EPSRC Faraday Partnership for Industrial Mathematics
- administered by the Smith Institute

Final Report on the Project August 30, 2005

Real-timE ModellIng of Nonlinear Data streams (REMIND)

Overview

Significant progress was made toward the objectives stated in the grant proposal, the most complete advances being in the analysis of data streams. There was some change of focus at the request of our Industrial partners, and new industrial connections were made, leading to the application of our approach to multi-model forecasts in an area not originally envisioned (electricity demand). Some of the original objectives proved overly ambitious given the quality of the data and models available; as a result the study was broadened to consider the effects of data quality. Specifically: data streams of the duration, sampling rates, and noise levels required to demonstrate the full value of the application of MCMC techniques are not currently available. We believe that the impact of these data quality issues may be generally under-appreciated within the MCMC parameter estimation community, and we have addressed the issue, suggesting tests of internal consistency. There is ongoing interest within industry to enhance operational data streams and exploit the techniques reported below. Explicit comments on each of the original objectives and our aims for the intended beneficiaries (as stated in the original proposal) are given at the conclusion of this report. The report follows the development of the research itself. In short, while we did not achieve a "unification of structural and probabilistic modelling which is robust in the face of real-world constraints", this was in large part due to huge differences in the utility between the two types of models in different applications. Of the two applications targeted in the proposal, a data-based time series model was clearly superior for one, while a stochastic structural simulation model was clearly superior in the other. Unification of the approaches remains a goal, given an application where models of both sorts have complementary skill; we have clearly illustrated how to proceed in the target applications, providing new methods and identifying the most relevant operational constraints on their deployment.

Summary

The project Real-timE Modelling of Nonlinear Data Streams (REMIND) had at its core the aim of melding advances in nonlinear dynamics with those of Bayesian statistical estimation; this was attempted in both mathematical systems and real industrial settings [2]. In each case, we maintained the realism of a data stream setting: a situation in which real-time reaction is required, and where data is provided at a rate that precludes complete reliance on online storage and post processing. The industrial data streams considered under REMIND consisted of (i) second by second observations of variations in grid frequency of the UK electricity grid and minute by minute variations in electricity demand, both provided by National Grid Transco (see [2] for details) and (ii) multivariate observations from a Hurricane Gas Turbine in various states of spin-up, spin-down and steady load states, collected by Dave Mellor of Intertec during a site visit to Siemens-ALSTOM labs in Lincoln with both the PI (LAS) and RA (LC) attending (see [1] for details). Data streams for mathematical

models came from a collection of low dimensional nonlinear deterministic models (see [4, 5, 6, 8] and nonlinear stochastic models (see [2, 4, 7]) with which the investigators have familiarity. Early in the project, connections within the energy sector led to an additional industrial application on forecasting electricity demand; while not a data stream application, this work led to new insights on the Bayesian methods of processing multi-model forecasts [3].

The first step was to develop the means for evaluating the relevance of model classes, for a given problem, and improved means of building deterministic nonlinear models for data streams. A new test for long-range persistence was developed (see [7]); passing this test would suggest using a model class that permits slowly decaying correlation functions. Despite widespread claims of long-range persistence, we found little indication that estimates of persistence were robust either in the grid data, or in other geophysical series analyzed by others. In order to apply traditional nonlinear local polynomial model techniques to data streams, we developed a method of selective incorporation [5] which has proven very effective for modelling chaotic systems. Neither Intertec nor NG Transco are interested in prediction *per se*, but rather in condition monitoring and parameter estimation, respectively. Condition monitoring requires the (early) detection of changes in operating characteristics. Under REMIND we have extended and applied a technique previously suggested by the PI in un-refereed literature [8] which uses low-dimensional reconstructions, with the ability to incorporate the forecast models of [5]. Much modern condition monitoring is based on frequency space techniques, using Fourier methods to detect changes in observed spectra; proposed nonlinear methods are often based upon the estimation of scaling statistics which are quite difficult in practice (see [7] and reference therein). The method developed by Clarke and Smith [1] under REMIND can detect and quantify changes that are literally invisible in frequency space, hence the title “Detecting Transparent Noise”. The technique was illustrated on data collected jointly with Intertec. Intertec is a small to medium size engineering company who want to extend their monitoring techniques in order to gain market differentiation. Our analysis technique complements existing frequency-based methods and can be applied using existing monitoring technologies. In order to prove its practical benefit in fault detection we would need to test the method on data for which a known fault had occurred; such data is not yet available. The method does successfully identify changes in Hurricane data stream as the turbine is placed in different settings. The development of both predictive monitoring [5] and data-library based monitoring [1] allows us to adjust to Intertec’s operational constraints (the computing and data storage limits of their detector). In short, our method is designed to complement existing frequency-based methods (and likely to replace the alternative nonlinear methods noted in [1]) and to extend the usefulness of condition monitoring with particular applications to data streams.

A major thrust of the REMIND project was the development of a Bayesian parameter estimation methodology, using MCMC methods, for both deterministic and stochastic models. A number of fairly fundamental difficulties with proposed applications have been documented under REMIND [9] and by others during this time [4]. Under REMIND we developed a simplified model of the UK's electricity grid [2] and went about parameter estimation using that model. When run in perfect model mode, with high sampling rates on the observable quantities, the MCMC approach worked well. When run using current operational sampling rates (a few hertz for grid frequency,

around 1 to 15 minutes for demand), the method did not converge; more precisely: common evaluation statistics of the MCMC method did stabalise but not about physically plausible parameter values . A good deal of time was spent investigating methods to apply the MCMC algorithm, to determine whether the problems were (a) in the MCMC method, (b) in the fact that the mathematical structure of our model was imperfect, or (c) that the quality of the operational data stream failed to provide the information required by our MCMC algorithm. In short, the answer was (c) as demonstrated in reference [2], where we not only address particular questions in the analysis of grid frequency, but also suggest sanity tests for establishing the *meaningful* convergence of MCMC analyses of time series in general.

Joint research continues with NGT, the possibility of acquiring data at higher sampling rates in the future is a real one; the research into MCMC modelling has attracted the interest of other Bayesian modelling groups (Jim Berger at Duke) and is likely to lead to a better understanding of the application of these methods to simulation models with a dominant deterministic component. Other examples are in preparation [9]; the combination of MCMC and alternative methods of parameter estimation will form the core of M Cuellar's PhD thesis which will be submitted by the end of 2005. (Her studentship was funded by NGT under REMIND.)

Objectives

We next consider each of the five goals stated in the proposal.

1. Construct a consistent approach to the unification of structural and probabilistic modelling which is robust in the face of real-world constraints and applicable to unending data-streams.

A methodology for combining structural and probabilistic modeling techniques was developed [2], but there was no clear "unification" since, in most cases, one of the different modeling approaches tended to dominate in each individual application. The Bayesian MCMC approach was shown to work given data sampled at sufficiently high sampling rates but was unable to perform given the data quality even in the perfect model setting (where the data was generated by the model itself, and then subsampled to reflect the properties of the operational data stream). The methodology for the NGT branch of the project was developed with large, but finite, data sets. *It remains to develop Bayesian parameter estimation algorithms that update online with each new observation; NG Transco has expressed interest in funding this development, once they obtain operational data streams of sufficient quality for the MCMC method to meaningfully converge.*

2. Advance our ability to account for uncertainty in the derivation, estimation and verification of nonlinear models, with emphasis on employing ensembles over models of different mathematical structure.

By incorporating stochastic elements into structural models of the UK electricity grid we have advanced our ability to account for uncertainty in both the derivation of our models and the estimation of model parameters[2]. The output from this structural model class, namely predictive distributions for quantities of interest and distributions for model parameters, enables us to quantify the uncertainty with respect to

verification. The poor quality of the empirical time series models precluded a multi-model approach in the case of NGT. *It remains to develop a real-time estimation system, this requires improvements in the operational data stream which NG Transco expects within two years.*

In terms of the Huricane turbine data, the single models developed in [1] appear to exploit the information in the datastream efficiently; this aspect of the condition monitoring problem is effectively solved until time series data on specific faults is available.

The electricity demand forecasting application [3] provided an opportunity for multi-model forecast interpretation. This produced quite interesting results, in that traditional Bayesian methods failed to outperform rather ad hoc methods. These insights have been transferred to the DIME Faraday project and will contribute to additional publications under that program. It has lead to significant conversation with the modelers (weather modelers at the European Centre for Medium-range Weather Forecasts in Reading) and Bayesian statisticians at SAMSI (see below). *It remains to determine how widely these results generalize, and the general effectiveness of Bayesian updating in the imperfect model scenario.*

3. Provide a viable real-time monitoring framework for grid frequency, able to rapidly identify loss of generation, estimate impact of a given loss of generation, and provide estimates of time varying parameters.

The project did not progress sufficiently to be able to provide a real-time monitoring system for grid frequency for NGT, although the general framework was been constructed. Practical issues in a real-time system will turn on the eventual data quality. *It remains to combine the two data stream approaches (the forecast methods [5] and the library based condition monitoring [1]) with the MCMC model based parameter estimation; given expected improvements in the quality of the operational datastream, this would be a fascinating project.*

4. Provide more general methods for novelty detection in rotating machinery.

A novel method for condition monitoring has been presented [1] that is able to detect changes that are invisible in frequency space and is designed specifically for data streams. It is more stable than other proposed 'nonlinear methods', can detect transparent noise that is literally invisible to any frequency space analysis and has been applied to the Hurricane turbine data. *This goal was fully achieved.*

In addition to the above, a further objective is to use the Smith Institute Faraday Partnership as a means of disseminating the new insights and techniques that the project will generate to other researchers and industrial organisations and to give them the opportunity to stimulate new applications in related areas.

The importance of the Smith Institute to the success of this project, in terms of interaction with the original industrial partners, provision of domain knowledge and distribution of new insights and techniques to the broader industrial community cannot be overstated. REMIND was the first project on which the PI had a funded "technology translator;" all the PI's future grants which hinge on collaboration with

industry will include requests to fund this role. [This includes a NERC Knowledge Transfer grant approved earlier this month.] Without question, the level of dissemination within industrial sectors was greatly enhanced.

Beneficiaries

The principle beneficiaries of this work are the individual partners linked directly to the project: National Grid Transco (NGT) and InterTec, although researchers at SAIC and the Cal ISO electricity demand forecast team, as well as academics at the Statistical and Applied Mathematical Sciences Institute (SAMSI), have incorporated aspects of REMIND research into their work. National Grid Transco have gained significant insight into Bayesian parameter estimation techniques applied to industrial problems and have come to better understand the data quality issues that prevent a full application of the method. The company is interested in pursuing this approach and is looking to work with LSE CATS beyond the end of the project on these methods. Through this work, InterTec have been exposed to new fault detection analysis techniques that, if adopted, could well lead to unique condition monitoring products. Beyond these immediate partners, there are more general beneficiaries in the condition monitoring industry and general industrial modelling community.

As many of the results are not yet in print, there has been limited chance for new collaboration at this point in time. The continued interest of NG Transco to potentially fund the research illustrates its perceived value; new collaboration with SAIC (Dr Mary Altalo) is also being pursued in the area of systems monitoring (and cross-pollinated with our DIME Faraday on applications of weather forecasting). No additional collaborations have been identified as yet.

Generalizations to better account for model inadequacy has produced significant academic interest, both in the UK and the US and were a major focus of discussion during the six month SAMSI program (www.samsi.info) in early 2005 where the PI was the 2005 University Fellow.

References (REMIND PAPERS are [1,2,3,5,6,7,9,10])

[1] L. Clarke and L.A. Smith (2005) Detecting transparent noise. *Mechanical systems and signal processing, Elsevier*. (In Review)

[2] M. Cuellar, L. Clarke, M. Brown, and L.A. Smith (2005) The Role of Operational Constraints on MCMC Parameter Estimation: The case of the UK Electricity Grid. *International Journal of Power Engineering and Energy Systems, Elsevier* (In Review).

[3] L.A. Smith, M.G. Altalo & C. Ziehmann (2004) Predictive Distributions from an Ensemble of Forecasts: Electricity Demand Forecasts from Imperfect Weather Models. *Physica D* (accepted with minor revisions)

[4] K. Judd and L.A. Smith (2004) Indistinguishable states II: The imperfect model scenario. *Physica D*, **196**:224-242. K. Judd (2003) Right Results for the Wrong Reasons? *Physical Rev* E, **67**, 026212.

[5] F. Kwasniok and L.A. Smith (2004) Model reconstruction from data streams. *Phys. Rev. Lett.*, 92(16).

[6] L.A. Smith, M. Cuellar, H. Du and K. Judd (2005) A Geometric Approach to Parameter Estimation in Nonlinear Systems. *Phys Rev Lett* (In Review).

[7] L.A. Smith and A. Guerrero (2005) A Maximum Likelihood estimator for Quantifying Long-range Persistence. *Physica A*, **355** 619-632.

[8] L.A. Smith, K.Godfrey, P.Fox, and K. Warwick. A new technique for fault detection in multi-sensor probes. In *IIE International Conference on Control `91*, volume 2, page 1062, 1991.

REMIND papers still in Preparation

[9] M. Cuellar and L.A. Smith (2005) On the Curious Behavior of MCMC Parameter estimation in the Logistic Map (to be submitted to *Phys Lett A*)

[10] L.A. Smith H. Du (2005) Using ISIS to Resolve the Problem of Chaotic Likelihoods (to be submitted to *Physica A*)