# Spatio-temporal inference for high-dimensional non-stationary processes

Jointly tackling challenges in large-scale data streams

**Jeremias Knoblauch**
j.knoblauch@warwick.ac.uk

**Abstract**

Inference for complex dynamical systems generating high-dimensional structured data is complicated by non-stationarity, changepoints, model uncertainty, misspecification and outliers. While the analysis of real-world data streams almost always needs to address these complications, tackling them jointly leads standard likelihood-based learning rules to break down. In contrast, my research shows that learning rules derived from Generalized Bayesian Inference can do this efficiently and effortlessly. Consequently, I am committed to advancing their application and theory for state space models, complex spatio-temporal point processes, sequential Monte Carlo and the related, long-standing issue of proposal sensitivity in Importance Sampling.

## Motivation

Every minute, Facebook processes more than 500,000 comments, 300,000 status updates and 150,000 photo uploads resulting from more than 1.5 billion daily interactions. Similarly, high-frequency trades are made in the span of Milliseconds, MasterCard processes more than 40,000 transactions a second, 4 million YouTube videos are being watched every minute and 24 billion emails sent every day. With complex large scale data streams being generated at such an incredible pace and scale, both commercial and public institutions are becoming increasingly interested in high-dimensional real-time inference. For example, within the Clean Air Project London, we are currently working with London's Major's Office to build monitoring tools for their extensive sensor networks As with most other modern day data streams, this analysis is severely complicated by four main challenges:

**Changes:** the measurement patterns change over time, often abruptly and in complex ways. For example, London's congestion charge introduction in 2003 affected air pollution levels because drivers changed their behaviors.

**Model Uncertainty/Misspecification:** The Data Generating Processes (DGPs) of most complex dynamical systems and data streams are impossible to describe even approximately with statistical models.

**Outliers:** Abnormal or unusual data points are abundant and cause traditional inference methods to fail. For instance, both foggy days and Christmas produce stunning irregularities in pollutant levels across London.

**On-line inference:** Modern applications often require real-time rather than retrospective inference. Case in point: While London's air pollutants are measured only every few minutes, the sensor network is so vast that re-analyzing the existing data from scratch would require weeks.

While each of these issues *individually* has an extensive modern literature attached to it, the corresponding methods are highly specialized and fail in the presence of more than one of these complications. My research tackles this by building on recent innovations within the field of Generalized Bayesian Inference (Bissiri et al., 2016; Jewson et al., 2018) to derive learning rules that function reliably in the joint presence of non-stationarity, model misspecification, outliers, and high dimensionality.

## Design Principles

To address outliers, on-line inference, permanent changes and model uncertainty in real-time inference, methods developed as part of this research adhere to the following principles:

**Interpretability:** Whenever one wants to shed light on a complex and poorly-understood DGP, the model parameters need to clearly translate relevant aspects of the underlying process. Beyond that, lack of interpretability can pose other severe drawbacks and has even been declared illegal within the European Union for a wide range of inference tasks since the General Data Protection Regulation has become active (see e.g. Goodman and Flaxman, 2016);

**Computational Efficiency:** The developed methods solve problems in real time, which constrains their computational complexity in a natural and meaningful way;

**Uncertainty Quantification:** Whenever possible and feasible, the uncertainty about inferred quantities and variables should be quantifiable, preferably in a principled Bayesian manner.

Together, the principles of interpretaility, computational efficiency and uncertainty quantification guide me. Consequently, black box methods are not a focal point of my research and Bayesian inference is preferred to frequentist methods whenever this is computationally viable.

# Theoretical basis

On-line inference in streaming data that is non-stationary, hard to model appropriately and full of outliers demands strong theoretical pillars for inference. The first of these is generalized Bayesian inference, which directly addresses the challenges of both model uncertainty/misspecification and outliers. The second pillar is state space modelling, which elegantly accounts for abrupt or continuous changes in the DGP and is well-suited for real-time inference.

## Generalized Bayesian inference

The assumption that the data model can describe the DGP exactly is at the heart of virtually all theoretical guarantees for statistical inference such as the sufficiency principle, unbiasedness or consistency (e.g., Casella and Berger, 1990). This supposition is often called the M-closed world assumption (Bernardo and Smith, 2001). To make this point notationally more concise, suppose you observe some data $\boldsymbol{y}$ generated from a real world DGP. The M-closed world assumption presumes that the $\boldsymbol{\theta}$-parameterized model $M$ can explain $\boldsymbol{y}$ exactly as

$$\boldsymbol{y} = M(\boldsymbol{\theta}) \tag{1}$$

In inferential practice of course, $M$ never describes the DGP exactly. Yet, as long as the model matches the DGP at least approximately, inference strategies and theoretical guarantees based on the M-closed world are still very useful. This observation was immortalized in Box's aphorism that *all models are wrong, but some are useful* and explains why the M-closed assumption remains an excellent principle in many traditional statistical inference tasks.

However, the M-closed assumption is ill-suited for inference in contemporary complex, noisy and high-dimensional data streams. For this class of inference problems, it is typically impossible to specify a model $M$ that can approximately match the DGP. Moreover, even if it were theoretically possible to design a model $M$ expressive enough to model the DGP, it would be computationally infeasible to do inference on its parameters. Invariably, this leads to inference on models that are intended as rough approximations to the typical behavior of $\boldsymbol{y}$ rather than as good descriptions of $\boldsymbol{y}$'s DGP. In this setting, performing inference relying on eq. (1) and the M-closed world assumption leads to a multitude of problems. Two issues become especially relevant:

**Model misspecification and outliers:** Under model misspecification, i.e. when $M$ and the DGP do not match closely, traditional inference becomes unreliable. In particular, standard likelihood-based inference focuses on outliers and *atypical* patterns in the observations (e.g., missing values, false entries, ...). This is optimal in the M-closed world, as in this case the unusual observations carry most information about the DGP. In contrast, this becomes detrimental when one wants to do inference with $M$ to obtain a rough approximation to the *typical* behavior of $\boldsymbol{y}$.

**High dimensionality:** As the dimension increases, so does the difficulty of finding an appropriate model $M$ reflecting all relevant aspects of the DGP. Moreover, computational constraints will become more important and force the use of a simplified approximate model in practice. To exacerbate the additional challenge, higher dimensionality also compounds the adverse effects associated with abnormal or outlying behaviour: Any observation need only be aberrant relative to the model $M$ in a *single* dimension to dominate inference outcomes.

My research builds on recent innovations within the field of Generalized Bayesian Inference (Bissiri et al., 2016; Jewson et al., 2018) to derive learning rules that can function reliably in the *M-open* world, i.e. under model misspecification, outliers, and high dimensionality. The key observation for this is that likelihood-based inference in the M-closed world minimizes the Kullback-Leibler Divergence (KLD) between $\boldsymbol{y}$ and $M$ in terms of $\boldsymbol{\theta}$. While standard Bayesian inference via the KLD is optimal in the M-closed world, this no longer holds in the M-open world. To perform appropriate Bayesian inference when $M$ and the DGP do not coincide, one can derive new Bayes rules

from *robust* divergences (Jewson et al., 2018), for example using the families of $\alpha$-, $\beta$- and $\gamma$-divergences (Cichocki and Amari, 2010). For a prior belief $p(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$, these generalized Bayes rules take the form

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}) \exp\left\{\ell^D(\boldsymbol{y}|\boldsymbol{\theta})\right\}}{\int p(\boldsymbol{\theta}) \exp\left\{\ell^D(\boldsymbol{y}|\boldsymbol{\theta})\right\} d\boldsymbol{\theta}}, \tag{2}$$

where $\ell^D$ is a divergence-specific function. For example, if $D = \text{KLD}$, then $\ell^D$ is the log-likelihood and it is easy to see that the standard Bayes Theorem is recovered. Similarly, $\alpha$-, $\beta$- and $\gamma$-Divergences have a one-to-one correspondence with some functions $\ell^\alpha, \ell^\beta$ and $\ell^\gamma$ that enable efficient parameter updating with respect to that divergence. To see how these divergences address misspecification and robustness, it is instructive to take a look at the corresponding loss functions. For example, the $\beta$-divergence that we are going to use as a running example in the application section has loss

$$\ell^\beta(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{\beta} p(\boldsymbol{y}|\boldsymbol{\theta})^\beta - \frac{1}{1+\beta} \int p(\boldsymbol{z}|\boldsymbol{\theta})^{1+\beta} d\boldsymbol{z}. \tag{3}$$

Noting that the function $\ell^\beta(\boldsymbol{y}|\boldsymbol{\theta})$ goes to $\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$ as $\beta \to 0$, it is clear that inference based on the $\beta$-divergence coincides with standard Bayesian inference in the limit. Focusing on the first term $\frac{1}{\beta} p(\boldsymbol{y}|\boldsymbol{\theta})^\beta$, it is also clear that for $\beta > 0$, $\ell^\beta$ downweights small values of $p(\boldsymbol{\theta}|\boldsymbol{y})$ relative to the log-loss. This decreases the influence outlying observations have on inference, an observation one can make concise via *influence functions* as in Figure 1. Influence functions give an indication of how much a single additional observation can influence the posterior belief distribution. Figure 1 shows how the influence profiles differ for a range of $\beta$-divergences and the KLD and how this translates into drastically different posteriors on a toy example.

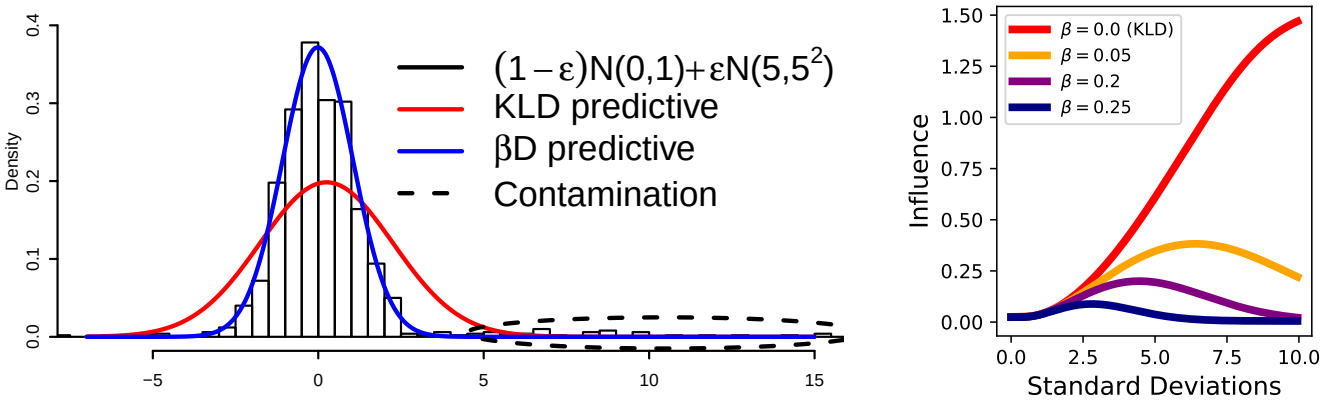


Figure 1: **Left:** Data with $\varepsilon = 5\%$ outlier contamination and the corresponding **standard (KLD)** and **$\beta$-divergence** ($\beta = 0.5$) posteriors. The $\beta$-divergence can recover the uncontaminated distribution while standard Bayesian inference fails. **Right:** Influence functions for different $\beta$ and the KLD expressed in terms of Standard Deviations from the posterior mean. As observations become less likely under the fitted model, standard Bayesian inference via the KLD assigns more and more influence. In contrast, the $\beta$-divergence only does so up to a point of maximum influence. If an observation occurs further away from the prescription of the model, it is assumed to be aberrant or outlying. Accordingly, its influence decays from that point onwards.

## State Space Models

State Space Models (SSMs) are flexible and feasible for a wide range of time series problems. Their main appeal for this project is twofold: Firstly, they are perfectly suited to model non-stationary behaviour through slow and abrupt changes across time via their latent state sequence $\boldsymbol{x}$. Secondly, inference can often be written recursively, which enables computation to proceed on-line and in real-time. Writing $\boldsymbol{y} = (y_1, y_2, \ldots, y_T) = y_{1:T}$ as the sequence of $T$ observations and $\boldsymbol{x} = (x_1, x_2, \ldots, x_T) = x_{1:T}$ as a sequence of $T$ unobserved variables related to $\boldsymbol{y}$ via the sequence $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_T) = \theta_{1:T}$ of parameters, a schematic view of SSMs is given in Figure 2. The Figure reveals that for SSMs, the model for $\boldsymbol{y}$ is not only a function of the parameters $\boldsymbol{\theta}$, but also of another hidden component $\boldsymbol{x}$ so that

$$\boldsymbol{y} = M_{\boldsymbol{x}}(\boldsymbol{\theta}) \tag{4}$$

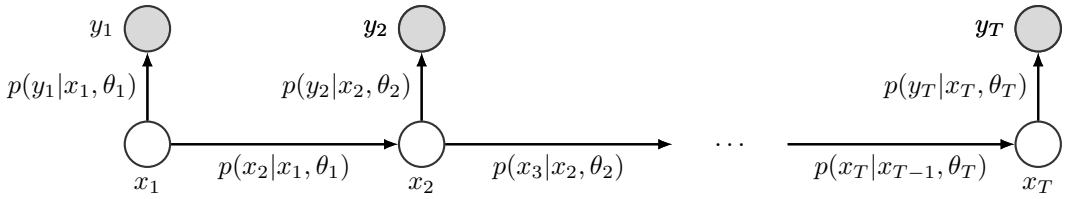


Figure 2: Schematic view of state space models. The observables $y_{1:T}$ corresponding to the gray nodes are conditionally independent of the latent sequence $x_{1:T}$ corresponding to the white nodes. Emissions $x_t \to y_t$ and transitions $x_t \to x_{t+1}$ depend on parameters $\theta_t$ which are typically unknown and have to be inferred, often on-line.

The sequence $\boldsymbol{x}$ indexes the *state* of the model and is what ultimately makes SSMs so versatile: Variations in $\boldsymbol{x}$ correspond to a (smooth or abrupt) change in the model for $\boldsymbol{y}$. If $\boldsymbol{\theta}$ is known, SSMs allow for recursive inference via

$$p(y_{1:t}, x_{1:t}|\theta_{1:t}) = p(y_{1:(t-1)}, x_{1:(t-1)}|\theta_{1:(t-1)}) \times \underbrace{p(x_t|x_{t-1},\theta_t)}_{\text{transition}} \times \underbrace{p(y_t|x_t,\theta_t)}_{\text{standard emission}} \tag{5}$$

While things get notationally and computationally more involved if $\boldsymbol{\theta}$ is unknown, the same recursive logic applies. The versatility of this model class for modelling $\boldsymbol{y}$ is reflected by the state space $\mathcal{X}$, which is the collection of values $x_t$ is allowed to take. If $\mathcal{X}$ is countable, this type of model can mimic recurrent behaviour (e.g., Beal et al., 2002; Caron et al., 2012) or abrupt changes (e.g., Adams and MacKay, 2007; Fearnhead and Liu, 2007; Knoblauch and Damoulas, 2018). For uncountable $\mathcal{X}$, one can also model slower transitions of $\boldsymbol{y}$, but inference typically becomes more challenging, too (e.g., Chopin et al., 2013).

Combining the SSM recursion of equation (5) with Generalized Bayes rules as in equation (2) yields a generic recipe for on-line recursive inference that is robust to model misspecification and outliers via

$$p^D(y_{1:t}, x_{1:t}|\theta_{1:t}) \propto p^D(y_{1:(t-1)}, x_{1:(t-1)}|\theta_{1:(t-1)}) \times \underbrace{p(x_t|x_{t-1},\theta_t)}_{\text{transition}} \times \underbrace{\exp\{\ell^D(y_t|x_t,\theta_t)\}}_{\text{robust emission}} \tag{6}$$

While this last step appears harmless, the resulting inference methods generate a range of new challenges that need to be addressed. First of, the parameters $\boldsymbol{\theta}$ will again be unknown in practice and have to be inferred on-line, too. Next, normalizing constants are *never* available in closed form for the new recursion, posing serious computational challenges. Another complication arises from the fact that most other divergences such as the $\alpha$-, $\beta$- and $\gamma$-divergences all depend on hyperparameters $\alpha, \beta$ and $\gamma$ (see e.g., Cichocki and Amari, 2010) that need to be elicited carefully. Moreover, inference ought to be on-line, meaning that the same should hold for choosing the divergences' hyperparameters. For example, Knoblauch et al. (2018) addresses parameter inference and intractable normalizing constants with a quasi-conjugate property of the $\beta$-divergence. The same paper also provides an on-line optimization scheme for choosing $\beta$ as minimizer of expected prediction error.

# Application examples

This section provides two short univariate examples of how my research addresses the problem complex of interest. While univariate examples make it is easier to pictorially illustrate challenges and solutions, the performance gain of the presented methods is even *more pronounced* for multivariate and high-dimensional data sets.

## Bayesian On-line Changepoint Detection

Knoblauch et al. (2018) shows how the application of equation (6) can lead to substantially improved performance for Bayesian On-line Changepoint Detection (BOCPD). By definition, data streams requiring changepoint models are affected by model uncertainty and sudden changes. Usually, these data sets are also full of outliers and have been generated by a complex DGP. Thus, re-designing BOCPD to deal with the complications of modern day data is both natural and necessary.

BOCPD is one of the most important Bayesian methods for tackling inference in non-stationary data streams. It defines a state variable $x_t \in \mathbb{N}_0$ called *run-length* that tracks the most recent abrupt change at time $t$. In other words, if $x_t = k$, then the DGP was subject to an abrupt change at time $t - k$. So whenever one infers $x_t = 0$, this amounts to inferring a permanent change at time $t$ so that the model for $y_t$ ought to be different from that for $y_{1:(t-1)}$. Figure 3 demonstrates the powerful effect of substituting the **standard** equation (5) by the **new** equation

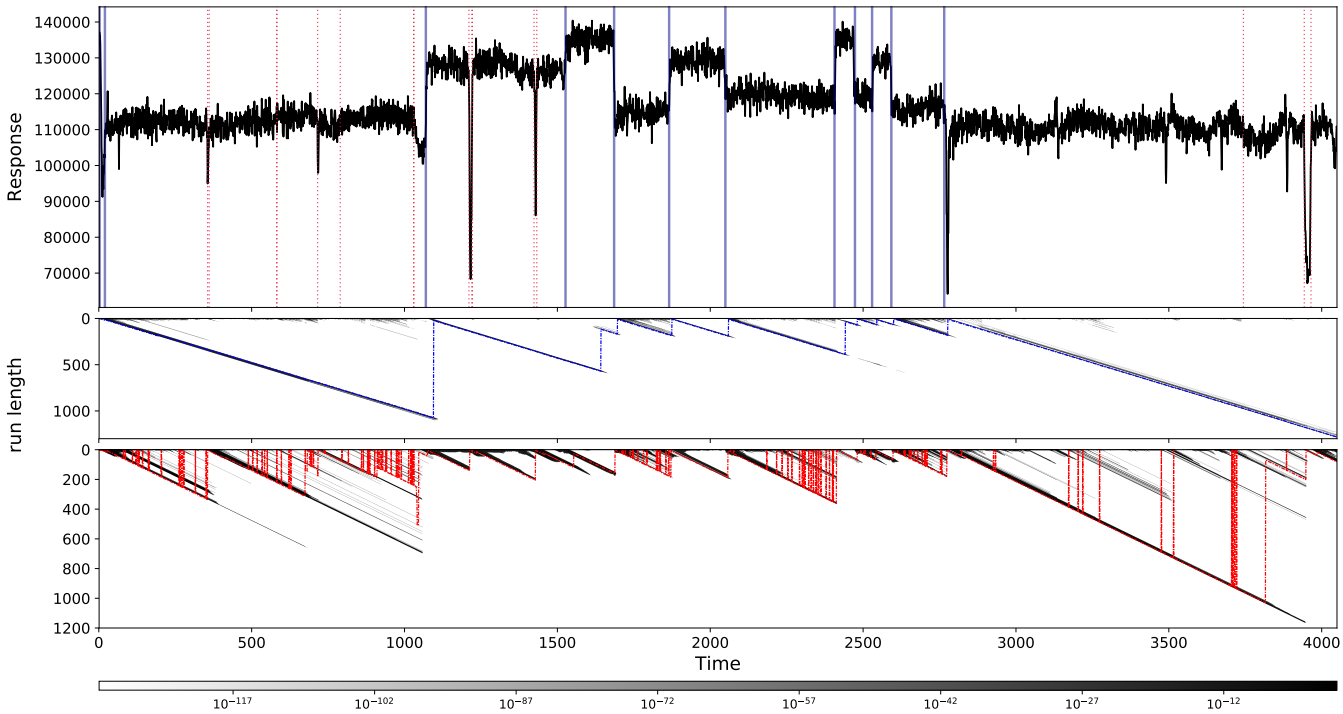


Figure 3: **Top:** The well-log data set with **new** and **standard** retrospective BOCPD Maximum A Posteriori segmentation. **Middle and Bottom:** **new** and **standard** run-length posteriors for $x_t$ in grayscale, with colored maximum. The corresponding False Discovery Rates for changepoints are 0% and $> 90\%$.

(6) with the $\beta$-divergence loss $\ell^\beta$ on a real world data set giving the nuclear magnetic response during the drilling of a well. This well-log data set has been used multiple times to illustrate standard BOCPD *after* removing its numerous outliers (see e.g., Adams and MacKay, 2007; Turner et al., 2009; Turner, 2012). Yet, applying preprocessing steps before running BOCPD makes the method on-line in name only. Furthermore, it is not obvious how to preprocess data which has more than one dimension. Without removing outliers first however, the well-log data causes **standard** BOCPD to fail: In the bottom panel, the maximum of the posterior run-length $x_t$ falsely drops to 0 more than 130 times, making for a False Discovery Rate of changepoints in excess of 90% with standard inference methods. In contrast, the middle panel shows that the **new** method has a False Discovery Rate of 0% and finds all of the true changepoints. While this comes at the cost of slightly increased changepoint detection latency, it has the effect of significantly improving prediction errors, too.

While other work on on-line robust changepoint detection exists (Pollak, 2010; Fearnhead and Rigaill, 2017; Cao and Xie, 2017), our method is significantly more general than any other existing approach. Most importantly, it is neither constrained to univariate data nor to detecting changepoints only in the mean function. In fact, it works for *any* probabilistic model, including discrete-valued and time-dependent ones. Furthermore, unlike existing approaches its outputs provide more information than just point estimates. Specifically, it is completely Bayesian and enables thorough uncertainty quantification.

## Sequential Monte Carlo

Amongst other things, I am currently also working on robust sequential Monte Carlo inference for slow, continuous changes in $\boldsymbol{x}$. Sequential Monte Carlo is particularly important for real-time inference in engineering, satellite tracking problems and financial econometrics. As they all have to deal with highly complex DGPs, methods of interest to these areas need to address misspecification and outliers.

While the results of my research are not ready to be published yet, the findings are very promising thus far. For example, Figure 2 compares results for inference on simulated data in a correctly specified and a moderately *misspecified* Linear Gaussian SSM with the **standard** and a **new** robust sequential Monte Carlo method[1]. The plots reveal a desirable asymmetry in the risk of minimizing a robust divergence instead of the KLD: On the one hand,

[1] In the moderately misspecified setting, the noise process in the emissions $x_t \to y_t$ is still linear, but Student's $t_{10}$ instead of Gaussian. Additionally, 5% of the observations are injected with asymmetric $\alpha$-stable noise.

the gain in predictive performance is substantial under misspecification: In this setting, the mean square prediction error for $x_t$ is reduced by $> 55\%$. On the other hand, if the model is correctly specified, the gain in mean square prediction error is $< 1\%$ when compared to the (in this case optimal) Kalman Filter. While the filter-based inference in Figure 2 is fully on-line, the depicted asymmetry gets even *more* pronounced for the associated off-line smoothing techniques. Exactly as with BOCPD, this effect again is even more pronounced for the higher-dimensional setting. Moreover, there are strong theoretical reasons to expect the new inference methodology to remedy the problem of particle degeneracy associated with Sequential Monte Carlo techniques.

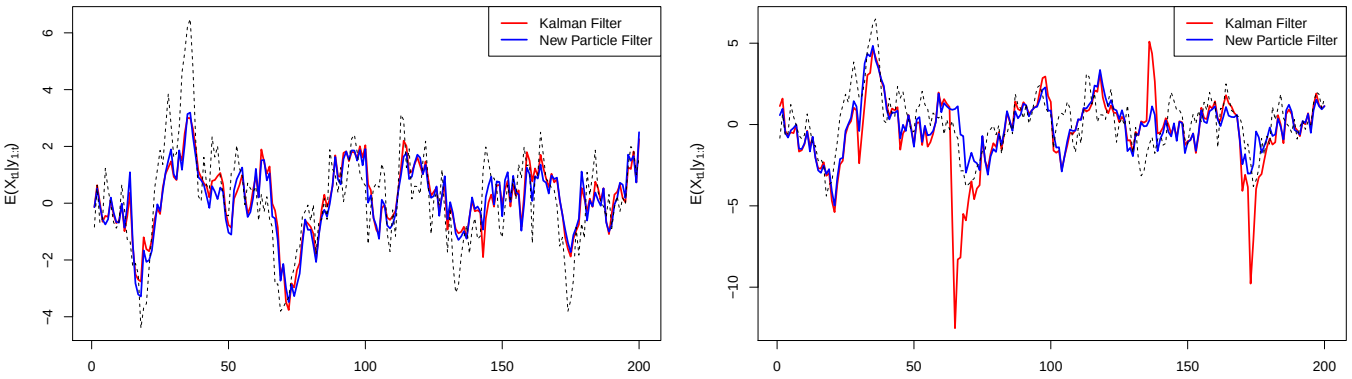


Figure 4: **Left:** **New** and **standard** method in a correctly specified linear SSM. **Right:** **New** and **standard** method in a misspecified linear SSM.

## Future research

In the future, I will apply the translation of equation (5) into (6) to other settings of particular interest for modern day large scale data streams. In particular, further research will address

**Point Processes:** On-line inference in complex spatio-temporal point processes such as Log-Gaussian Cox Processes by itself is already an important problem lacking a good solution. I want to find principled real-time solutions to this problem. In a second step, this will be extended into dealing with misspecification, non-stationarity and outliers via the attractive framework of Bayesian On-line Changepoint Detection. Work on this has already started and will intensify over the course of a 6-month Masters thesis I will supervise from January on.

**Quasi-conjugate Robustness:** Sampling-based approaches for inference through equation (6) such as Sequential Monte Carlo can violate computational efficiency requirements for real-time inference in very high-dimensional problems. In Knoblauch et al. (2018), we resolved this problem using an efficient structural variational approximation that can be generalized. This exciting generalization has three extremely convenient advantages for scalable and robust inference in high dimensions: Firstly, it retains parameter dependence and even is exact in an asymptotic sense. Secondly, it avoids sampling. Thirdly, its objective function has closed form for a range of models. The theoretical groundwork for this project is well underway and is applicable to exponential families as well as Gaussian Processes. I expect a paper to be submitted around June as result of another Master thesis that I will supervise.

**Importance Sampling:** (Sequential) Importance Sampling is of fundamental importance for any recursive Bayesian inference scheme. For example, it is the engine at the heart of any Sequential Monte Carlo Method. High-dimensional problems pose a serious challenge to Importance Sampling because efficient proposal distributions become extremely difficult to design. The result is a substantially increased sensitivity to the importance distribution, which results in extremely large variances of the importance weights. This in turn means that importance weights are either close to zero or close to one, implying that complex densities are represented by a very small number of points. The machinery of Generalized Bayesian Inference allows this problem to be addressed in a theoretically appealing way without requiring black box parameter tuning. This project is in its initial stage and will be my focal point once research on the new Sequential Monte Carlo methods has come to a close.

# References


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

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584.

Bernardo, J. M. and Smith, A. F. (2001). Bayesian theory.

Bissiri, P. G., Holmes, C. C., and 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.

Cao, Y. and Xie, Y. (2017). Robust sequential change-point detection by convex optimization. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1287–1291. IEEE.

Caron, F., Doucet, A., and Gottardo, R. (2012). On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2):579–595.

Casella, G. and Berger, R. (1990). *Statistical Inference*. Duxbury advanced series. Brooks/Cole Publishing Company.

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). Smc2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426.

Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

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

Fearnhead, P. and Rigaill, G. (2017). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, (just-accepted).

Goodman, B. and Flaxman, S. (2016). European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.

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

Knoblauch, J. and Damoulas, T. (2018). Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2018). Doubly robust bayesian inference for non-stationary streaming data with $\beta$-divergences. *Neural Information Processing Systems (NIPS)*.

Pollak, M. (2010). A robust changepoint detection method. *Sequential Analysis*, 29(2):146–161.

Turner, R., Saatci, Y., and Rasmussen, C. E. (2009). Adaptive sequential Bayesian change point detection. In *Temporal Segmentation Workshop at NIPS*.

Turner, R. D. (2012). *Gaussian processes for state space models and change point detection*. PhD thesis, University of Cambridge.