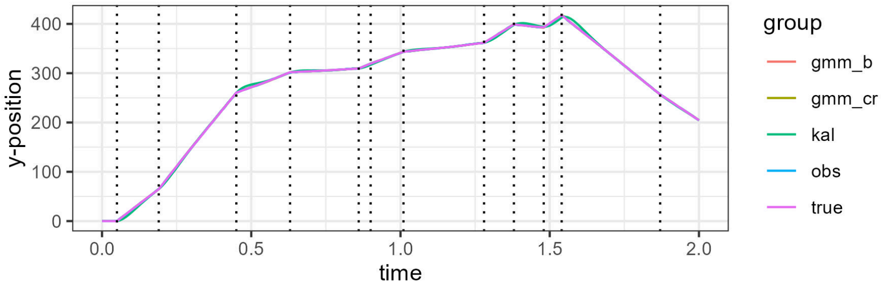


Figure A.24.: Analysis means and y-position for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

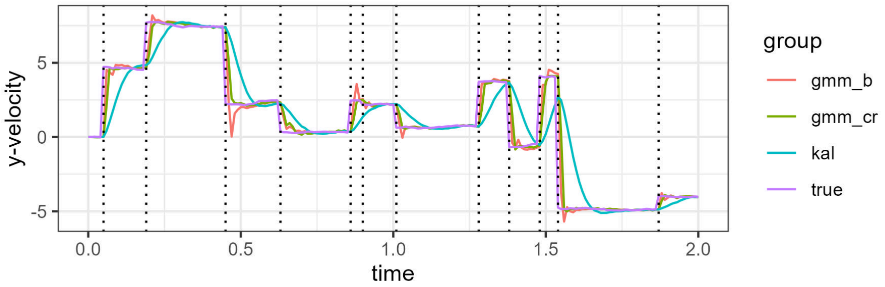


Figure A.25.: Analysis means and y-velocity for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

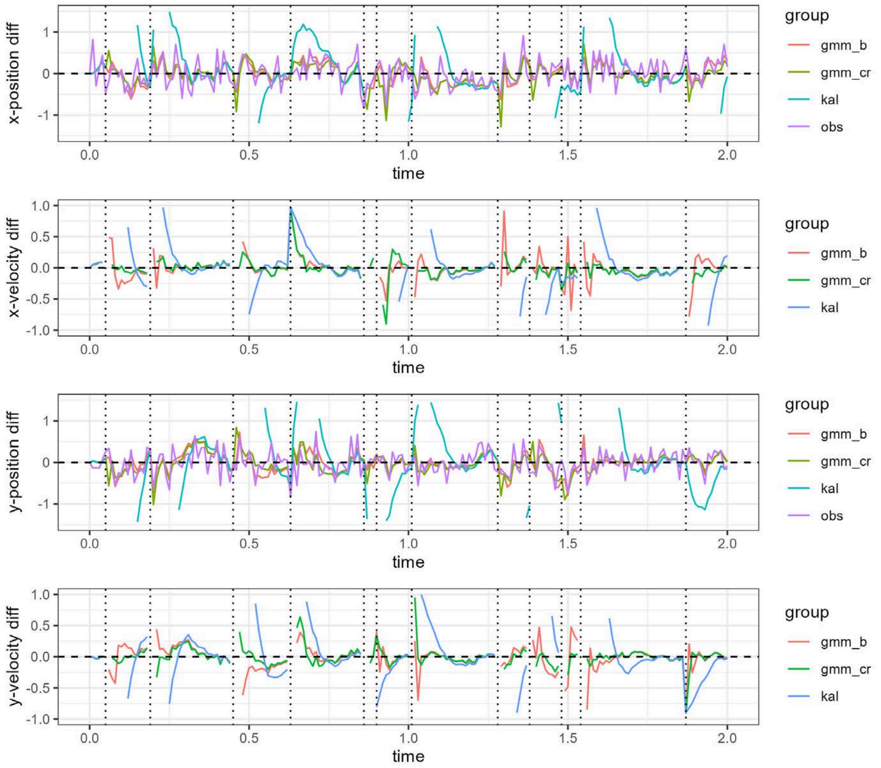


Figure A.26.: Side-by-side difference of analysis means and true trajectory for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

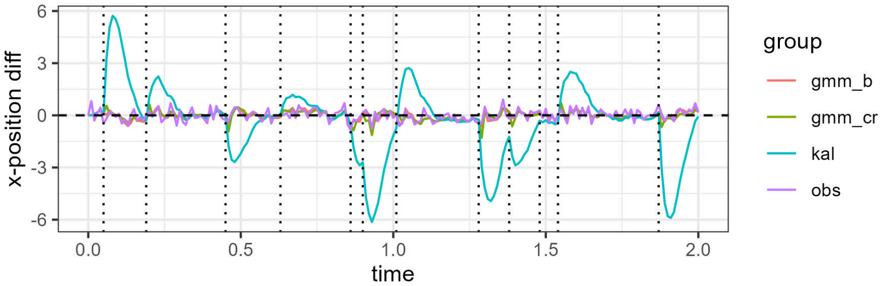


Figure A.27.: Difference of analysis means and x-position for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

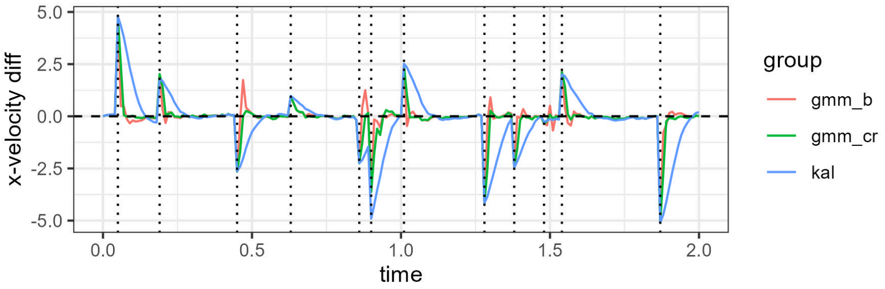


Figure A.28.: Difference of analysis means and x-velocity for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

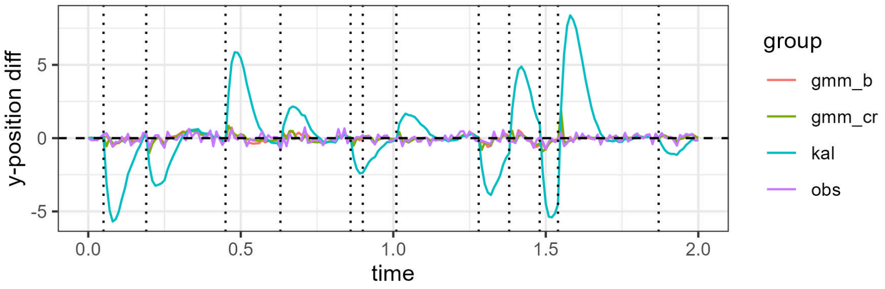


Figure A.29.: Difference of analysis means and y-position for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

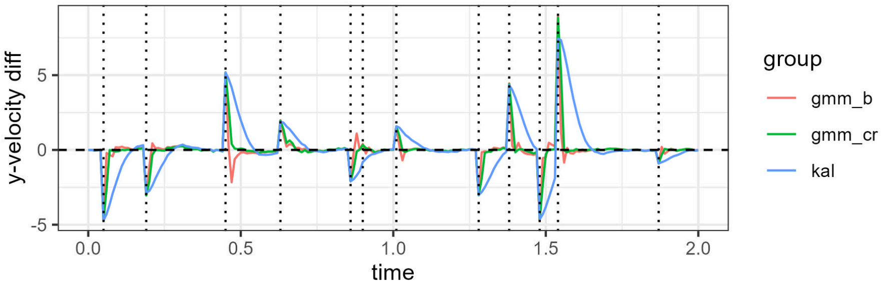


Figure A.30.: Difference of analysis means and y-velocity for the regular KF, the BOCPD GMM-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

A.2.3. Experiment B: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

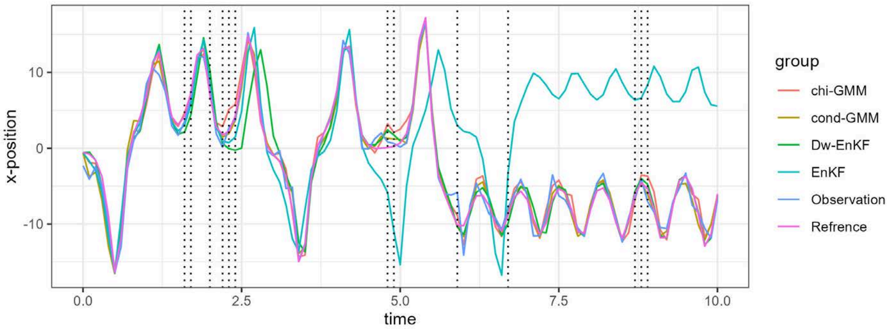


Figure A.31.: First variable of the simulated signal, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF with conditional probabilities and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion. Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

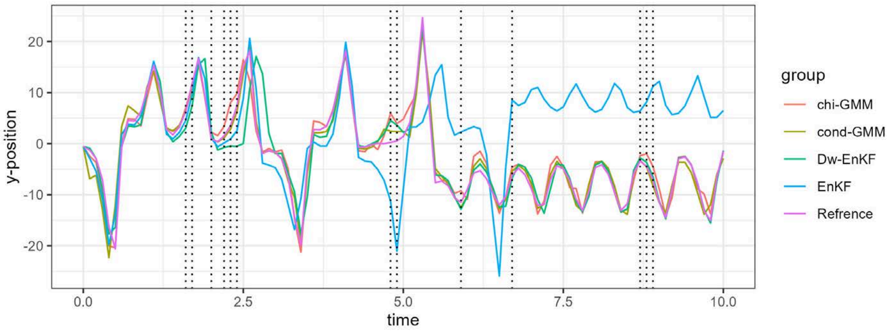


Figure A.32.: Second variable of the simulated signal, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF with conditional probabilities and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion. Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

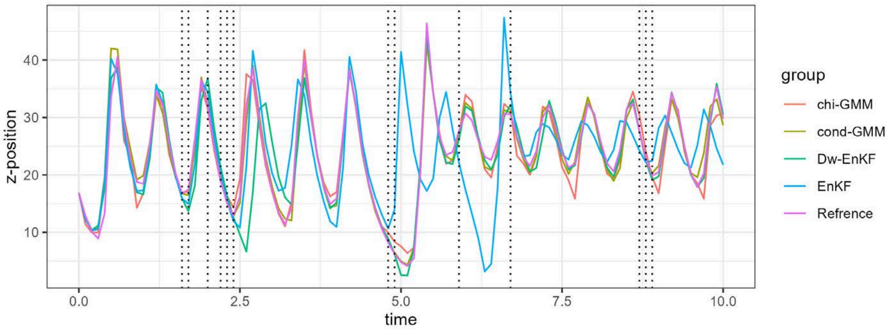


Figure A.33.: Third variable of the simulated signal, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF with conditional probabilities and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion. Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

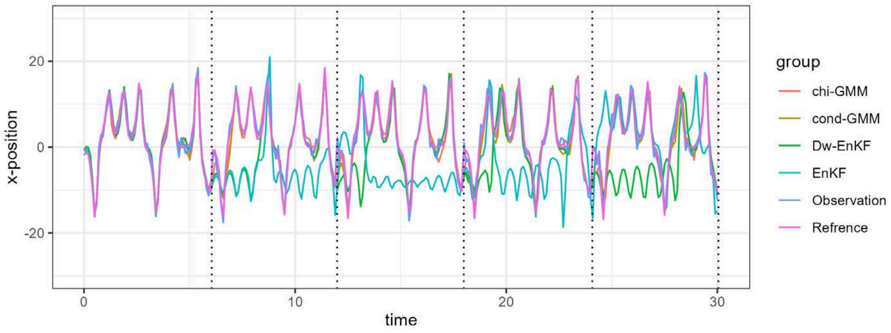


Figure A.34.: First variable of the simulated signal with jumps, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF with conditional probabilities and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion. Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

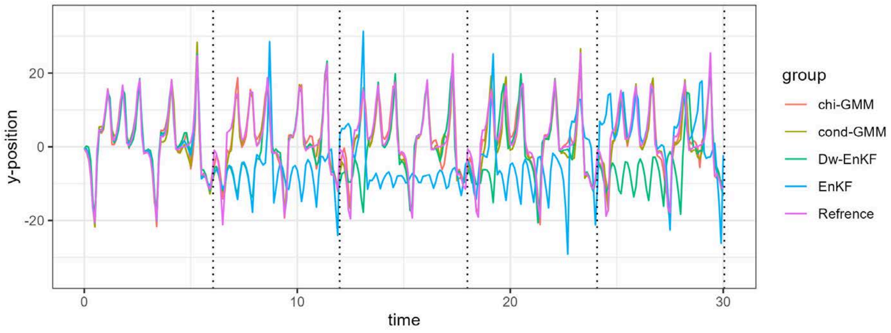


Figure A.35.: Second variable of the simulated signal with jumps, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF with conditional probabilities and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion. Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

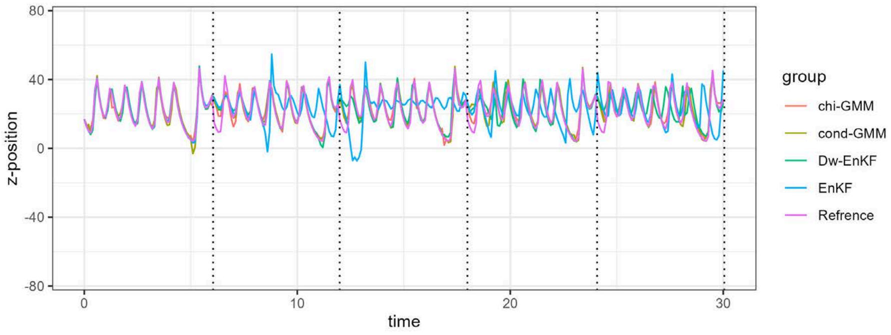


Figure A.36.: Third variable of the simulated signal with jumps, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF with conditional probabilities and the CR-BOCPD GMM-EnKF with χ^2-GLR criterion. Dotted lines indicate restarts of the conditional CR-BOCPD GMM-EnKF.

A.3. Chapter 4

A.3.1. Experiment A: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

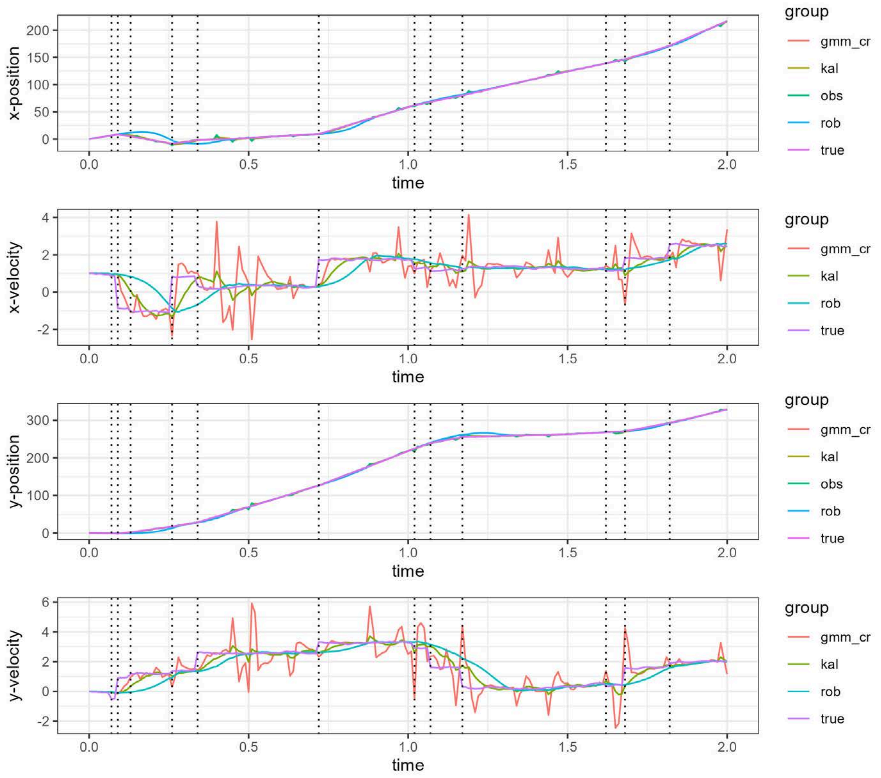


Figure A.37.: Side-by-side of analysis means and true trajectory for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

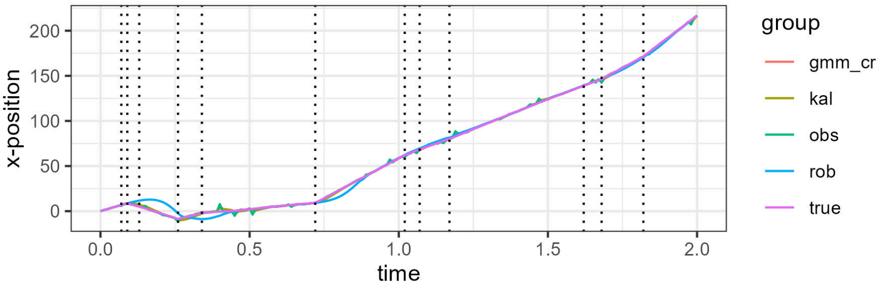


Figure A.38.: Analysis means and x-position for the regular regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

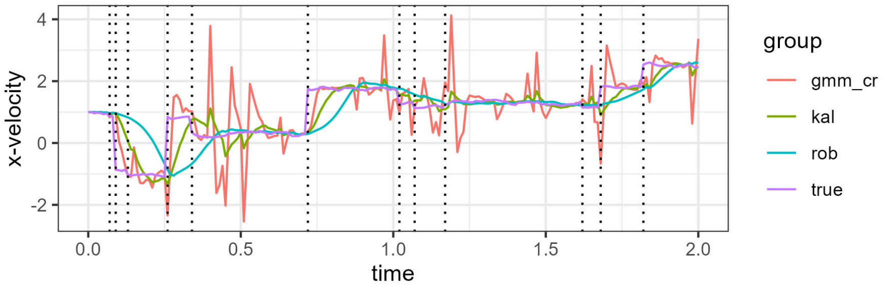


Figure A.39.: Analysis means and x-velocity for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

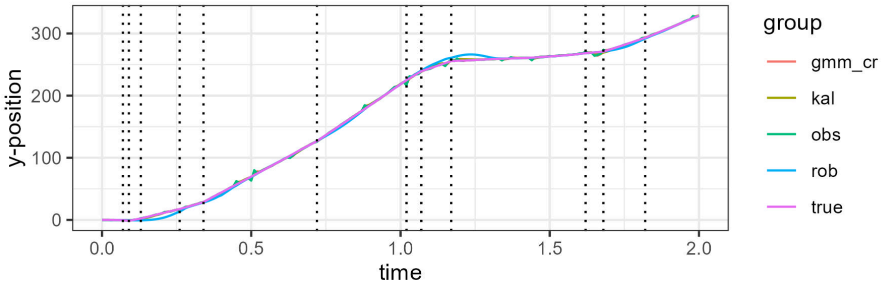


Figure A.40.: Analysis means and y-position for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

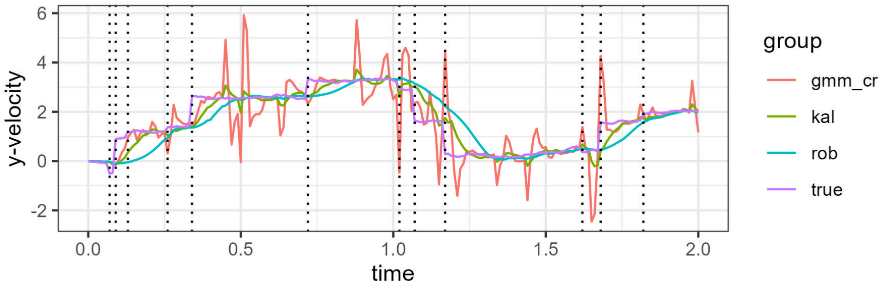


Figure A.41.: Analysis means and y-velocity for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

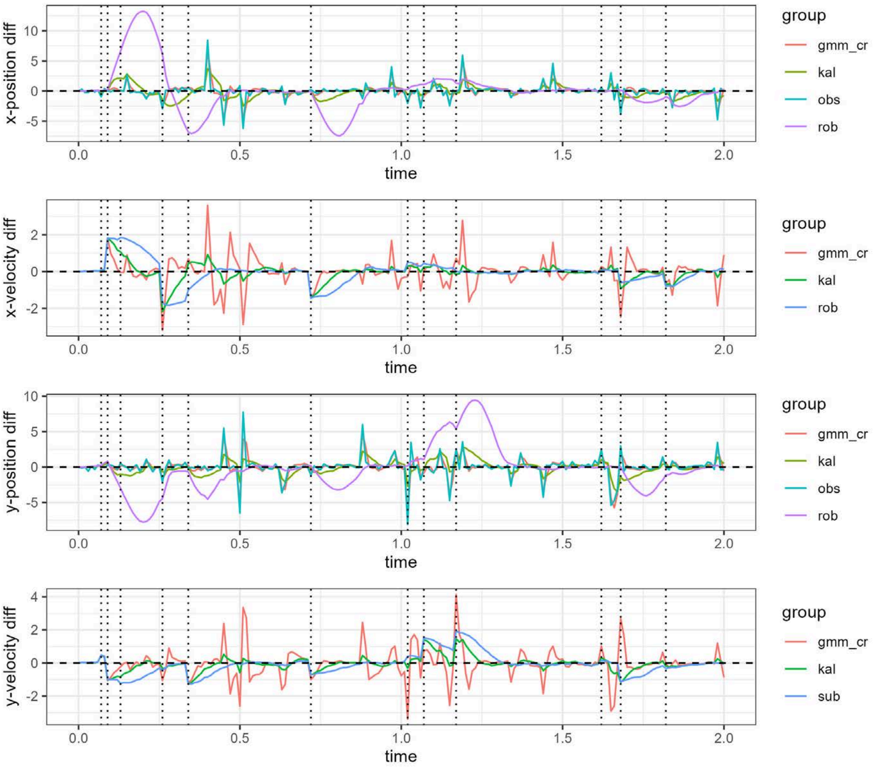


Figure A.42.: Side-by-side difference of analysis means and true trajectory for theregular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

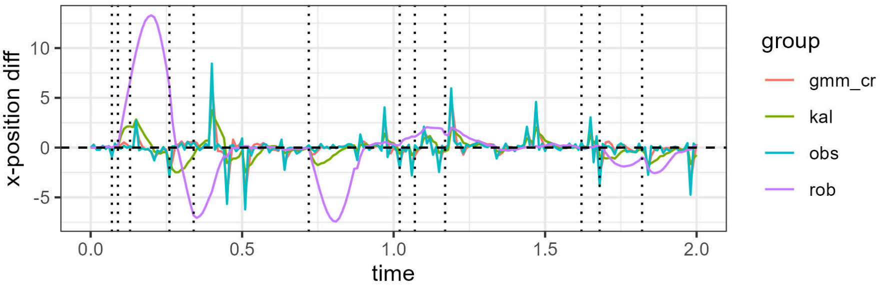


Figure A.43.: Difference of analysis means and x-position for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

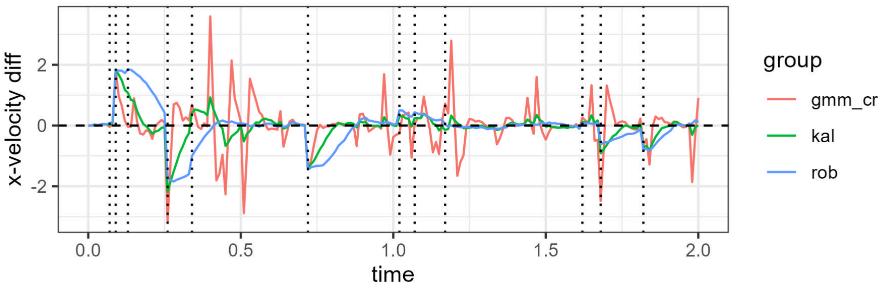


Figure A.44.: Difference of analysis means and x-velocity for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

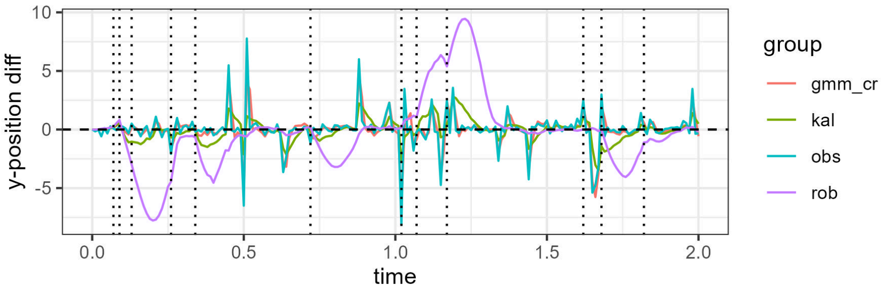


Figure A.45.: Difference of analysis means and y-position for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

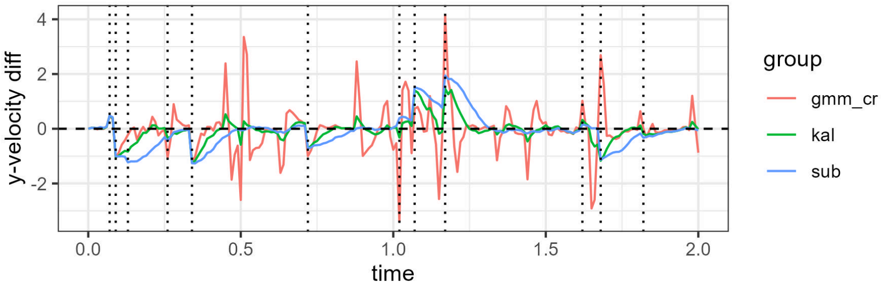


Figure A.46.: Difference of analysis means and y-velocity for the regular KF, $\mathcal{D}_w$-KF and CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

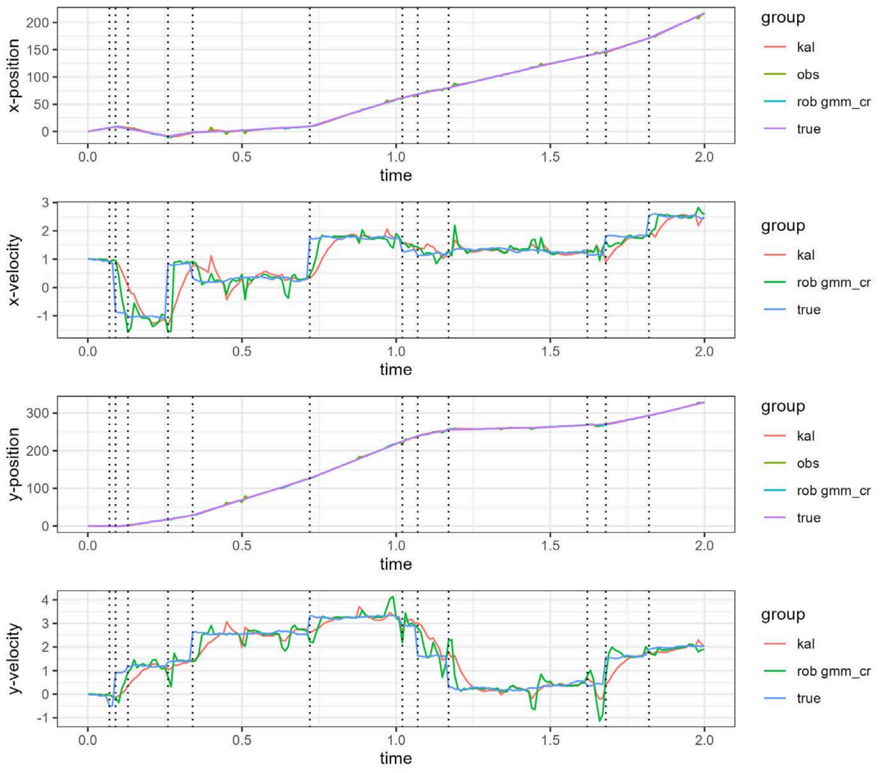


Figure A.47.: Side-by-side of analysis means and true trajectory for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

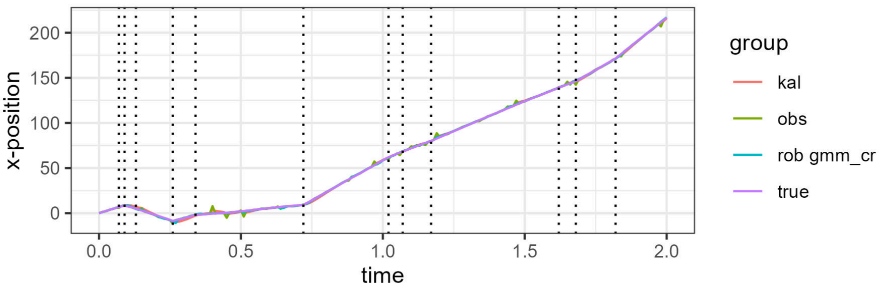


Figure A.48.: Analysis means and x-position for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

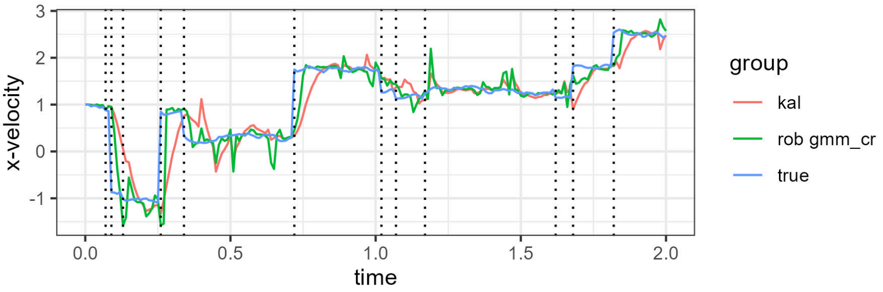


Figure A.49.: Analysis means and x-velocity for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

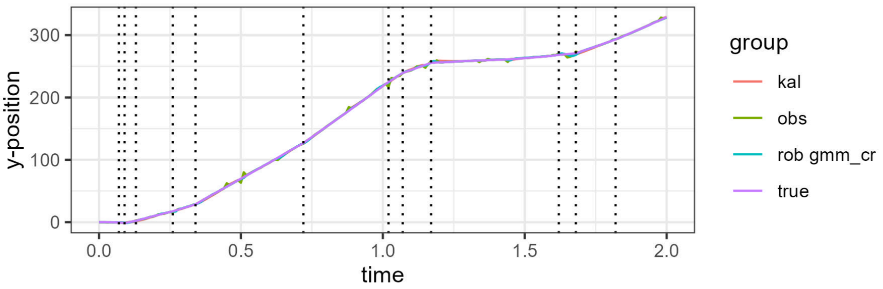


Figure A.50.: Analysis means and y-position for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

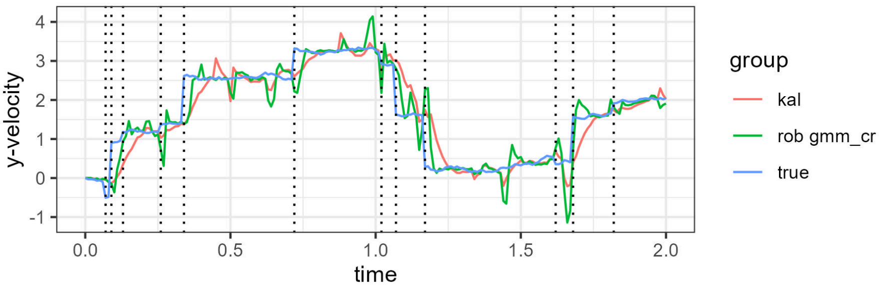


Figure A.51.: Analysis means and y-velocity for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

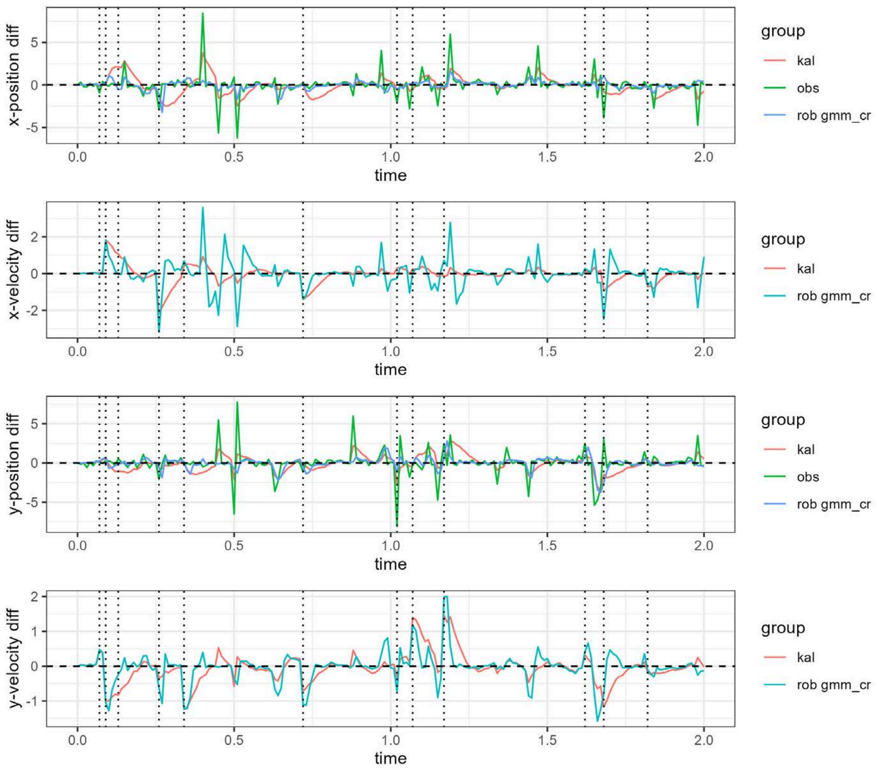


Figure A.52.: Side-by-side difference of analysis means and true trajectory for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

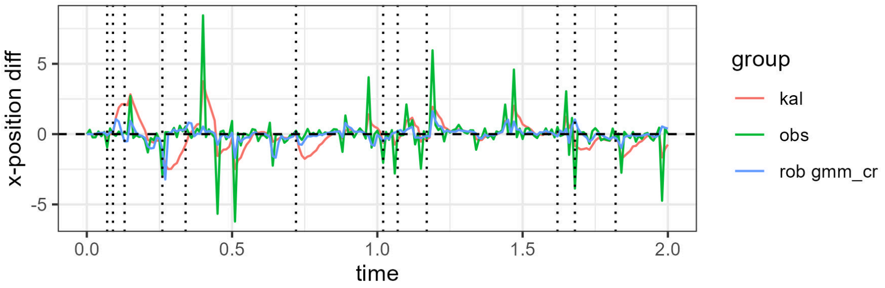


Figure A.53.: Difference of analysis means and x-position for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

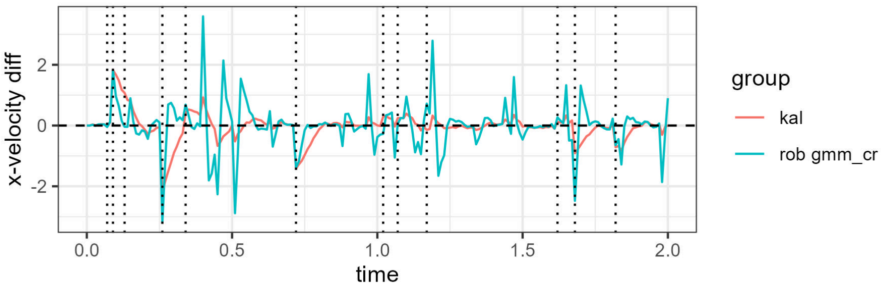


Figure A.54.: Difference of analysis means and x-velocity for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

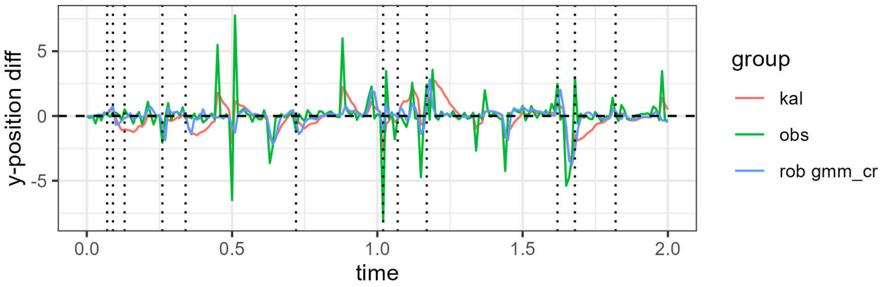


Figure A.55.: Difference of analysis means and y-position for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

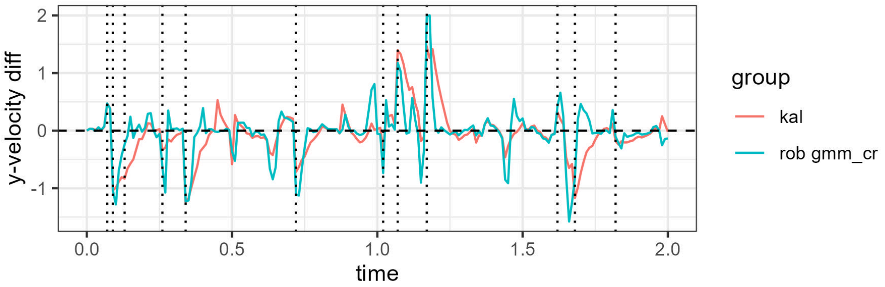


Figure A.56.: Difference of analysis means and y-velocity for the regular KF and the combined robust CR-BOCPD GMM-KF. Dotted lines indicate instances of change.

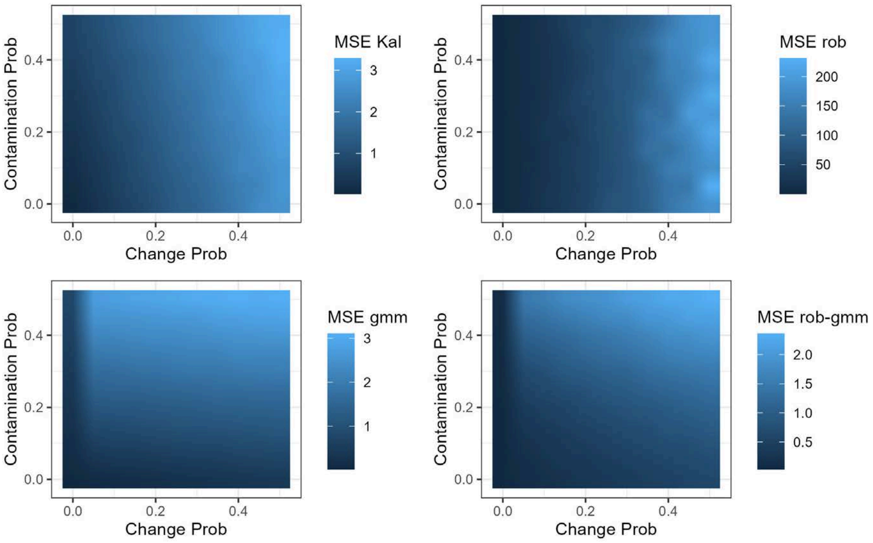


Figure A.57.: Tile plot with interpolation of change point probability φ and contamination probability ε regarding mean MSE over repeated Monte Carlo simulations for fixed covaraince factors $\tilde{u}$ and $\tilde{c}$ and each individual method: the regular KF (top-left), the $\mathcal{D}_{\sqsupseteq}$-KF (top-right), the CR-BOCPD GMM-KF (bottom-left) and the robust CR-BOCPD GMM-KF (bottom-right).

For the second set of scaling experiments, the respective probabilities are fixed at $\varphi = \varepsilon = 0.15$ with the respective covariance factors scaling from 5 to 200 in steps of 15.

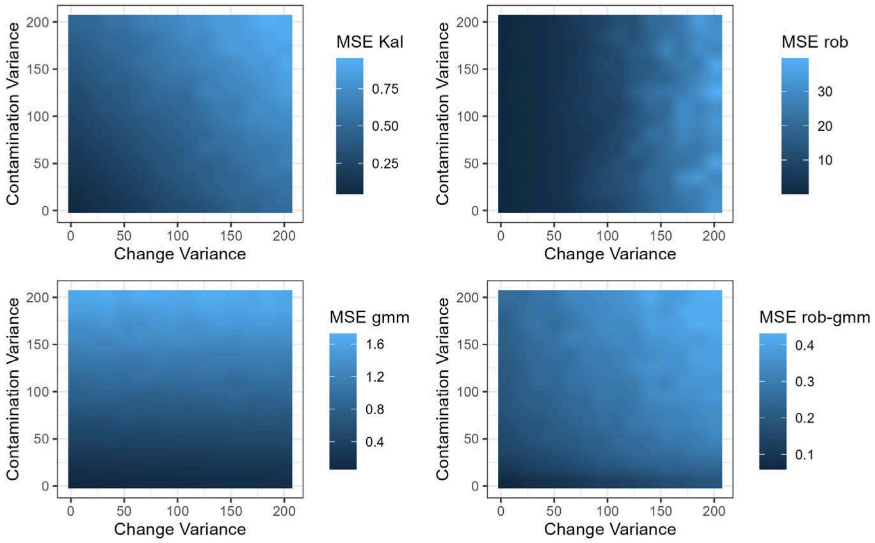


Figure A.58.: Tile plot with interpolation of change point covariance factor $\tilde{u}$ and contamination covariance factor $\tilde{c}$ regarding median information criterion score over repeated Monte Carlo simulations for fixed probabilities φ and ε and each individual method: the regular KF (top-left), the $\mathcal{D}_{\sqsupseteq}$-KF (top-right), the CR-BOCPD GMM-KF (bottom-left) and the robust CR-BOCPD GMM-KF (bottom-right).

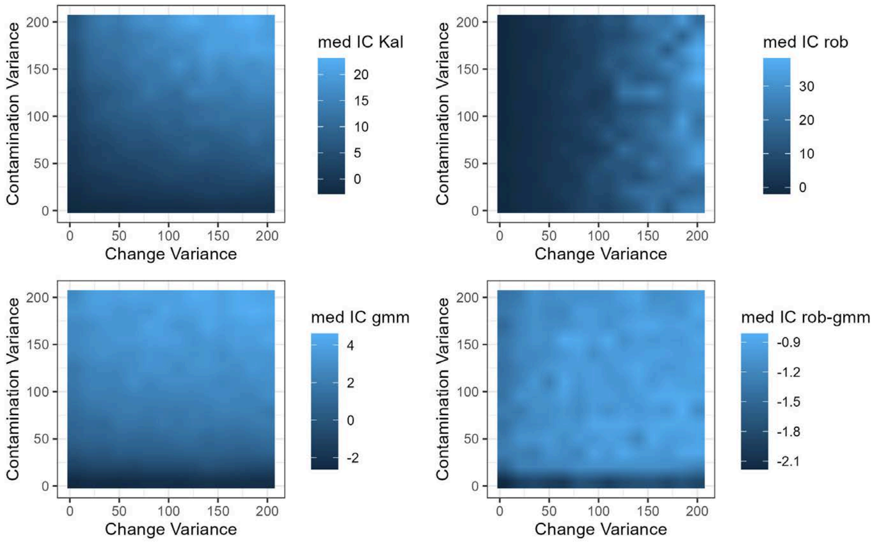


Figure A.59.: Tile plot with interpolation of change point covariance factor $\tilde{u}$ and contamination covariance factor $\tilde{c}$ regarding mean MSE over repeated Monte Carlo simulations for fixed probabilities φ and ε and each individual method: the regular KF (top-left), the $\mathcal{D}_{\sqsupseteq}$-KF (top-right), the CR-BOCPD GMM-KF (bottom-left) and the robust CR-BOCPD GMM-KF (bottom-right)

A.3.2. Experiment B: Additional Graphs

The model, simulation and parameter choices as well as notation are described in the respective section.

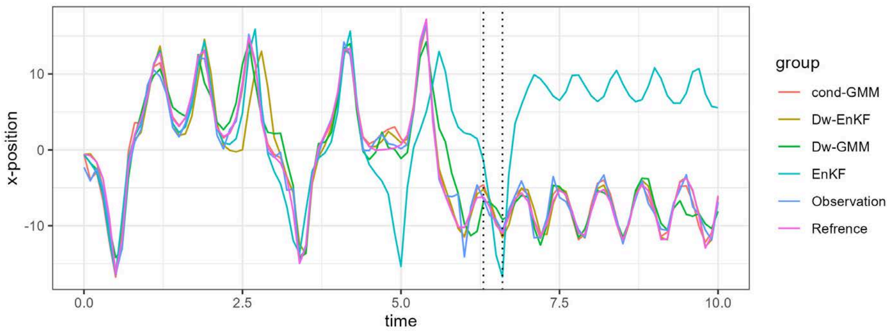


Figure A.60.: First variable of the simulated signal, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF and the robust CR-BOCPD GMM-EnKF. Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

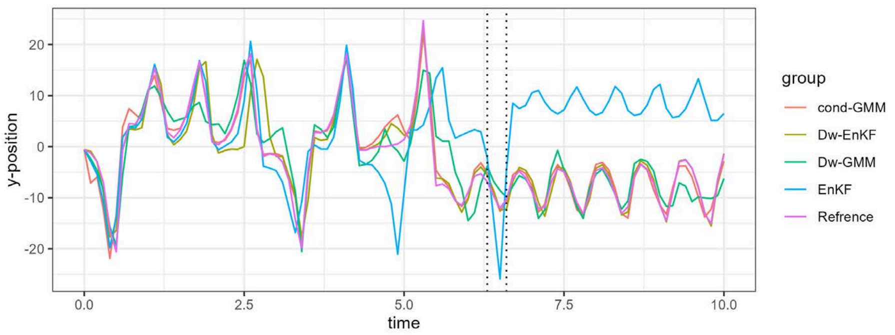


Figure A.61.: Second variable of the simulated signal, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF and the robust CR-BOCPD GMM-EnKF. Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

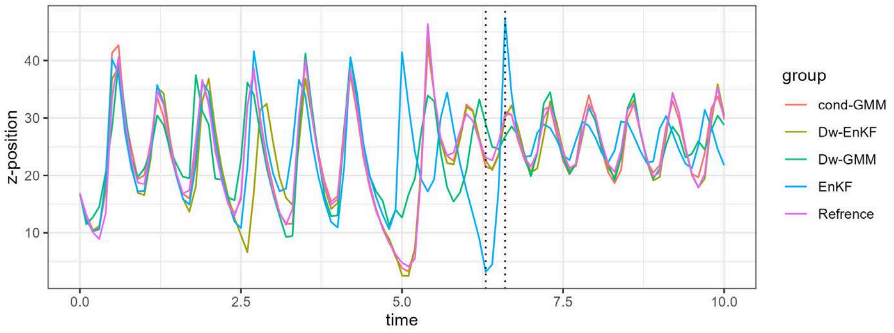


Figure A.62.: Third variable of the simulated signal, generated observations, the analysis mean of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF and the robust CR-BOCPD GMM-EnKF. Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

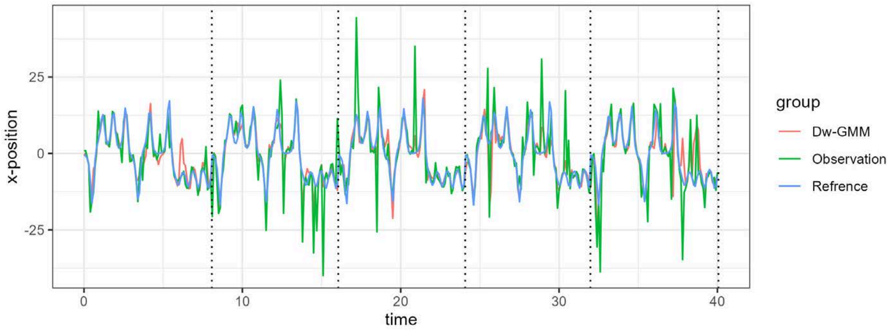


Figure A.63.: First variable of the simulated signal with jumps and contaminated observations as well as the analysis mean of the robust CR-BOCPD GMM-EnKF. Dotted lines indicate instances of change via a restart of the system.

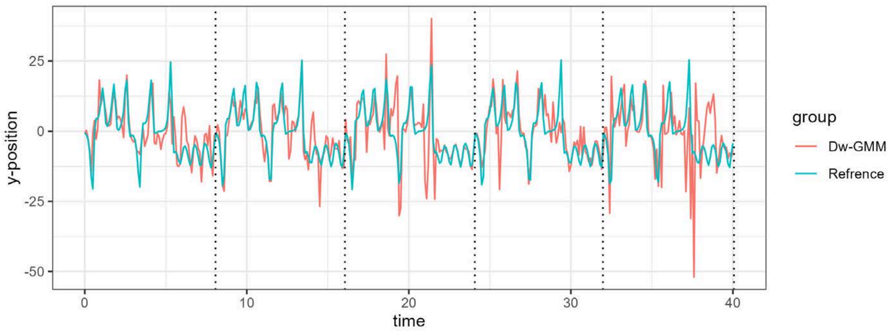


Figure A.64.: Second variable of the simulated signal with jumps and contaminated observations as well as the analysis mean of the robust CR-BOCPD GMM-EnKF. Dotted lines indicate instances of change via a restart of the system.

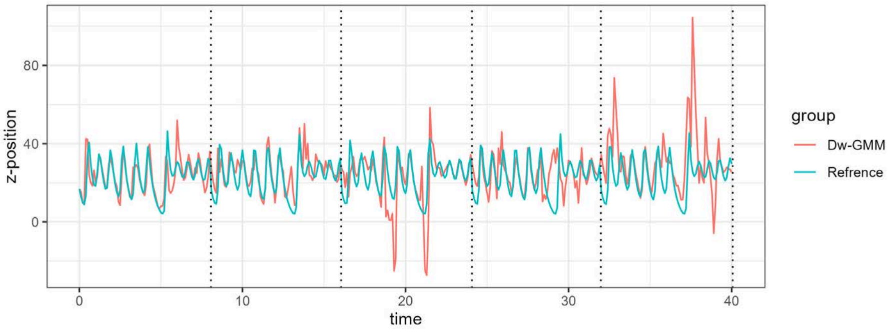


Figure A.65.: Third variable of the simulated signal with jumps and contaminated observations as well as the analysis mean of the robust CR-BOCPD GMM-EnKF. Dotted lines indicate instances of change via a restart of the system.

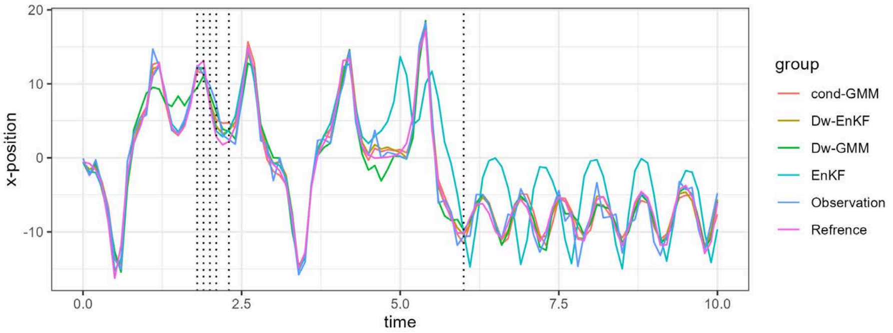


Figure A.66.: First variable of the simulated signal, generated observations, the analysis mean with equalized number of ensemble members of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF and the robust CR-BOCPD GMM-EnKF. Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

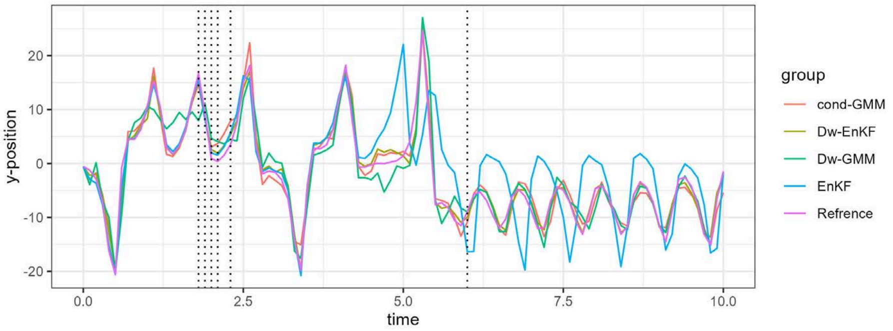


Figure A.67.: Second variable of the simulated signal, generated observations, the analysis mean with equalized number of ensemble members of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF and the robust CR-BOCPD GMM-EnKF. Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

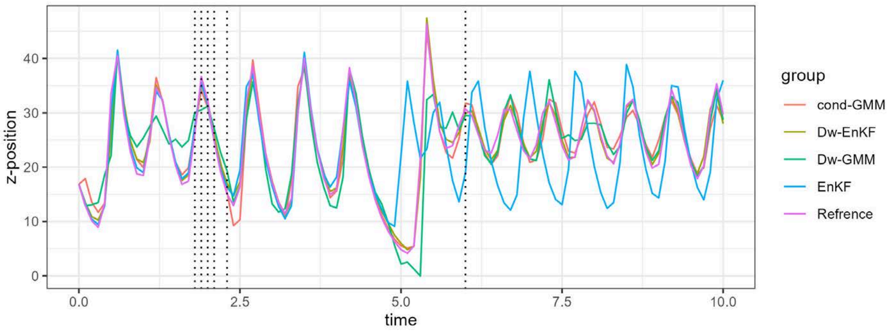


Figure A.68.: Third variable of the simulated signal, generated observations, the analysis mean with equalized number of ensemble members of the regular EnKF, the default $\mathcal{D}_w$-EnKF, the CR-BOCPD GMM-EnKF and the robust CR-BOCPD GMM-EnKF. Dotted lines indicate restarts of the robust CR-BOCPD GMM-EnKF.

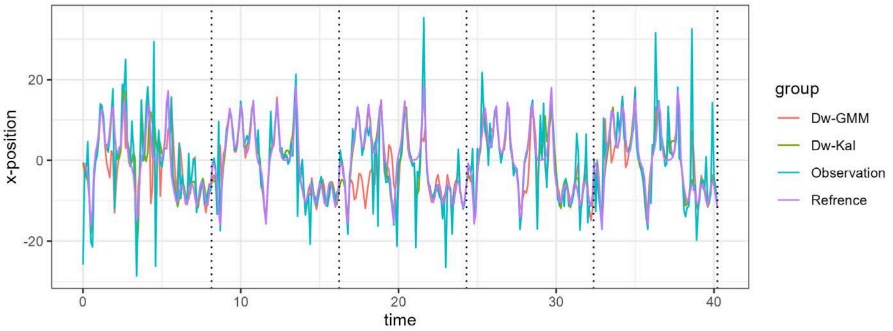


Figure A.69.: First variable of the simulated signal with jumps and contaminated observations as well as the analysis mean with equalized number of ensemble members of the $\mathcal{D}_w$-EnKF and robust CR-BOCPD GMM-EnKF. Dotted lines indicate instances of change via a restart of the system.

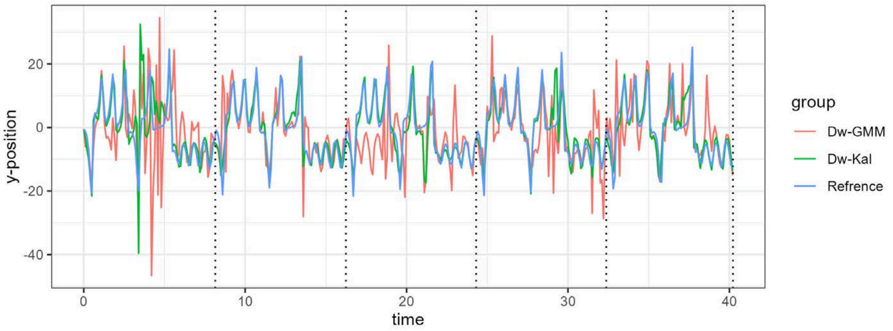


Figure A.70.: Second variable of the simulated signal with jumps and contaminated observations as well as the analysis mean with equalized number of ensemble members of the $\mathcal{D}_w$-EnKF and robust CR-BOCPD GMM-EnKF. Dotted lines indicate instances of change via a restart of the system.

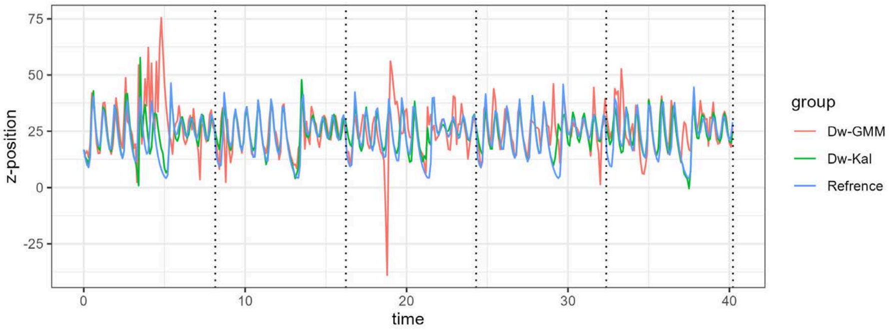


Figure A.71.: Third variable of the simulated signal with jumps and contaminated observations as well as the analysis mean with equalized number of ensemble members of the $\mathcal{D}_w$-EnKF and robust CR-BOCPD GMM-EnKF. Dotted lines indicate instances of change via a restart of the system.

Examples of numerical implementation of the derived algorithms also used for the simulation experiments can be provided. Please do not hesitate to get in touch via hans.reimann.97@gmail.com for any questions.