Examination Report on PhD Thesis “Inference and Characterisation of Planar Trajectories” by Zhanglong Cao

1 Summary

This thesis comprises a novel and coherent study on using spline methods and Markov Chain Monte Carlo methods to tackle the online trajectory reconstruction problem. The candidate has demonstrated his broad knowledge of the literature and his capabilities in harnessing complicated mathematical tools for problem-solving. The simulation results also indicate that the proposed algorithms achieved some better or competitive performance in comparison studies.

In terms of presentation quality, this thesis needs to be improved through revision. The original contributions of the thesis are not clearly spelled, and the organization of the thesis at times seems confusing and convoluted. It seems to me that the thesis title and content are not a good match. The term ”planar trajectories” in the title, and the way the thesis opens the Introduction, imply that it is a study on GPS trajectory reconstruction, but this does not seem to be the case. Only in Chapter 2 a GPS trajectory dataset was used. Chapter 5 used ”some observed data”, without giving any further information about the source and nature of the data. Chapter 3 and Chapter 4 only used some simple simulated data. Empirically, the significance of the work can be leveraged by including more results obtained from using real-world data.

Does the thesis comprise a coherent investigation of the chosen topic? This thesis focuses on the inference for streaming time-series GPS data. Existing appropriate methods were intensively reviewed and some novel techniques were proposed and examined. Does the thesis deal with a topic of sufficient range and depth to meet the requirements of the degree? Zhanglong proposed some novel approaches and investigate the properties theoretically and numerically. When those ideas were firstly introduced and verified, they were applied to simple models then more complex cases such as irregularly sampled times series data were dealt. It was also demonstrated that the V-spline can be estimated by a Bayesian approach in a certain reproducing kernel Hilbert space. Overall, I think this thesis meets the level of PhD qualification. Does the thesis make an original contribution to knowledge in its field and contain material suitable for publication in an appropriate academic journal? Two novel approaches are proposed in this thesis. The first approach is the V-spline method in which the model consists of both location and velocity information (through basis function) and the smoothing parameter to capture the curvatures. Model parameters are estimated using the cross-validation score. Through some simulated examples and real data, it is demonstrated that the V-spline method was able to reconstruct the trajectory of time-series data. Secondly the cutoff and threshold ideas are adapted in the DAMH algorithm to avoid a poor model approximation and a conditional prediction for the Bayesian inference. A particular formed covariance matrix made easy in forecasting as it is described in Chapter 5. The posterior simulation results were compared based on the efficiency measures of the Markov chain and it was observed that the cutoff and threshold criterion were helpful for the Markov chain to explore the state space. This chapter begins with a simple state-space model then it is extended to the OC process (for irregularly sampled times series data). With the particular form of the precision matrix 1 A B1 = B1 B , the joint and conditional distributions were easily computed. I think that with some text and argument improvement both approaches are suitable for publication in an appropriate academic journal. Does the thesis meet internationally recognised standards for the conduct and presentation of research in the field? With my knowledge, I have not seen both above approaches in literature. If the ideas are developed bit further and the simulation study includes various case studies, I think there is no problem in publishing this work in international journals. Does the thesis demonstrate both a thorough knowledge of the literature relevant to its subject and general field and the candidates ability to exercise critical and analytical judgement of that literature? The thesis begins with literature review on the two well known approaches for trajectory reconstruction; smooth-spline and Bayesian inference. Then, the relationship between 1 polynomial smoothing spline and the Bayes estimate and, the sequential Monte Carlo methods are intensively revised in Chapters 3 and 4. From this profound study, the V-spline approach and the sliding window MCMC method were proposed and their prop- erties were investigated theoretically and numerically. This demonstrates that Zhanglong has ability to investigate the problem, study literature, propose appropriate methods and critically revise them comparing to other methods. 6. Does the thesis display mastery of appropriate methodology and/or theoretical material? Overall I agree the methodologies and arguments presented in this thesis. There were some parts that I had difficulties in understanding and, I listed them as the major comments. Some explanation or relevant work supporting the arguments would be good to make a strong thesis.

This thesis presents statistical methodology for estimating the entire trajectory of a moving vehicle using sparse noisy measurements, namely GPS. Two disparate class of techniques are considered to reconstruct the vehicles trajectory, namely spline methods and Bayesian state-space models, although they are not directly compared. The novel endeavour is the application of the techniques to the vehicle tracking problem. The work is of merit but not yet at a sufficiently polished stage. This is primarily true of the Bayesian state-space approach. The computational technique employed appears to be a heuristic derived from a combination of ideas in the literature. The algorithm (in Chapter 5) has not been sufficiently well described (see detailed comments below). The thesis would benefit from a chapter covering the detailed statistical modeling of the application, i.e. the description of the sensors, the GPS data and measurement model, the parametric vehicle motion model along with its unknown parameters which are to be inferred jointly with its trajectory from the available data. Once the precise model has been described, appropriate Bayesian and non-Bayesian computational methods can then be introduced. This will also improve reproducibility of the results should anyone desire to do so.

This thesis comprises a novel and coherent study on using spline methods and Markov Chain Monte Carlo methods to tackle the online trajectory reconstruction problem. The candidate has demonstrated his broad knowledge of the literature and his capabilities in harnessing complicated mathematical tools for problem-solving. The simulation results also indicate that the proposed algorithms achieved some better or competitive performance in comparison studies. In terms of presentation quality, this thesis needs to be improved through revision. The orig inal contributions of the thesis are not clearly spelt, and the organization of the thesis at times seems confusing and convoluted. It seems to me that the thesis title and content are not a good match. The term "planar trajectories" in the title, and the way the thesis opens the Introduction, implythatitisastudyonGPStrajectoryreconstruction,butthisdoesnotseemtobethecase. Only in Chapter 2 a GPS trajectory dataset was used. Chapter 5 used "some observed data", without giving any further information about the source and nature of the data. Chapter 3 and Chapter 4 only used some simple simulated data. Empirically, the significance of the work can be leveraged by including more results obtained from using real-world data. There are numerous grammatical or presentational flaws that need to be corrected. Please refer to the following detailed comments, and the markings in the returned hard-copy.

2 Details

Overall, please pay attention to the following minor, but important points for revision: Be consistent with the use of capitalizations, e.g., "Table 2.1", "Figure 3.4", "Algorithm 5.2". Reproduce the figures so that the keys/legends and labels are readable. This applies also to the figures in the appendices. Ensure the proper use of articles. Reduce unnecessary repetitions. For instance, Equations 5.9, 5.29, 5.45, and 5.59(?) seem identical. Chapter 1 Correct the citation "(Trevor Hastie, 2009)". Please indicate that this the 2nd edition of the book, which is authored by Hastie, Tibshirani, and Friedman. Also follow the good practice to give the exact section/pages when citing from a book. As this citation was made throughout the thesis, make sure you correct all its occurrences. Before Section 1.6, please add a section outlining your research problem(s), and listing all the significant contributions that you have made in your thesis work.

Overall, please pay attention to the following minor, but important points for revision: Be consistent with the use of capitalizations, e.g., "Table 2.1", "Figure 3.4", "Algorithm 5.2". Reproduce the figures so that the keys/legends and labels are readable. This applies also to the figures in the appendices. Ensure the proper use of articles. Reduce unnecessary repetitions. For instance, Equations 5.9, 5.29, 5.45, and 5.59(?) seem iden tical.

3 Major Comments

3.1 Chapter 1

Chapter 1. In general, from my examining experience, the Introduction should be a thorough review of the main computational paradigms the thesis will draw on. In this case the review of Monte Carlo computational methods are cursory at best. There is no proper mathematical formulation and concepts and mathematics are introduced somewhat on the hoof, without proper thought or planning. There are many methods mentioned with little or no detail. Some of the methods do not appear to be relevant to the work of the thesis, namely Hamiltonian Monte Carlo, Zig-zag samplers. I would recommend a much more detailed focus on methods the thesis does eventually employ, for example, particle filters, adaptive MCMC, etc.

Chapter 1 Correct the citation "(Trevor Hastie, 2009)". Please indicate that this the 2nd edition of the book, which is authored by Hastie, Tibshirani, and Friedman. Also follow the good practice to give the exact section/pages when citing from a book. As this citation was made throughout the thesis, make sure you correct all its occurrences. Before Section 1.6, please add a section outlining your research problem(s), and listing all the significant contributions that you have made in your thesis work.

3.2 Chapter 2

Chapter 2. The spline methodology is formulaic in its structure, although it has to be developed in detail in every instance of a new application. This is indeed one major contribution by the author. The chapter describes a spline solution to vehicle tracking, dubbed V-spline, that incorporates noisy position and velocity observations and penalizes both data misfit and spline roughness. I regard this the most complete and significant chapter of the thesis. 1 Theorem 1 describes the basic feature of the optimal V-spline but the linear outside knots feature of the optimal solution is mathematically undefined or ambiguous. The Hermite interpolation in sec 2.2.1 would benefit from an illustrative figure. The domain of the function f should be declared. To better understand the contribution, a clear statement at the beginning of the chapter that declares which of the theorems are original should be supplied. Section 2.2.2 is terse and difficult to follow but otherwise appears mathematically correct. Section 2.5, application to real dataset, should clearly indicate if velocity and bearing measurements are somewhat distinct from position measurements and not derived from the positions themselves. What is the boom and boom status? Why say pi and si are d-dimensional instead of d = 2? Why si for velocity and not vi? Also, why arent bearing measurements used for the reconstruction? What is the loss in fitting splines for each dimension separately? Why havent the results of this chapter compared with any other competing method since state-space models for tracking with GPS with dropouts, inertial sensors etc are well documented, especially for aeronautical applications?

3.3 Chapter 3

Chapter 3. This is by far the most technical and difficult chapter to follow. Not all terms are adequately defined, for example I did not see a definition of a reproducing kernel. My best attempt to comprehend the chapter is that the main theorem appears to be Theorem 7 which relates the best V-spline solution to the mean of a certain Bayesian posterior. I am unclear as to the practical significance of the result. I am used to very technical expositions but this chapter eludes me. I think more effort could have been expended to define terms better and accompany the mathematics with a suitable narrative to help the reader along the way. The chapter is dense with mathematical derivation and has a veneer of correctness about it. But then again I am no expert in the subject matter of the chapter.

3.4 Chapter 4

Chapter 4. The intention of this chapter is to present an overview of the discrete time state-space formulation and related Monte Carlo computational methods. It is important to properly understand the limitations of the various computational paradigms before committing to one of them to take forward, which is presumably the aim of the chapter. The chapter presents the particle filter for online state estimation, sequential Markov chain Monte Carlo methods, sequential Importance sampling for static parameter estimation, the ensemble Kalman filter, online pseudo-likelihood method. Overall this is a good coverage of methods which is then followed up by a numerical comparison. This is indeed a useful pre-cursor to Chapter 5 where the state-space formulation and its associated computational method will the applied to the application focus of this thesis. The numerical example is very simple though, it concerns an Gaussian AR(1) model which already has been covered in numerous other works in the literature. It would have been better if a more substantial test case could have been used. The conclusion of the chapter does not present a summary of the outcomes and justify the method the author intends to take forward to Chapter 5. Also the Chapter 4 uses Algorithm 5.2 which is not discussed at all in the chapter. I myself struggle to draw a conclusion from the trace plots and box plots presented as the difference of the quality of the estimates are very small. The main issue with the chapter is that it has numerous typographical errors, some of which I have listed below. The other issue I have is that some techniques seem out of place, namely the ensemble Kalman filter, which is a method for inference in high- dimensional state-space models where the state and observation at each time can be vectors comprised of thousands of variables each. This does not appear to be true of the application the author has in mind. The second is that the author has not explained how the application (inferring a planar trajectory with GPS measurements) fits into the state-space modeling framework. Indeed I am unsure what the static parameter would be for the vehicle planar trajectory application.

3.5 Chapter 5

Chapter 5. This chapter applies the paradigm of Chapter 4 to inferring a planar trajec- tory with GPS measurements. The chapter contains original research but is somewhat confusing and meandering. The chapter suprisingly starts with a linear Gaussian state-space with unknown co-variance matrices paramerised by . This is a different and far simpler model compared to the general discussion in Chapter 4. How the discussion in Sec 5.2.1 relates to the parameterised covariance matrices of the previous section is unclear to me. This sec- tion needs to be carefully rewritten to better explain its flow of ideas and its context. I cannot see how (5.9) relates to the model in (5.1) and (5.2). I do not understand why conjugate priors were not used for the unknown covariance matrices R and Q in which case a Gibbs sampler can be implemented exactlty. Section 5.2.3 needs to be better justified and its context better explained.

Algorithm 5.1 is introduced without any run-in. It does not appear to have been discussed in Chapter 4 or alluded to therein. This makes Chapter 4 slightly redundant since no algorithm from there has been caried forward. Alg 5.1 appears to me somekind of population MCMC method which is adapted. The adaptation strategy (line 5) is not a conventional strategy I recognise, e.g. does not employ a stochastic approximation method. Also, adaptation does not appear to be diminishing, which can severly alter its convergence. Again a new concept, delayed acceptance, is introduced without any foreword. I am not sure why delayed acceptance is relevant for the main application of the thesis. Section 5.4.2 introduces a new concept, a continuous time-series models. Its purpose or relevance is unclear to me. (Irregularly spaced data can be adequately analysed in a discrete time framework which is, hiterto, the framework of choice of the thesis.) All these concepts then converge with the statement of the main algorithm of the thesis, Alg 5.2, which is a sliding-window delayed acceptance method. Why havent the results of this chapter compared with the spline solution of Chapter 2? Overall the chapter has merit but is confusing. Computational methodology (e.g. population MCMC, adaptive MCMC) are conflated with the modelling of the appli- cation and neither have been adequately described. It needs to be rewritten in its entirety, trimmed of any superfluous material. To enhance clarity, there needs to be a dedicated chapter to the main application of the thesis, covering the mounted sensors, the available measurements and then the overall mathematical formulation (discrete and continuous time) in detail. This is essentially a modelling task and should be free of any computational considerations. Computational methods can then be argued for and introduce subsequently to the model being properly declared.

In the introduction, the motivation of this thesis and the usage of real GPS data are clearly written. However, there is no separate description of the data. It would be easier to engage simulation studies if there is a subsection about the data. In page 92, (Section 5.2.4), does the equation 5.6 mean that a joint distribution of (x1,x2,...,xT,y1,y2,...,yT) is a multivariate normal distribution with zero mean and covariance matrix of ? Given a generic model in the equation 5.1 and 5.2, I feel that a joint distribution of N(0,) sounds like a very particular model. Could you write types of models in which has this particular form of the joint distribution? Inpage 92, under the assumptionoff(x1,x2,...,xT,y1,y2,...,yT—)=N(0,),apreci- sion matrix 1 is formed then the covariance matrix is recovered. Could you write the 1 A B1 motivation of using the particular precision matrix = B1 B and explain what this form means in the model for (x1,x2,...,xT,y1,y2,...,yT)? For example, if 1 is a matrix with two blocks (A and B and zero-off diagonals), we safely say that x and y are independent. In Equation 5.9, 0.5 log(—1 —) = 0.5 log(tr(B))log(tr(L))+log(tr(R)). I think that this YY is only true when B, L and R are diagonal matrices. Are they always diagonal matrices? Doesnt it rather depend on models? In page 104, the local trend model is used as an example.yt—xt N(xt,2), xt—xt1 N(xt1,2) I do not understand why the joint distribution (x0,x1,...,xT,y1,...,yT) is N(0,) and, havent seen a reference with it. Do you have a reference? Or could you explain? Following the above question in Section 5.4.1, I do not understand how you get this particular form of precision matrix and how its inverse represent the local trend model. If this is your calculation, could you describe it? If not, could you add relevant references? The above two questions are related to the other considered models in Chapter 5. Could you give some explanation or add references? Have you used the real GPS data in Chapter 5? Then can you clearly indicate which result is for the real data? 2 If you used the simulated data, please write the true values with the inference result to diagnose the result easily.

3.6 Others

4 Minor

There are numerous grammatical or presentational flaws that need to be corrected. Please refer to the following detailed comments, and the markings in the returned hard-copy.

Minor comments In Section 1,please check the references. The first example is that in Section 1.2, the first sentence should be Smoothing spline ..... reconstruction. See Eubank (2004) and Durbin and Koopman (2012) for details. There are many of these in Section 1. Page 5, Algorithm 1.1 : the notation tk+1, t2k+1, . . . is not distinguishable to t2, ... Please use a different notation. Page 5, Second paragraph in Section 1.4 : Please rewrite For example a posterior estimation of xt p(xt—y1:t) ...for incorrect estimates ...... Page 7 : Describe Rt and Qt. Page 9 : The first equation, E[p(yt—y1:t1,)]? Page 23, Section 2.2.2 : Please rewrite the sentence N1(t1) = 1, ... This is confusing. Page 25, Section 2.2.3 : What is m in m2/t2? Page 26, Equation (2.24) : Isnt it 1/(n 1) instead of 1/n? Page 81, Algorithm 4.8 : Draw (i) p(—y1:k) seems to duplicate to the previous sentence. Page 83 : Please describe the reason of your choice of sufficient statistics st. Page 84 : It was observed that the LW filter has a larger distance to the true parameter, can you describe the reason? Page 92, Equation for : Change B to Bt. Page 104, 1 is (2t+1)(2t+1) and expression is for 2t2t. Please change it for the correct dimension. Page 120. Section 5.5.5 : What is Eff? Page 120, Second sentence in Section 5.5.5 : What is same dataset? Please write in details. Page 129 : Figures 5.15-5.17 are not discussed in the section. Appendix C : It is not clear how Appendix C is relevant to the main part. Please describe it in the main sections. Typographical/grammatical/stylistic Page 183: Please write the full details for the publication, Atchade, Y. et al (2009) Page 185: Please write the full details for the publication, Bodewig, E. (1956) Page 188: Please write the details correctly for the publication, Doucet, A. and Johansen , A. M. (2009) 3 Page 190: Please write the details correctly for the publication, Gelman, A, et al. (2096) Page 194: Please write the full details for the publication, Kokkala, J, et al. (2016) Page 197. Roberts, G. O. et al (1997) : Change to Annals of Applied Probability. Page 199: Please write the full details for the publication, Sorenson, H. W. (1985) Page 200: Please write the full details for the publication, Syed, A. R. (2011) Page 201: Please write the full details for the publication, Wahba, G. (1990) Page 202: Please write the full details for the publication, Wolberg, G. (1988) In Reference, many journal publications are listed without volumn/number.

The thesis details an important application for which there are novel and worthy contributions. The execution though is poor, which is especially true of the chapters on the state-space approach and less so for the spline approach. Part to the thesis is on a contemporary problem area concerning the enhancement of the particle filter methodology and draws on recent advancements and best practise from the literature. The thesis attempts to present an account of the improvements although not convincing by virtue of being poorly written up. The author has developed skills in a range of statistical methods and has produced potentially publishable work - the second and fifth chapters that is. Typographical Errors and Issues to be Resolved Chapter 4. (1) Section 4.2 makes reference to a target-tracking system which is undefined. (2) P66: xk1instead of xt1 in displayed equation. (3) ... are drawn from the target probability density function p(x). What is p(x)? Final displayed equation has no p(x). (4) Sec 4.2.2 diplayed equation has dxx? (5) Final displayed eqn before sec 4.2.3 has p(y — x(i)) instead of p(y — x(i)). 1:t 0:t t t (6) Firsline of sec 4.2.3 states qt must be of the displayed form so that the importance weight can be updated sequentially. This contradicts the sequential update rule of the importance weights given in final displayed eqn before sec 4.2.3. (7) Page 69 onwards uses w(i) for the normalised importance weight whereas eqn t (4.4) uses tilde for the normalised weight. Alg 4.1 line 3 then suggests the tilde weight is unnormalised. Similar issue in Alg 4.6. (8) Page 62 equantion missing brackets. (9) Alg 4.3 line 5 in defn of : how is p(xk, xk1 — y1:t) available to be used? (10) Alg 4.7: in line 3 the observation yt+1depends on the sufficient stat instead of xt+1? No particle indices present in line 4.