Modelling Sequential Healthcare Data: a Preliminary Survey

Alex Bird

January 2017

Contents

Chapter 1

Probabilistic Models for Monitoring Data

When considering sequential data, the dependence between observations must be modelled effectively. While independence may be assumed for some models (e.g. Shumway and Stoffer [2010] §2), typically the assumption is violated significantly, with a corresponding loss in predictive performance. In this section we first address the question of what dependence can be assumed while retaining some mathematical tractability, and secondly how these models have been extended and adapted for applications in medical time series.

1.1 Sequential models

1.1.1 Autoregressive process

We choose autoregressive (AR) processes to model the most basic assumption of dependence between observations. For $y \in \mathbb{R}$, an AR(p) process is defined:

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, q)$$

which can be written equivalently as a vector AR(1) process for $\triangleq (y_t, y_{t-1}, \ldots, y_{t-p})^\mathsf{T}$,

$$= A\mathbf{y_{t-1}} + G\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, q),$$

$$A = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_p \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^\mathsf{T}.$$

Inductively, any AR(p) process (vector or not) can be written as a vector AR(1), or Markov process. Thus while long-term dependencies can be captured, mathematically we only need consider first order dependencies. The appropriate graphical model is given in Figure 1.1. In the general (unconstrained) case, the model parameter A can be learned using multivariate least squares.

The autoregressive model captures the idea that observations that occur in similar positions in time have similar values. However the process is primarily driven by mathematical convenience rather than practical significance. In practice, data often exhibit non-linearity or non-stationarity in relationship to previous values. Furthermore, the model assumes that the process is perfectly observed, with no measurement error. There are many instances where sequential data are not fully observed. For example, consider a body acting under Newtonian mechanics, such as a car. One might consider that the acceleration follows

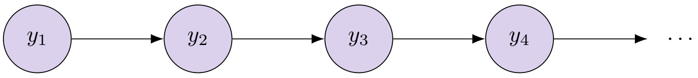


Figure 1.1: AR(1) Process

an auto-regressive process according to the forces that are acting on it in an observation window. And so the spatial co-ordinates will not submit to a predictive model in this regard, it is the 'unobserved' second derivatives that may be modelled as an AR process.

It is also worth considering why the model is to be fitted to the data. Sometimes we are interested in predictive performance, in which case we require that the model is a powerful description of the generative process. However, at other times the dynamics might be a nuisance process which must be eliminated before inference can be performed. For example, in credit risk, time inhomogeneity in covariate response may in some cases be alleviated by fitting a time varying component corresponding to unknown macroeconomic factors. The distinction is important, since many time series models may be fit to the data, but as for the case of a random walk model, it does not imply that the model has captured a description of the data. If prediction of the dynamic process is not crucial, a simple model of an AR process may be entirely appropriate.

There is a lot more to say here about classical time series models and their extensions: constraints, differencing, exogenous inputs (covariates), and various different models. In the interest of brevity, the reader is referred to (for example) Shumway and Stoffer [2010] for an overview of this area.

1.1.2 Linear Dynamical Systems

A linear dynamical system (LDS) is a model for which we believe the underlying generative process is an autoregressive one, but the observation is a noisy transformation of it. In the language of control theory, the unobserved value at a given point in time is known as the *state*.

This model was first discovered by Kalman in the seminal paper of 1960 Kalman [1960] in the Engineering community. It has enjoyed phenomenal success in tracking problems (for example, radar and weather forecasting) and autonomous guidance (for example in the Apollo program) among others. This survey will take a particular Bayesian interpretation for continuity of presentation, but it should be emphasised that it was introduced simply as the solution to the least squares problem for state estimation. As with linear regression, the Kalman Filter enjoys the property of being the Best Linear Unbiased Estimator (BLUE) regardless of the distributions involved[1].

The approach developed by Kalman for inference of the state variables is equivalent to inference using the sum-product algorithm of the model:

$$p(\mathbf{x}, \mathbf{y}) = \prod_{t=1}^{T} p(\mathbf{x}_t | \mathbf{x}_{t-1}) p(\mathbf{y}_t | \mathbf{x}_t) \tag{1.1}$$

where $\mathbf{x}_0 \triangleq \varnothing$, $\mathbf{x}_t \in \mathbb{R}^n, \mathbf{y}_t \in \mathbb{R}^d$. All distributions are assumed Gaussian; the transition distribution:

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(A\mathbf{x}_{t-1}, Q)$$

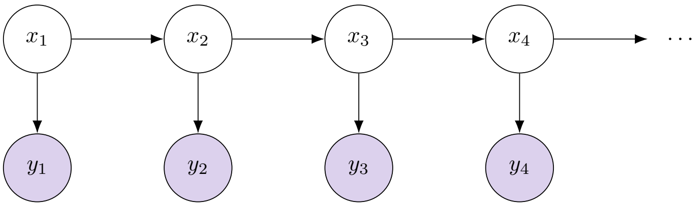


Figure 1.2: Latent Markov Model

[1] It is assumed that the errors are uncorrelated with mean zero and homoscedastic (finite) variance.

and the emission distribution:

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(H\mathbf{x}_t, R).$$

See Figure 1.2 for the graphical model. The inferential procedure to estimate the latent states is given in appendix B.

A similar model was proposed by Baum and Petrie [1966] which uses discrete latent states instead of continuous ones (as in the LDS), known as the Hidden Markov Model (HMM). In some ways the HMM is much more expressive than an LDS in that it is able to capture arbitrarily complex relationships in the state evolution. Unfortunately this comes at an exponential cost in the number of states: when needing to retain n bits of information in the latent state, 2^n states are required. If it is known that all latent states evolve and interact linearly, clearly using an LDS is advantageous. Nevertheless, HMMs have proved extremely effective in tackling problems such as handwriting recognition, speech transcription, and gene prediction among many others.

1.2 Regime-switching sequential models

Linear dynamical systems are extremely powerful in the instance where the dynamics follow a fixed pattern (such as a linear physical law), but are unable to capture changes in the dynamics, or different regimes in the action of the latent state. This form of time-homogeneity is only of use for modelling gradually evolving or 'steady-state' models; for our applications in healthcare we require more expressive models.

1.2.1 Auto-regressive Hidden Markov Models

Perhaps the simplest such 'switching' model that permits dynamic transitions of a time series is an auto-regressive Hidden Markov Model (ARHMM). (It is usually inadvisable to use HMMs directly since the observations are assumed independent given the discrete latent state; a sequential mixture of Gaussians.) ARHMMs have gained notable success in econometric modelling and speech recognition.

The model is similar to an HMM, but the observations follow a conditional autoregressive process (see Figure 1.3). For S regimes, $\mathbf{y}_t \in \mathbb{R}^d$ and $x_t \in \{1, \ldots, S\}$:

$$\begin{aligned} p(x_t = i|x_{t-1} = j) &= \phi_{ij}, \\ p(\mathbf{y}_t|\mathbf{y}_{t-1}, x_t = k) &= \mathcal{N}(\mathbf{y}_t|A^{(k)}\mathbf{y}_{t-1} + \mathbf{b}^{(k)}, R^{(k)}). \end{aligned}$$

The model above follows an AR(1) process in the observations: this can be extended by adding in additional previous observations, but as previously observed, all AR(p) processes may be cast in AR(1) form. Inference can be performed in the model in the same way as an HMM, but the emission distribution gains dependencies to the previous observation(s). The filtering (forward) messages can be derived as:

$$\begin{aligned} \alpha_t(i) = p(x_t = i, \mathbf{y}_{1:t}) = \sum_{x_{t-1}} p(x_t = i, x_{t-1}, \mathbf{y}_t, \mathbf{y}_{1:t-1}) \\ = p(\mathbf{y}_t|x_t = i, \mathbf{y}_{t-1}) \sum_{x_{t-1}} p(x_t = i|x_{t-1}) p(x_{t-1}, \mathbf{y}_{1:t-1}) \\ = \mathcal{N}(\mathbf{y}_t|A^{(i)}\mathbf{y}_{t-1} + \mathbf{b}^{(i)}, R^{(i)}) \;\; \boldsymbol{\phi}_{i,\cdot}\boldsymbol{\alpha}_{t-1} \end{aligned}$$

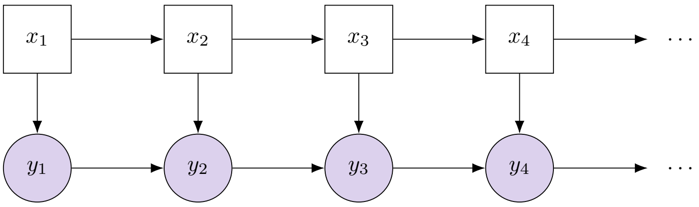


Figure 1.3: Autoregressive AR(1) Hidden Markov Model

where $\boldsymbol{\phi}_{i,\cdot}$ is the ith row of the transition matrix, and the usual re-normalisation is advised for numerical stability. Backward messages can be amended from the HMM in a similar way. Barber [2012] suggests in the case of high frequency data to limit the switching to a subset of time points $\mathcal{T}_s \subset \{1, \dots, T\}$ by setting the transition matrix to the identity for $\{1, \dots, T\} \setminus \mathcal{T}_s$, (presumably) for computational reasons.

Stanculescu, Williams, and Freer [2014] showed how sepsis may be detected early from neonatal ICU data using an ARHMM (with a discrete emission distribution). Two regimes were used (S = 2) corresponding to $\{$*normal*, *sepsis*$\}$, and the model was learned in a 'supervised' way using annotations of the data by clinicians. A well-documented deficiency of HMMs is the geometric distribution of state duration which was ameliorated by duplicating states in the hidden topology.

1.2.2 Switching LDS

A switching LDS (SLDS) is a piecewise fixed LDS, where a discrete latent state chooses the regime to follow according to a Markov process. Whereas the ARHMM is is an AR process conditioned on a latent Markov switch process, the SLDS is an LDS conditioned on the Markov switch process. The model is shown in Figure 1.4 and for $\mathbf{x}_t \in \mathbb{R}^d$, $\mathbf{y}_t \in \mathbb{R}^n$, and $s_t \in \{1, \dots, S\}$ is described by:

$$\begin{aligned} p(s_t = i | s_{t-1} = j) &= \phi_{ij}, \\ p(\mathbf{x}_t | \mathbf{x}_{t-1}, s_t = k) &= \mathcal{N}(A^{(k)} \mathbf{x}_{t-1},\ Q^{(k)}), \\ p(\mathbf{y}_t | \mathbf{x}_t, s_t = k)) &= \mathcal{N}(H^{(k)} \mathbf{x}_t,\ R^{(k)}). \end{aligned}$$

The SLDS is a frequent extension of the LDS, and applications include aeroplane tracking, and anomaly detection at industrial plants. The similarity with ARHMMs is obvious, but while SLDS is a more general model, it suffers from greater computational complexity. A different deficiency shared with the ARHMM is the inability to model multiple regimes simultaneously. We will see how this can be addressed in the following section of this report.

Inference in an SLDS is more challenging than in the LDS. In an (AR)HMM we are able to use an efficient algorithm for state space inference that is quadratic in the number of states $O(S^2 T)$, despite the presence of exponentially many possible paths (S^T). Unfortunately in the case of the SLDS, the same dynamic programming (DP) approach cannot work. Mathematically, the marginalisation step at time t is

$$p(s_t, \mathbf{x}_t | \mathbf{y}_{1:t}) = \sum_{k=1}^{S} \int \mathrm{d}\mathbf{x}_{t-1} p(s_t, \mathbf{x}_t | s_{t-1} = k, \mathbf{x}_{t-1}, \mathbf{y}_t) p(s_{t-1} = k, \mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}).$$

By induction on time $t = 1$ which is a mixture of S Gaussians, the filtered posterior at time t has S^t Gaussian components. In the ARHMM, the continuous process is observed, and only the discrete transition must be marginalised, which can be done exactly using DP. It is the marginalisation over the combined (discrete, continuous) process which causes computational intractability, and therefore requires approximation[2]. The approximation usually takes the form of particle filtering or assumed density filtering – minimising the KL-divergence of the exact posterior to that of a simpler posterior (typically a mixture of Gaussians with a fixed number of components).

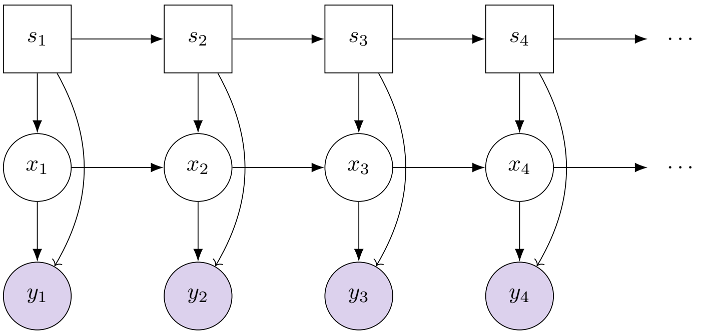


Figure 1.4: Switching Linear Dynamical System

[2] for a technical discussion of the hardness of inference in an SLDS see Lerner and Parr, 2001.

1.2.3 Factorial Switching LDS

The Factorial Switching LDS (FSLDS) attempts to improve the efficiency of modelling multiple hidden latent chains that may be considered to evolve independently. If m such chains are required, the SLDS would require 2^m switch settings to accommodate all possible interactions (for binary switches). The FSLDS reduces this burden under the assumption that each latent chain transitions independently and contributes the same input to the observation function independent of the state of the others. This assumption decreases the total number of states for binary switches to $2m$, a considerable improvement. Figure 1.5 shows an example of such a graphical model. Unfortunately, even though the generative process is assumed independent, the posterior over the latent variables becomes coupled through 'explaining away' (see e.g. Koller and Friedman [2009] §3.3), and exact inference is no easier than in the equivalent SLDS.

For m latent chains, the basic FSLDS is defined for observations $\mathbf{y}_t \in \mathbb{R}^d$, the continuous AR latent chains $\mathbf{x}_t^{(k)} \in \mathbb{R}_k^d,\ \ k \in 1, \ldots, m$, and switch variables $s_t^{(k)} \in \{1, \ldots, S_k\},\ \ k \in 1, \ldots, m$ as

$$\begin{aligned} p(s_t^{(k)} = i | s_{t-1}^{(k)} = j) &= \phi_{ij}^{(k)}, && k \in \{1, \ldots, m\}, \\ p(\mathbf{x}_t^{(k)} | \mathbf{x}_{t-1}^{(k)}, s_t^{(k)}) &= \mathcal{N}(A^{(k, s_t^{(k)})} \mathbf{x}_{t-1},\ Q^{(k, s_t^{(k)})}), && k \in \{1, \ldots, m\} \\ p(\mathbf{y}_t | \mathbf{x}_t^{(1)}, s_t^{(1)}, \ldots, \mathbf{x}_t^{(m)}, s_t^{(m)}) &= \mathcal{N}(H^{(1)} \mathbf{x}_t^{(1, s_t^{(1)})} + \ldots + H^{(m, s_t^{(m)})} \mathbf{x}_t^{(m)},\ R^{(s_t^{(1)}, \ldots, s_t^{(m)})}). \end{aligned}$$

Some independence assumptions may be dropped, such as independence between chains, and if overwriting rules are encoded (see discussion below) may make inference easier.

Application to clinical monitoring data

Working with similar monitoring data to Stanculescu et al. [2014], Quinn, Williams, and McIntosh [2009] took a different approach in learning a generative model for the neonatal data. This was approached using an FSLDS to model each of the known effects, both physiological and artifactual with a separate latent chain (see table 1.1). Additional unknown effects are modelled with a so-called 'X-factor', an inflated covariance variant of the steady-state dynamics to determine un-modelled outliers.

Access to carefully annotated data, giving ground truth for each of the latent effects conferred several advantages. Firstly, the authors were able to encode various rules such as which dimensions of the observations were affected by the different effects, as well as a hierarchy of the effects: which effects were capable of 'overwriting' the signal. This means certain features in a subset of channels (such as dropouts)

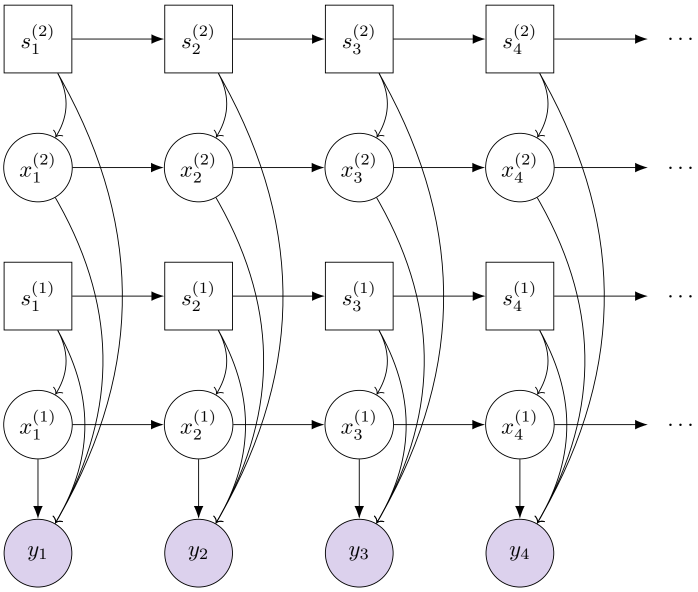


Figure 1.5: Factorial Switching Linear Dynamical System (example)

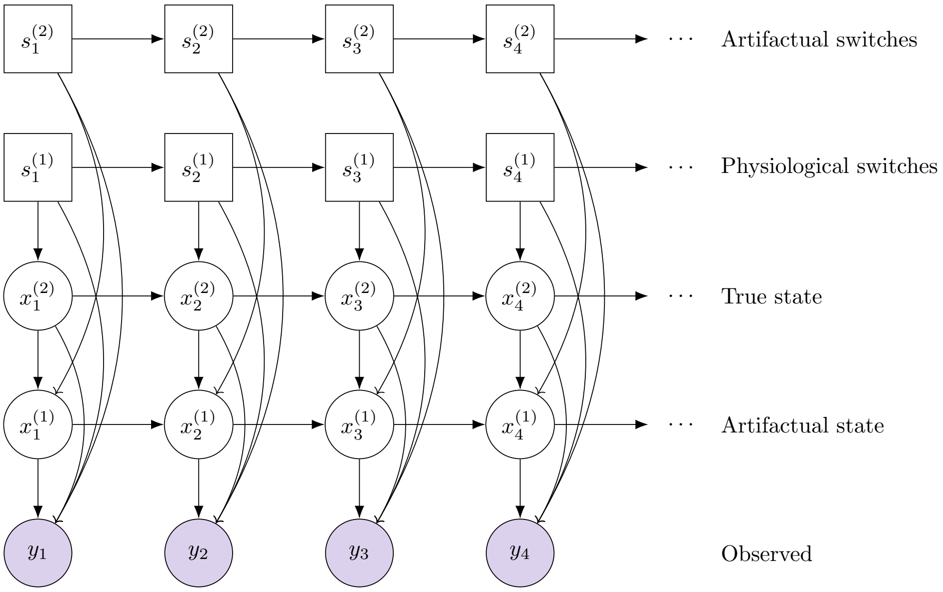


Figure 1.6: FSLDS model used in Quinn, Williams, and McIntosh, 2009

may be assigned by priority to a single effect without requiring expensive inference. Secondly, the model could be trained in a 'supervised' way: conditioned on the (known) switch values, the model can be learned in the same way as an LDS. While this is not necessary in principle, the size of the training set would be prohibitive of training a meaningful model in an entirely unsupervised fashion.

The resulting graphical model shown in Figure 1.6 is slightly different to the canonical FSLDS, with the artifactual representation of physiology subordinate to the 'true' latent physiological state. The generative process is of true physiological effects assimilated by the patient, corrupted with various artifacts and observed noisily. Also, in order to reduce computational burden, the transitions are constrained in such a way that only one factor can change its setting per time step (not shown). The Gaussian-sum approximation was chosen for the inference, as it performed considerably better than particle filtering for the allotted computational budget.

In practice, the parameters of each conditional LDS was fitted using domain knowledge from the properties of each event. Steady state dynamics are assumed to follow various low order ARIMA models across each channel. Temperature decay during probe removal was modelled using exponential decay to the ambient temperature. Blood sample artifacts were modelled using a linear model with drift. TCP recalibration has several distinct stages and was modelled as a Markov model with fixed state transition. Because effects are well understood, and their interactions known, learning in this constrained way mitigates concerns of overfitting.

Event	Type	Brief description
Probe dropout	artifactual	no monitoring data on a channel due to removal or malfunctioning of the monitoring device.
Temperature probe removal	artifactual	slow decrease of the temperature channel due to the removal or disconnection of the temperature probe. The temperature decline follows physical law.
Blood Sample	artifactual	diversion of blood for sampling, causing heart rate readings to desist and an artifactual ramp in the blood pressure due to the saline pump against the sensor.
Incubator open	artifactual	drop in humidity and temperature due to opening of incubator by medical personnel potentially leading to intervention.
Bradycardia	physiological	brief episode of heart rate decline common amongst babies for both benign and serious reasons.
TCP recalibration	artifactual	transcutaneous probes must be recalibrated every few hours to alleviate potential measurement drift – resulting in a distinctive dropout patten.

Table 1.1: Clinical events annotated in neonatal monitoring data (adapted from Quinn, Williams, and McIntosh [2009].)

1.2.4 Discriminative Switching LDS

The Discriminative Switching LDS (DSLDS, Georgatzis and Williams, 2015) was developed in response to the FSLDS under the rationale that greater statistical efficiency could be found through a discriminative inference of the latent effects than a generative one. The DSLDS uses a *conditional random field*[3] approach by omitting a model for the observations (and features derived thereof), and optimising $p(s, \mathbf{x}|\mathbf{y})$ rather than $p(s, \mathbf{x}, \mathbf{y})$. A classifier is trained on the observations at each time t according to a sliding window over indices[4] $t-l:t-r$ to estimate $s_t, t \in \{1, \dots, T\}$. The full model can be written as:

$$p(\mathbf{s}, \mathbf{x}|\mathbf{y}) = \left(\prod_{t=1}^{T} \prod_{k=1}^{m} p(s_t^{(k)}|\mathbf{y}_{t-l:t+r}) \right) \left(p(\mathbf{x}_1|s_1^{(1)}, \dots s_1^{(m)}, \mathbf{y}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t^{(1)}, \dots s_t^{(m)}, \mathbf{y}_t) \right)$$

where $y_t = \varnothing$ for $t \notin \{1, \dots, T\}$. Note that the switches are no longer time-dependent in the DSLDS; this is enforced weakly by the moving window of the classifier. Inference is made easier by the assumption that the posterior latent switch states are independent given the observations. Once the artifactual/physiological effect state is inferred, the 'true' state of the underlying physiology is inferred using the usual LDS updates conditioned on the switch states. The LDS parameters are learned in a similar way to the FSLDS.

Inference of $p(s_t^{(k)}|\mathbf{y}_{t-l:t+r})$ was performed using a random forest for each k. Various features were added to the raw data including linear fit coefficients of multiple subsets, smoothing, differencing, and summary statistics such as the mean and maximum. Inferring the underlying vital signs ($p(\mathbf{x}_1|s_1, \mathbf{y}_1)$) is stil intractable, so the Gaussian-sum approximation was used as before.

The results of the DSLDS were in general better than those of the FSLDS, but not universally, and the DSLDS performed significantly worse for discerning unlabelled abnormalities (the X-factor). An α-mixture of the DSLDS and FSLDS, proposed also in Georgatzis and Williams [2015] was shown to combine the classification of the the two models in a manner that improved on both. An α-mixture of two models p_1 and p_2 is defined by Amari and Nagaoka [2007] as:

$$p_\alpha(s_t) = c \left(p_1(s_t)^{(1-\alpha)/2} + p_2(s_t)^{(1-\alpha)/2} \right)^{2/(1-\alpha)}.$$

Using an $\alpha = 0.5$ mixture of the FSLDS and DSLDS, the classification accuracy of the latent events was uniformly better than either of the individual models.

1.3 Models for control

Dynamical Systems may be augmented with control inputs in order to model system response under external control. We will briefly introduce LDS with control inputs as one of the simplest such models, and is well studied in *control theory* literature. This will motivate the final model in this section, the IONLDS, which is intended to aid medical professionals in administering drug dosages.

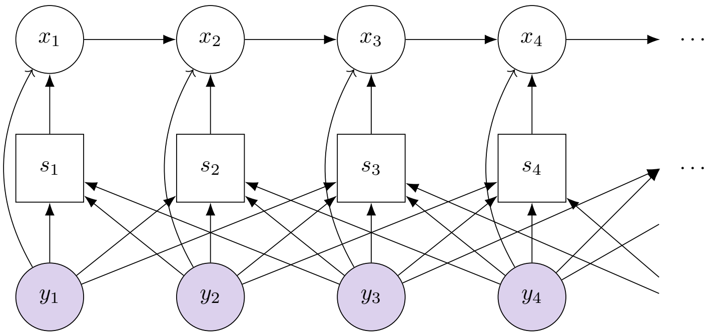


Figure 1.7: Discriminative Switching Linear Dynamical System

[3] see Sutton and McCallum, 2011 for an introduction.

[4] using the shorthand $a_1 : a_2 \triangleq \{a_1, a_1 + 1, \dots, a_2 - 1, a_2\}$ for $a_1 \in \mathbb{Z}, a_2 \in \mathbb{Z}$.

1.3.1 Linear Dynamical System with Control Inputs

Control inputs can exert influence on an LDS in both the latent chain and the emission. This is a canonical model in the engineering community (see e.g. Glad and Ljung, 2000) and used frequently in autonomous navigation and control. The usual form of such a model has the following dynamics:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(A\mathbf{x}_{t-1} + B\mathbf{u}_t, Q)$$

and emission distribution:

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(H\mathbf{x}_t + D\mathbf{u}_t, R)$$

for control signals $\mathbf{u}_t$, $t = 1, \dots T$. Frequently however, only one of the transition and the emission is affected by control inputs. See Figure 1.8 for the graphical model of the latter. For a given set of control inputs and outputs, the model is the same as that of an LDS except with biases in the transitions and emissions as relevant. Due to the compositional closure of Gaussian distributions under not only linear but affine transformations, inference and learning look very similar to the standard LDS.

1.3.2 Input-Output Nonlinear Dynamical System

The IONLDS was developed to model the impact of drug infusion on patient physiology and monitoring data. It is a relaxation of the current state-of-the-art (PKPD model[5]) consisting of a compartmental ODE model of the diffusion of the administered drug about the body, and a nonlinear sigmoid mapping drug concentration to effect on vital signs. Since the ODE can be discretised and written in state space form, the IONLDS is the relaxation from the PKPD dynamics to the maximum likelihood dynamics. Figure 1.8 also shows the graphical model for the IONLDS, since graphical models do not capture the nature of dependencies, and it is defined mathematically for $\mathbf{x}_t \in \mathbb{R}^d$, $\mathbf{y}_t \in \mathbb{R}^n$, $\mathbf{u}_t \in \mathbb{R}^\ell$ with conformable matrices A, B, C, Q, R:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{u}_t) &= \mathcal{N}(A\mathbf{x}_{t-1} + B\mathbf{u}_t, Q), \\ p(\mathbf{y}_t|\mathbf{x}_t) &= \mathcal{N}(h(\mathbf{x}_t), R), \\ h(\mathbf{z}) &= m + \frac{M - m}{(1 + \exp\{-\gamma C\mathbf{z}\})^{1/\nu}}. \end{aligned}$$

While the model is conceptually simpler than that of the FSLDS and DSLDS, intractability is still encountered due to the nonlinearity in the emission distribution. Inference was performed using the Unscented Transform for efficiency reasons against alternative approximations such as Gauss-Hermite methods and the Rao-Blackwellised Particle Filter. Särkkä [2013] is a very readable reference of different approximations to nonlinear filtering problems.

It is left to the experimenter to choose the dimension of the latent space, d. In the experiments in Georgatzis, Williams, and Hawthorne [2016], $d = 4$ was chosen for having the lowest BIC score over the candidate model sizes. Unlike the monitoring models, the parameters of the IONLDS have been

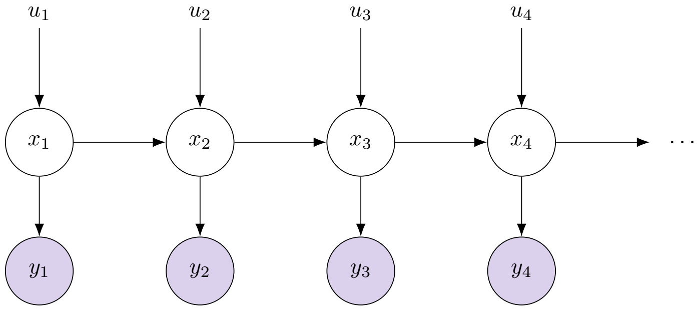


Figure 1.8: Chain model with control in latent space

[5] The Pharmacokinetic-Pharmacodynamic (PKPD) model was developed by Gepts, Camu, Cockshott, and Douglas, 1987, and subsequently developed into a covariate-sensitive model by various authors.

learned in an entirely unsupervised way using Expectation Maximisation (EM). Due to its non-linearity, the parameters of the emission distribution cannot be optimised in closed-form, and an iterative scheme such as BFGS must be employed. A good initialisation must be found for the nonlinear parameters to find a sensible local optimum.

Models trained on each patient have been observed to outperform the current PKPD model. In future work we hope to show how a model tuned only by patient covariates known in advance can adapt effectively to a patient's physiology.

References

Amari, S. & Nagaoka, H. (2007). *Methods of Information Geometry*. American Mathematical Soc.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Baum, L. E. & Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, *37*(6), 1554–1563.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*.

Georgatzis, K. & Williams, C. K. I. (2015). Discriminative Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring. *Uncertainty in Artificial Intelligence (UAI)*.

Georgatzis, K., Williams, C. K. I., & Hawthorne, C. (2016). Input-Output Non-Linear Dynamical Systems applied to Physiological Condition Monitoring. *Machine Learning for Healthcare*.

Gepts, E., Camu, F., Cockshott, I., & Douglas, E. (1987). Disposition of Propofol Administered as Constant Rate Intravenous Infusions in Humans. *Anesthesia & Analgesia*, *66*(12), 1256–1263.

Glad, T. & Ljung, L. (2000). *Control Theory*. CRC press.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Lerner, U. & Parr, R. (2001). Inference in Hybrid Networks: Theoretical Limits and Practical Algorithms. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 310–318). Morgan Kaufmann Publishers Inc.

Quinn, J. A., Williams, C. K. I., & McIntosh, N. (2009). Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(9), 1537–1551.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Shumway, R. H. & Stoffer, D. S. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Science & Business Media.

Sorenson, H. W. & Alspach, D. L. (1971). Recursive Bayesian Estimation Using Gaussian Sums. *Automatica*, *7*(4), 465–479.

Stanculescu, I., Williams, C. K. I., & Freer, Y. (2014). Autoregressive Hidden Markov Models for the Early Detection of Neonatal Sepsis. *IEEE journal of biomedical and health informatics*, *18*(5), 1560–1570.

Sutton, C. & McCallum, A. (2011). An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning*, *4*, 267–373.

Chapter 2

The BFGS Algorithm for Nonlinear Optimisation

In the previous chapter, we introduced the IONLDS and the challenge of optimising the nonlinear output function. This chapter contains a practical introduction to performing unconstrained optimisation using the BFGS algorithm. The BFGS algorithm has proved to be one of the most effective general numerical optimisation algorithms, although a stochastic variant has yet to be widely adopted in the large scale context (see for example Keskar and Berahas, 2015). An understanding of the algorithm is useful in its own right, but it is also the subroutine used to optimise the nonlinear output function in the IONLDS. This exposition primarily follows Fletcher [1987], with additions from Nocedal and Wright [2006]. The objective here is to present the main results and intuition; for a more in-depth treatment, the reader is referred to these references.

2.1 Unconstrained Optimisation

We consider the objective of unconstrained optimisation to be:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

where maximisation can be achieved through negating the objective function f. It is assumed that $f \in C^2$ and is bounded below. Define $\mathbf{x}^*$ to be a local minimiser if it satisfies

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) \qquad\qquad \mathbf{x} \in N(\mathbf{x}^*)$$

where $N(\mathbf{x}^*)$ is some neighbourhood around $\mathbf{x}^*$ such as the radius r open ball $B_r(\mathbf{x}^*)$. $\mathbf{x}^*$ is a *strict* local minimiser of f if the inequality holds strictly. We will concentrate on finding local minima, not necessarily global minima which in practice (for nonconvex functions) may be very hard to verify. There are two necessary conditions for $\mathbf{x}^*$ to be a local minima, given $\mathbf{g}(\mathbf{x}) \triangleq \nabla f(\mathbf{x})$ and $G(\mathbf{x}) \triangleq \nabla^2 f(\mathbf{x})$:

$$\mathbf{g}(\mathbf{x}^*) = \mathbf{0} \qquad \text{and} \qquad \mathbf{s}^\mathsf{T} G(\mathbf{x}^*)\mathbf{s} \geq 0 \quad \forall\, \mathbf{s}$$

that is, G is positive semi-definite. These conditions are sufficient for $\mathbf{x}^*$ to be a strict local minimser provided that G is positive *definite.*

2.1.1 The traditional ‘prototype algorithm’

Many methods of unconstrained optimisation can be cast into the same prototype. This includes steepest descent, co-ordinate descent, conjugate gradients and Newton-like algorithms. A user supplies an initial guess $\mathbf{x}^{(0)}$ and the following is iterated until convergence:

(a) determine a search direction $\mathbf{s}^{(k)}$.

(b) minimise $f(\mathbf{x}^{(k)} + \alpha \mathbf{s}^{(k)})$ with respect to α.

(c) set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{s}^{(k)}$ where $\alpha^{(k)}$ is the minimiser of (b).

Part (a) is trivial for both coordinate descent and steepest descent, but due to the lack of sophistication of this choice, both methods typically suffer from oscillation and slow convergence. Newton and Quasi-Newton methods use second order information to inform the search direction, taking into account curvature as well as instantaneous gradient. Convergence is usually defined by either $||f(\mathbf{x}^{(k+1)} - f(\mathbf{x}^{(k)})|| < \epsilon$ or $||\nabla f(\mathbf{x}^{(k)}|| < \epsilon$ for a specified ϵ.

Notation

In order to reduce clutter, the following quantities are defined for use in this chapter:

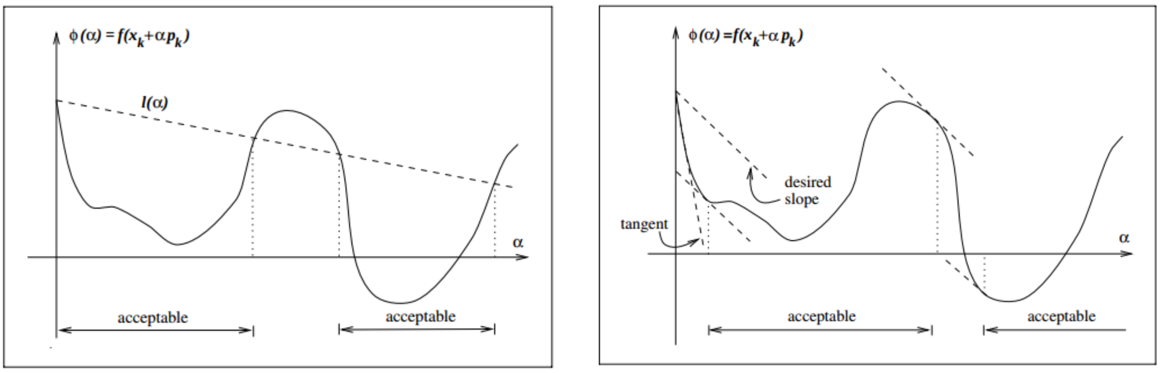
$$\begin{aligned} f^{(\xi)} &= f(\mathbf{x}^{(\xi)}) \\ \mathbf{g}^{(\xi)} &= \mathbf{g}(\mathbf{x}^{(\xi)}) \\ G^{(\xi)} &= G(\mathbf{x}^{(\xi)}) \\ \phi_\xi(\alpha) &= f(\mathbf{x}^{(\xi)} + \alpha \mathbf{s}^{(\xi)}) \end{aligned}$$

2.1.2 Line search

Part (b) of the prototype algorithm is often quite challenging. For some objectives/directions, the one-dimensional minimisation may be performed in closed form but for most interesting cases, a numerical search must be performed. Even though only local minima are sought, wasting valuable resources finding the minimum exactly should be avoided, particularly when the search direction is chosen from a fairly poor model of the function. In order to trade off the competing objectives of fast evaluation and precise minimisation, practical line search algorithms look to minimise inexactly within a region satisfying:

$$\begin{array}{lrcl} \text{(Armijo condition)} & \phi(\alpha) & \leq & \phi(0) + c_1\, \alpha\, \phi'(0) \\ \text{(Curvature condition)} & \phi'(\alpha) & \geq & c_2\, \phi'(0) \end{array}$$

for $0 < c_1 < c_2 < 1$. Collectively these are known as the *Wolfe conditions*. The Armijo condition requires sufficient decrease in the objective to avoid convergence to a non-stationary point. The Curvature condition avoids arbitrarily small steps which are permitted by Armijo. The intuition is that if the gradient is still similar to that at $\alpha = 0$, a better minimum may be found by continuing in the same direction. See figure 2.1.



(a) Armijo Condition

(b) Curvature Condition

Figure 2.1: Wolfe Conditions for Line search. Diagrams 3.3, 3.4 from Nocedal and Wright [2006], copied without permission. The Wolfe conditions are the intersection of the acceptable regions in (a) and (b).

2.2 Newton's Method

Most traditional search methods can be motivated by a truncated Taylor series expansion at the current iterate. First order methods (such as steepest descent) model the objective around the current iterate as a linear function and use this gradient in the prototype algorithm above. Newton and Newton-like methods use a second-order model of the objective,

$$f(\mathbf{x}^{(k)} + \boldsymbol{\delta}) \approx f^{(k)} + \mathbf{g}^{(k)\mathsf{T}}\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^{\mathsf{T}} G^{(k)} \boldsymbol{\delta} \tag{2.1}$$

and use the minimising $\boldsymbol{\delta}$ as the direction. Of course this approximation only has a unique minimiser if $G^{(k)}$ is positive definite. For many practical cases, no such guarantee can be given for every iterate. For strategies to overcome this difficulty, chapter 6 of Nocedal and Wright [2006] provides a useful overview, but we make the assumption here that this situation is not encountered. Differentiating 2.1 with respect to δ we have:

$$\begin{aligned} \nabla f(\mathbf{x}^{(k)} + \boldsymbol{\delta}) &= \mathbf{g}^{(k)} + G^{(k)}\boldsymbol{\delta} \\ \Rightarrow \qquad G^{(k)}\boldsymbol{\delta} &= -\mathbf{g}^{(k)} \end{aligned}$$

at the minimum. The canonical Newton method solves for $\boldsymbol{\delta}$ and (implicitly) uses a step size of 1 in the line search phase. However, this step size is only really justified when optimising a quadratic function, or one that is sufficiently close locally to one. Practical implementations of Newton's method typically also include a line search phase following the prototype algorithm.

Newton's method is sometimes said to have quadratic convergence, at least close to the minimum. This property is in reference to the fact, that sufficiently close to a minimum, the distance to the minimiser $\mathbf{h}^{(k)} \triangleq \mathbf{x}^{(k)} - \mathbf{x}^*$ is bounded by

$$||\mathbf{h}^{(k+1)}|| \leq c||\mathbf{h}^{(k)}||^2.$$

However, to determine if $\mathbf{x}^{(k)}$ is *sufficiently close*, convergence requires that $||\mathbf{x}^{(k)} - \mathbf{x}^*|| < 1/\sqrt{c}$, determined by the distance to the optimum, $||\mathbf{h}^{(k)}||$, which is unknown, and the closeness of the quadratic approximation to the objective function at $\mathbf{x}^{(k)}$ which may be investigated. Therefore, the proven quadratic convergence is of little practical relevance for general functions unless it has a fortuitous start.

2.2.1 Issues with Newton's method

Newton's method is rarely used in practice. Some reasons for this are:

1. **Non positive definiteness of G**. Unless the function is particularly well behaved, we have already noted that $G^{(k)}$ may not always be positive definite. A number of strategies exist for dealing with this behaviour, such as incorporating the steepest descent direction.
2. **Non-descent**. Even if $G^{(k)}$ is positive definite, f may not be reduced by a Newton step; the convergence proof only holds for x close to x^*.
3. **Computational burden**. Even in moderately high dimensional problems, the computational overhead associated with calculating the Hessian and performing a linear solve will dominate. If the method does converge to a local minimum, a quasi-newton or even first order method will often be faster.

If Hessian modification strategies are used to ensure positive definiteness and descent direction, then (1) is no longer a problem. With a guaranteed descent direction obtained this way, (2) can be ensured using the direction in the prototype algorithm in §2.1.1.

2.3 Quasi-Newton Methods

Quasi-Newton methods look similar to Newton methods, but use first order information to *approximate* the Hessian of the objective. In approximating the Hessian, positive definiteness can also be enforced to remove problem (1) in the above. The structure of a Quasi-Newton method adds an extra step to the prototype algorithm. Given an initial estimate of the minimiser, $\mathbf{x}_0$, and an estimate of the inverse Hessian $H^{(0)}$:

1. set $\mathbf{s}^{(k)} = -H^{(k)}\mathbf{g}^{(k)}$.
2. minimise $f(\mathbf{x}^{(k)} + \alpha\mathbf{s}^{(k)})$ with respect to α.
3. set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\mathbf{s}^{(k)}$.
4. Update $H^{(k+1)}$.

$H^{(0)}$ is often taken to be I, but more sophisticated choices are possible. Note that it is the *inverse* Hessian that is updated to avoid the cost of inversion. Of paramount interest is the update step in (4); the other steps are as before. In the interest of developing a criterion for updating H, define the following quantities:

$$\begin{aligned} \boldsymbol{\delta}^{(k)} &= \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \quad = \alpha^{(k)}\mathbf{s}^{(k)}, \\ \boldsymbol{\gamma}^{(k)} &= \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}, \end{aligned}$$

and expanding around the gradient at $x^{(k)}$:

$$\begin{aligned} \underbrace{\mathbf{g}(\mathbf{x}^{(k)} + \boldsymbol{\delta}^{(k)})}_{\mathbf{g}^{(k+1)}} &= \mathbf{g}(\mathbf{x}^{(k)}) + G^{(k)}\boldsymbol{\delta}^{(k)} + O(||\boldsymbol{\delta}^{(k)}||^2) \\ \Rightarrow \qquad \boldsymbol{\gamma}^{(k)} &= G^{(k)}\boldsymbol{\delta}^{(k)} + O(||\boldsymbol{\delta}^{(k)}||^2). \end{aligned}$$

For convenience, we neglect the higher order terms and simply use the quadratic model. Now a reasonable criterion for H is a matrix that satisfies $H^{(k)}\boldsymbol{\gamma}^{(k)} = \boldsymbol{\delta}^{(k)}$, but since $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\delta}^{(k)}$ are only available after the line search in step (b), the following becomes the Quasi-Newton Criterion (QNC):

$$H^{(k+1)}\boldsymbol{\gamma}^{(k)} = \boldsymbol{\delta}^{(k)} \tag{2.2}$$

in this we ensure the one-step behind inverse Hessian is exact for quadratic functions.

2.3.1 (SR1) The rank-one update

The simplest update that satisfies the QNC is the (symmetric) rank one update (SR1):

$$H^{(k+1)} = H^{(k)} + a\mathbf{u}\mathbf{u}^{\mathsf{T}}$$

by the QNC we must have that:

$$H^{(k)}\boldsymbol{\gamma}^{(k)} + a\mathbf{u}\mathbf{u}^{\mathsf{T}}\boldsymbol{\gamma}^{(k)} = \boldsymbol{\delta}^{(k)}.$$

Observing that a and $\mathbf{u}^{\mathsf{T}}\boldsymbol{\gamma}^{(k)}$ are scalars, we choose

$$\begin{aligned} u &= \boldsymbol{\delta}^{(k)} - H^{(k)}\boldsymbol{\gamma}^{(k)}, \\ a\mathbf{u}^{\mathsf{T}}\boldsymbol{\gamma}^{(k)} &= 1. \end{aligned}$$

The rank one update is then:

$$H \leftarrow H + \frac{\mathbf{u}\mathbf{u}^{\mathsf{T}}}{\mathbf{u}^{\mathsf{T}}\boldsymbol{\gamma}} = H + \frac{(\boldsymbol{\delta} - H\boldsymbol{\gamma})(\boldsymbol{\delta} - H\boldsymbol{\gamma})^{\mathsf{T}}}{(\boldsymbol{\delta} - H\boldsymbol{\gamma})^{\mathsf{T}}\boldsymbol{\gamma}} \tag{SR1}$$

As is convention, the iteration number is dropped for clarity. The SR1 is extremely appealing in its simplicity and has been discovered by a number of different authors. The algorithm is important in terms of analysis, but deficient in practice since it does not guarantee that H stays positive definite. While the Curvature condition implies that $\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\gamma} > 0$, this can not be guaranteed for $(\boldsymbol{\delta} - H\boldsymbol{\gamma})^{\mathsf{T}}\boldsymbol{\gamma}$

2.3.2 Rank two updates

The symmetric rank 2 update, also known as the Davidson-Fletcher-Powell (DFP) algorithm can be derived in a very similar way.

$$H^{(k+1)} = H^{(k)} + a\mathbf{u}\mathbf{u}^{\mathsf{T}} + b\mathbf{v}\mathbf{v}^{\mathsf{T}}$$

and by the QNC we must have that:

$$H^{(k)}\gamma^{(k)} + a\mathbf{u}\mathbf{u}^{\mathsf{T}}\gamma^{(k)} + b\mathbf{v}\mathbf{v}^{\mathsf{T}}\boldsymbol{\gamma}^{(k)} = \boldsymbol{\delta}^{(k)}.$$

u and **v** can no longer be chosen uniquely, but an obvious choice is

$$\begin{aligned} \mathbf{u} &= \boldsymbol{\delta}^{(k)}, & \mathbf{v} &= H^{(k)}\boldsymbol{\gamma}^{(k)}, \\ a\mathbf{u}^{\mathsf{T}}\boldsymbol{\gamma}^{(k)} &= 1, & b\mathbf{v}^{\mathsf{T}}\boldsymbol{\gamma}^{(k)} &= -1 \end{aligned}$$

$$\Rightarrow H \leftarrow H + \frac{\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}}{\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\gamma}} - \frac{H\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}H}{\boldsymbol{\gamma}^{\mathsf{T}}H\boldsymbol{\gamma}} \tag{DFP}$$

Unlike the SR1, the DFP algorithm provably retains positive definiteness of the inverse Hessian H. The proof relies critically on satisfaction of the Curvature condition during the line search phase. The DFP also enjoys a superlinear order of convergence for general functions. Hence with the DFP, virtually all of the convergence and computational problems of Newton's method have been alleviated while keeping a competitive order of convergence.

The BFGS Algorithm

The BFGS algorithm is simply the dual of the DFP; the Hessian approximation is iteratively updated rather than the inverse Hessian. Following the same method as above, one can easily derive:

$$B \leftarrow B + \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}}{\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\delta}} - \frac{B\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}B}{\boldsymbol{\delta}^{\mathsf{T}}B\boldsymbol{\delta}}$$

where B denotes the approximation to the Hessian G. However by careful use of matrix inversion identities, one can arrive at an update for the inverse Hessian,

$$H \leftarrow H + \left(1 + \frac{\boldsymbol{\gamma}^{\mathsf{T}}H\boldsymbol{\gamma}}{\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\gamma}}\right)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}}{\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\gamma}} - \frac{\boldsymbol{\delta}\boldsymbol{\gamma}^{\mathsf{T}}H + H\boldsymbol{\gamma}\boldsymbol{\delta}^{\mathsf{T}}}{\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\gamma}}. \tag{BFGS}$$

This is the famous BFGS update (Broyden-Fletcher-Goldfarb-Shanno) for approximating the inverse Hessian, and one can show that it is the solution to the following minimsation problem:

$$\begin{aligned} & \min_{H} \|H - H_k\| \\ \text{subject to} \quad & H \succ 0\,, \quad H\boldsymbol{\delta}^{(k)} = \boldsymbol{\gamma}^{(k)} \end{aligned}$$

where the norm is the weighted Frobenius norm $||A|| = ||W^{1/2}AW^{1/2}||_F$ for any matrix W satisfying the QNC.

The Broyden Class

Naturally BFGS and DFP are not the only rank two updates satisfying the QNC. A family of rank $r \leq 2$ solutions can be found as an affine combination of the DFP and BFGS solutions:

$$H_{\phi}^{(k+1)} = (1-\phi)H_{\text{DFP}}^{(k+1)} + \phi H_{\text{BFGS}}^{(k+1)}$$

which not only includes the BFGS and DFP solutions, but also the SR1 ($\phi = (\delta^{\mathsf{T}}\gamma)/(\delta - H\gamma)^{\mathsf{T}}\gamma$) and Hoshino formula ($\phi = 1/(1 \pm \gamma^{\mathsf{T}}H\gamma/\delta^{\mathsf{T}}\gamma)$). However, to preserve positive definiteness, a sufficient condition is that $\phi \geq 0$. If the class is instead restricted to be a convex combination (that is $\phi \in [0,1]$), this is known as the *restricted Broyden class.* While many results exist claiming superiority of certain members of the class in certain situations, the BFGS update has proved empirically to work very well in almost all situations, and other members are only infrequently used.

2.4 Using BFGS for learning the IONLDS

An iterative scheme is required to optimise the likelihood in the IONLDS. By taking advantage of the Quasi-Newton Criterion, the BFGS algorithm can achieve a superior order of convergence than gradient ascent. Within an Expectation Maximisation (EM) scheme, typically the parameters A, Q, R will still be updated via co-ordinate-like ascent after an E-step. The remaining parameters can be optimised iteratively using BFGS, and require only the derivatives of each parameter to be passed into a generic BFGS wrapper. See Appendix C for the required derivatives.

References

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley & Sons.

Keskar, N. S. & Berahas, A. S. (2015). adaQN: An Adaptive Quasi-Newton Algorithm for Training RNNs. *arXiv preprint arXiv:1511.01169*.

Nocedal, J. & Wright, S. (2006). *Numerical Optimization*. Springer Science & Business Media.

Chapter 3

Ethical Considerations for Collection and Use of Data

In this section we attend to the ethical considerations for data collection and processing. It is intended to be of general applicability to data science projects, not just those specified in chapter 1. The primary ethical issue is that of privacy, and the conflict inherent with data science. It is taken for granted that the reader understands the benefits of medical research. We will consider some of the arguments for and against the use of sensitive patient data, as well as some legal and policy aspects. Computational privacy is a field in and of itself, and technical issues including de-identification and privacy preserving analysis are not in scope for this report, engaging though they may be.

3.1 Privacy and the UK Data Protection Act

The Data Protection Act (DPA) came into force in the UK in 1998 Act of Parliament, UK [1998] to comply in broad terms with the EU's Data Protection Directive (DPD) in 1995. The DPA lists 8 principles in Appendix 1 which must be complied with for anyone processing personal information. The principles are:

1. Personal data shall be processed fairly and lawfully and, in particular, shall not be processed unless:
   - at least one of the conditions in Schedule 2 is met, and
   - in the case of sensitive personal data, at least one of the conditions in Schedule 3 is also met.
2. Personal data shall be obtained only for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.
3. Personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed.
4. Personal data shall be accurate and, where necessary, kept up to date.
5. Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes.
6. Personal data shall be processed in accordance with the rights of data subjects under this Act.
7. Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.
8. Personal data shall not be transferred to a country or territory outside the European Economic Area unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data.

Schedules 2 and 3 will not be reproduced in this document, but a sufficient condition for compliance with principle (1) is for the data subject to have given explicit consent for the processing of their data. The strength of the EU's DPD and the UK's DPA does not solely lie in the statutes of law, but the DPD (and by extension the DPA) was in part a codification of the right to privacy enshrined in the European Convention on Human Rights. Thus the above principles may be seen as a fundamental right. And more pragmatically, according to Lowrance [2012], (p. 3):

> Simply and profoundly: Privacy should be respected because people should be respected. [. . .] attending assiduously to privacy and the relations of confidentiality that serve it is essential to earning the trust that encourages the public to become involved in research

The consequences of a privacy breach are wide-ranging and sometimes pernicious, particularly as it relates to health data. Some examples are given in Mackie and Bradburn [2000] §3:

> Disclosure of such information may result in being arrested for a crime, being denied eligibility for welfare or Medicaid, . . . losing a job or an election, failing to qualify for a mortgage, or having trouble getting into college. Disclosure of a history of alcoholism, mental illness, venereal disease, or illegitimacy can result in embarrassment and loss of reputation. Less directly, research results based on personal data can cause harm by affecting perceptions about a group to which a person belongs.

3.2 Challenges to compliance

If data are collected with consent for use in a single project at a single institution, there are no substantial ethical questions to address. That is, provided data are stored with appropriate security, and if applicable, encrypted. However, it is becoming increasingly challenging to perform analysis in this way. Technology companies are argued in Mills [2016] to violate three of the principles listed above:

1. The data subject has consented to the processing of their data.
2. Personal data shall be obtained only for one or more specified and lawful purposes.
3. Personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed.

Even though most private and public organisations will obtain consent from data subjects, Mills argues that this 'consent' is often fairly meaningless. In part this is due to privacy policies that obfuscate the terms of engagement Pollach [2005], and are anyway not usually read. But it is even more obviously so if the use of their data is continually evolving, and the consent has taken the form of a blank cheque. This is often the case due to the challenges associated with principle 2. If a condition of access to a service is unfettered use of the subject's data, then we must be careful of coercion inherent in obtaining consent, which is particularly egregious in large social media organisations, and healthcare.

The advent of widespread and large-scale data science has shifted the operation of many companies to the collection of as much data as possible, with the purposes and intentions for processing in perpetual development. Even for research, academics will want to reuse data for as yet unknown projects, and obtaining additional consent forms from the data subject will be a time and cost-intensive process. In contravention of principle 3, the typical machine learning workflow consists of obtaining as many dimensions of data as possible first, and asking questions of the data later. We do not wish to make any a priori assumptions about important variables without interrogating the data first. While smaller studies could be done to determine important variables before larger scale data collection, this may preclude beneficial re-use of the data for other projects.

3.3 Compliance is not the whole story

We have looked at some compelling reasons why privacy should not be violated. However, if we are to follow the DPA to the letter, this would require explicit permissions from data subjects every time a new use of their data is required. This is obviously inimical to research and technological progress. While these are not sufficient reasons to ignore individuals' privacy, it does ask the question – particularly for medical research – whether trade-offs might be beneficial? On the policy of requiring explicit informed consent, Mackie and Bradburn [2000] asserts:

> such a policy is focused solely on an individual right and ignores individual responsibilities. Governments must collect personal information to function, and members of society have a civic duty to cooperate.

The 2015 Nuffield report Richards, Anderson, Hinde, and Kaye [2015] (p. xx (Roman)) discusses the contradiction inherent in fulfilling two practically irreconcilable objectives for biomedical research:

- to generate, use and extend access to data (because doing so is expected to advance research and make public services more efficient); and at the same time,
- to protect privacy as this is a similarly strong imperative, and a requirement of human rights law (and the more access is extended the greater the risks of abuse).

The controls over data required for healthcare are even more stringent. On these, Lowrance [2012] (p. 36) adds

> To the extent that they unjustifiably impede research, they impede the realization of the health benefits that might accrue from the research, and their complexity complicates the protection of privacy and confidentiality.

There is a clear moral imperative to uphold the right to privacy of individuals, but this cannot be held as an absolute, regardless of the opportunity cost.

Lowrance [2012] (chap. 6) argues that the strict burden for consent - regarding sustained education, understanding, defined purpose, written signatures etc. was drafted to be applicable to interventionist research on patients. This continual consent process may be challenged when the researchers are simply using already existing data or collecting it in a benign way. This is not to say that consent should be avoided, but that the presentation could be different.

Holm and Madsen [2009] suggest that explicit consent in data usage from clinical trials is largely ignored by patients anyway. In their trial, seven reasons for consent are given, with two relating to the physical trial, and five relating to the use of data. Holm and Madsen assert that a medical professional is unlikely to convey meaningfully the implications of the data related permissions. They present qualitative data that suggests patients fixate on the implications of the medical permissions and pay scant attention to the rest. The consent regarding data privacy ultimately violates the DPA principle of informed consent.

Consent as an obligation

Perhaps in the face of the 'rational, self-interested' man of economics, Harris [2005] claims that people are often motivated to participate in medical research because they perceive they have an obligation. Harris goes on to claim that such an obligation is justified, in part on appeal to basic fairness. If a person does not participate in medical research, accepting medical treatment is 'free-riding' on the contributions of others, and by definition not fair.

John [2009] offers the criticism that 'free-riding' is only a valid argument if it respects the sequential nature of the research; that is, "free-riding does not occur when we enjoy benefits which would exist whatever we do now". However, he develops the argument further, that it is the future *system* which should be regarded as the benefit, whether or not the actor subsequently has reason to use it. Thus since the system is continuously being improved by medical trials, non-willingness to participate in the present still constitutes a 'free-ride'. Of course willingness to participate must be carefully qualified and balanced against competing priorities in a subject's life, but it is difficult to justify non participation in situations which at most trivially inconvenience a subject and possess minimal risk. John advances that the current system of 'no questions asked' denial of consent implicitly favours the opt-out position.

3.4 Reform required for healthcare research to flourish?

The current status quo in data collection and processing is to enshrine the privacy and autonomous rights of individuals. There is much to be commended about this position. However, we have not clarified what privacy is, and whether it is an *absolute* right. There is great difficulty and little consensus in defining privacy; it appears to be the conflation of several different concepts. Daniel Solove notes in his book, *Understanding Privacy* (p. 1),

> Currently, privacy is a sweeping concept, encompassing (among other things) freedom of thought, control over one's body, solitude in one's home, control over personal information, freedom from surveillance, protection of one's reputation, and protection from searches and seizures.

The concepts pertinent to use of healthcare data are

- control over personal information;
- freedom from surveillance;
- protection of one's reputation.

One is necessarily under surveillance in hospital, so the right is evidently not absolute, and post-hoc analysis of data should not intrude heavily on this right. Protection of reputation is related to the trust one has in the hospital and the data controller. Again, perhaps by necessity, the trust has already been given to the hospital, so this right is protected provided the information is not exposed to third parties. Finally, control over use of data is given to the hospital in a *carte blanche* manner, implicitly under the trust that it will not be used for nefarious purposes. What is clear is that the three relevant rights are not absolute and are not uncommonly traded off against other requirements in a person's life. Indeed the UK government has radically violated freedom from surveillance in the so-called *snooper's charter* with surprisingly little resistance from the population (see The Guardian, MacAskill, 2016).

It also cannot be argued that the Data Protection Act is a faithful mapping of the rights inherent in privacy to a set of necessary and sufficient principles. It is a realisation based on the OECD guidelines of 1980 which worked well in the time it came into force, and still works well in most cases today. However, to identify a violation of the DPA as a violation of privacy per se is not necessarily justified. In particular, principles 2 and 3 (specified purposes and minimal data requirements) could be relaxed under appropriate guarantees, such as non-surveillance.

Healthcare data controls in the UK

There are a large number of controls on the use of healthcare data in the UK, and they are fragmented, at times outdated, and lengthy (Introduction AMS [2011]). Such controls on data, while valuable in the original intent, stifle research into health. In general, the UK requires adherence to the DPA, but additional bodies are involved in setting additional guidelines and approving compliance. Lowrance [2012] (p. 65) summarises these as:

> A system of officers called Caldicott Guardians oversee the use of clinical data in NHS operating units. Regulations relating to the use of NHS patient data without consent, issued under the NHS Act and the Health and Social Care Act, are administered by a National Information Governance Board. Research Ethics Committees apply guidance in reviewing project protocols. Rules are imposed as conditions of funding by the Medical Research Council . . . , the ESRC, and charitable organizations such as the Wellcome Trust and Cancer Research UK. Guidance is issued by the General Medical Council. . . , the royal medical colleges. . . , and other authoritative organizations such as the Human Genetics Commission and the Nuffield Council on Bioethics. The regulatory structures and rules in the four countries of the UK resemble each other, but they differ in specifics and procedures, cross-UK research can require considerable negotiation and coordination.

AMS [2011] suggests a major impediment to research is

> . . . obtaining research permissions across *multiple* NHS sites [emphasis added] is inefficient and inconsistent, characterised by NHS Trusts reinterpreting assessments already undertaken by regulators such as the National Research Ethics Service and duplicating checks that could be done once across a study.

Given the obstacles to performing research it is unsurprising that reform is being sought in the UK as well as many other countries including the US, Canada and Australia.

3.5 Conclusions

It is clearly extremely important that patient data is not divulged to a third party, either by design or attack. While the DPA permits a *data controller* to outsource 'data transactions' with a third party (*data processor*), the data controller is still responsible for adherence of the data processor to the DPA. It is nevertheless true that restricting access to patient data is harmful to the public good. In fact both precedent (Lee, Heilig, and White, 2012) and moral imperative exist for citizens to allow use of their data. Certain public bodies and legislation further inhibits the public benefit that may be realised through data science, such as the patchwork medical committees and funding institutions, and the specificity and minimalism requirements of the DPA.

Putting aside recommendations for reform, two key ethical considerations are apparent:

- a perfect obligation of the researcher(s) to respect the privacy of the individual as well as securing the data against attack,
- an imperfect obligation for data subjects to consent to medical research, particularly when there is no additional risk to the subject.

The critical points here are trust in the researcher(s), and the level of security used in storing and processing the data. Trust requires integrity on the part of the researcher(s) and continual engagement and transparency regarding the use of data and the benefits realised. Security is an ongoing concern, and we must be mindful of the risks of storing sensitive information. This certainly does expose the data subjects to greater risk. Appropriate safeguards may be required, such as minimal data requirements within the scope of each project, encryption, query audits. For some datasets and usages, anonymisation can reduce information leakage. See T. Li and Li [2009] for an overview of some of the measures used for attempting to remove identifying features from the dataset such as k-anonymity and ℓ-diversity. McGraw [2013] outlines and discusses the current 'safe-harbor' rules for de-identifying health data in the US. However, anonymisation is gained only by reducing the utility of the data, in analogy to the risk-reward tradeoff of financial investment (T. Li and Li, 2009). Developments in this area may continue to alter the boundaries of the tradeoff, such as differentially private synthetic datasets (for example Charest, 2011).

The motivation from the moral imperative to participate will be challenged by the risks involved in securing the data. The extent to which this can and should do so is not in scope for this report. The Nuffield report (Richards et al., 2015) concludes that the ideal trade-off should be achieved by a priori identifying all actors who stand to lose or gain from the actions and outcomes of the data activity. In discussion with samples of these groups of people, a set of norms may be discovered which are considered 'reasonable', and are publicly meaningful. Then, while policy may be impossible, there may be a framework under which to 'require' sharing of data when it is unquestionably in the common good, and to have various levels of consent required where it is less obvious.

In the case that the risk can be kept to a minimal level, proposing an opt-out scheme rather than an opt-in scheme might be justified. If agreement regarding the moral imperative exists, we should consider acceptance the default behaviour. There will undoubtedly be valid reasons why persons should not be part of such a database, and further practical and ethical issues surrounding sufficient communication must be considered.

References

Act of Parliament, UK. (1998). Data Protection Act (DPA). http://www.legislation.gov.uk/ukpga/1998/29/contents#pt4-l1g36.

AMS. (2011). Academy of Medical Sciences (UK), A new pathway for the regulation and governance of health research. http://www.acmedsci.ac.uk/policy/policy-projects/a-new-pathway-for-the-regulation-and-governance-of-health-research/.

Charest, A.-S. (2011). How Can We Analyze Differentially-Private Synthetic Datasets? *Journal of Privacy and Confidentiality*, *2*(2), 3.

Harris, J. (2005). Scientific research is a moral duty. *Journal of Medical Ethics*, *31*(4), 242–248.

Holm, S. & Madsen, S. (2009). Informed consent in medical research – A procedure stretched beyond breaking point. In O. Corrigan, J. McMillan, K. Liddell, M. Richards, & C. Weijer (Eds.), *The Limits of Consent: A socio-ethical approach to human subject research in medicine* (Chap. 1, pp. 11–24). Oxford University Press.

John, S. (2009). Is there an obligation to participate in medical research? In O. Corrigan, J. McMillan, K. Liddell, M. Richards, & C. Weijer (Eds.), *The Limits of Consent: A socio-ethical approach to human subject research in medicine* (Chap. 7, pp. 115–132). Oxford University Press.

Lee, L. M., Heilig, C. M., & White, A. (2012). Ethical Justification for Conducting Public Health Surveillance Without Patient Consent. *American Journal of Public Health*, *102*(1), 38–44.

Li, T. & Li, N. (2009). On the Tradeoff Between Privacy and Utility in Data Publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 517–526). ACM.

Lowrance, W. W. (2012). *Privacy, Confidentiality, and Health Research*. Cambridge University Press.

MacAskill, E. (2016, November 19). "Extreme surveillance' becomes UK law with barely a whimper'. *The Guardian*. Retrieved from https://www.theguardian.com/world/2016/nov/19/extreme-surveillance-becomes-uk-law-with-barely-a-whimper

Mackie, C., Bradburn, N. et al. (2000). *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. National Academies Press.

McGraw, D. (2013). Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *Journal of the American Medical Informatics Association*, *20*(1), 29–34.

Mills, P. (2016). Ethical Reuse of Data from Health Care: Data, Persons and Interests. In *The Ethics of Biomedical Big Data* (pp. 429–444). Springer.

Pollach, I. (2005). A Typology of Communicative Strategies in Online Privacy Policies: Ethics, Power and Informed Consent. *Journal of Business Ethics*, *62*(3), 221–235.

Richards, M., Anderson, S., Hinde, J., Kaye, J., et al. (2015). The collection, linking and use of data in biomedical research and health care: ethical issues. *London, UK: Nuffield Council on Bioethics.*

Appendix A

Gaussian Identities

Two useful multivariate Gaussian identities are derived below. These identities simplify derivation of the Kalman Filter considerably, and follow the derivations in Bishop, 2006.

A.1 Conditional → joint

Since the Gaussian distribution is defined only by the first and second moments, we may use the law of total expectation to calculate all the moments required from the conditionals.

Let:

$$x \sim \mathcal{N}(m, P), \qquad\qquad y|x \sim \mathcal{N}(Hx + u, R)$$

Then,

$$\begin{aligned}
\mathbb{E}[y] &= \mathbb{E}[\mathbb{E}[y|x]] \;=\; H\,\mathbb{E}[x] + u \;=\; Hm + u, \\[1ex]
\mathrm{Var}[y] &= \mathrm{Var}[\mathbb{E}[y|x]] + \mathbb{E}[\mathrm{Var}[y|x]] \\
&= HPH^\mathsf{T} + R, \\[1ex]
\mathrm{Cov}(x, y) &= \mathbb{E}_{x,y}[xy^\mathsf{T}] - \mathbb{E}[x]\,\mathbb{E}[y]^\mathsf{T} \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}[xy^\mathsf{T}|x]] - \mathbb{E}[x]\,\mathbb{E}[y]^\mathsf{T} \\
&= \mathbb{E}_x[x(Hx+u)^\mathsf{T}] - m(Hm+u)^\mathsf{T} \\
&= \Big[(\mathrm{Var}(x) + mm^\mathsf{T})H^\mathsf{T} + mu^\mathsf{T}\Big] - mm^\mathsf{T}H^\mathsf{T} - mu^\mathsf{T} \\
&= PH^\mathsf{T}.
\end{aligned}$$

And so,

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m \\ Hm + u \end{pmatrix}, \begin{pmatrix} P & PH^\mathsf{T} \\ HP & HPH^\mathsf{T} + R \end{pmatrix} \right). \tag{A.1}$$

A.2 Joint → conditional

Here we will need an identity linking a block matrix to its inverse:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix} \tag{A.2}$$

where

$$M = (A - BC^{-1}D)^{-1}, \quad \text{(the Schur Complement)}$$

We follow Bishop, 2006's derivation. Given a joint distribution specified by mean μ and precision (inverse covariance) Λ:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix}^{-1}\right)$$

we seek an expression for the conditioned distributions $p(x|y)$ and $p(y|x)$. By expressing the Gaussian in terms of the precision matrix, it allows the exponent of the density function to be written down immediately:

$$\begin{aligned} &-\frac{1}{2}(x-\mu_x)^\intercal \Lambda_{xx}(x-\mu_x) - \frac{1}{2}(x-\mu_x)^\intercal \Lambda_{xy}(y-\mu_y) \\ &-\frac{1}{2}(y-\mu_y)^\intercal \Lambda_{yx}(x-\mu_x) - \frac{1}{2}(y-\mu_y)^\intercal \Lambda_{yy}(y-\mu_y) \end{aligned}$$

Collecting terms which depend on x we have:

$$\begin{aligned} &= -\frac{1}{2}x^\intercal \Lambda_{xx} x + x^\intercal \Lambda_{xx}\mu_x - x^\intercal \Lambda_{xy}(y-\mu_y) + \text{const} \\ &= -\frac{1}{2}x^\intercal \Lambda_{xx} x + x^\intercal \left(\Lambda_{xx}\mu_x - \Lambda_{xy}(y-\mu_y)\right) + \text{const} \end{aligned}$$

From comparison with the general quadratic form of a single (possibly vector-valued) random variable, we can read off that $\Lambda_{x|y} = \Lambda_{xx}$. (Remember that $\Lambda_{xx}^{-1} \neq \Sigma_{xx}$ in general).

$$\begin{aligned} \Rightarrow \Sigma_{x|y} &= \Lambda_{xx}^{-1} \\ &= \left(\left(\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)^{-1}\right)^{-1} \\ &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \end{aligned}$$

where the second line follows from (A.2). This is the Schur Complement of the covariance matrix.

Next the mean vector $\mu_{x|y}$ can be read off by comparison to the general quadratic form:

$$\begin{aligned} \Rightarrow \Sigma_{x|y}^{-1}\mu_{x|y} &= \Lambda_{xx}\mu_x - \Lambda_{xy}(y-\mu_y) \\ &= \Lambda_{xx}\mu_x + \Lambda_{xx}\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y) \\ \Rightarrow \mu_{x|y} &= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y) \end{aligned}$$

with the second line again following from (A.2). We can derive equations for $p(y|x)$ in an entirely analogous manner.

Therefore,

$$\begin{aligned} x|y &\sim \mathcal{N}\left(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right), \\ y|x &\sim \mathcal{N}\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right) \end{aligned} \tag{A.3}$$

Equations (A.1) and (A.3) will be used extensively for the derivations in appendix B.

Appendix B

Inference in Linear Dynamical Systems

B.1 Filtering

The filtering problem is to define the marginal posterior of a latent state given all the preceding and current observations.

$$\begin{aligned}
\alpha_t(\mathbf{x}_t) \quad &\triangleq \quad p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \\
&= \frac{\int \mathrm{d}\mathbf{x}_{t-1} \; p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{y}_t|\mathbf{y}_{1:t})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \\
&= \frac{\int \mathrm{d}\mathbf{x}_{t-1} \; p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \\
&= \frac{p(\mathbf{y}_t|\mathbf{x}_t) \int \mathrm{d}\mathbf{x}_{t-1} \; p(\mathbf{x}_t|\mathbf{x}_{t-1})\alpha_{t-1}(\mathbf{x}_{t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \qquad \text{(B.1)}
\end{aligned}$$

We follow the notation used in Särkkä, 2013. Let $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^n$ with the following distributions:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(A_{t-1}\mathbf{x}_{t-1}, Q_{t-1}), \qquad \text{(B.2a)}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(H_t\mathbf{x}_t, R_t), \qquad \text{(B.2b)}$$

$$p(\mathbf{x}_1|\varnothing) = \mathcal{N}(\mathbf{m}_1^-, P_1^-) \qquad \text{(B.2c)}$$

and we assume for the purposes of induction the following distribution is known:

$$p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{m}_{t-1}, P_{t-1}). \qquad \text{(B.2d)}$$

We build up the solution of (B.1) for Linear Dynamical Systems as follows:

1. Using Gaussian identity (A.1) with above equations (B.2d) and (B.2a), we have the one-step ahead prediction:

$$\begin{aligned}
p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \;&\triangleq\; \mathcal{N}(\mathbf{m}_t^-, P_t^-) \\
&= \mathcal{N}(A_{t-1}\mathbf{m}_{t-1}, \;\; A_{t-1}P_{t-1}A_{t-1}^\mathsf{T} + Q_{t-1}). \qquad \text{(B.3)}
\end{aligned}$$

2. Again, using (A.1) to construct the joint with the next emission from $p(\mathbf{y}_t|\mathbf{x}_t)$:

$$p\left(\begin{pmatrix}\mathbf{x}_t \\ \mathbf{y}_t\end{pmatrix} \middle| \; \mathbf{y}_{1:t-1}\right) = \mathcal{N}\left(\begin{pmatrix}\mathbf{m}_t^- \\ H_t\mathbf{m}_t^-\end{pmatrix}, \; \begin{pmatrix}P_t^- & P_t^- H_t^\mathsf{T} \\ H_tP_t^- & H_tP_t^-H_t^\mathsf{T} + R_t\end{pmatrix}\right). \qquad \text{(B.4)}$$

3. Finally, we condition on $\mathbf{y}_t$ or equivalently, normalise by $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ using (A.3):

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \triangleq \mathcal{N}(\mathbf{m}_t, P_t) \tag{B.5a}$$

where

$$\begin{aligned} \mathbf{m}_t &= \mathbf{m}_t^- + P_t^- H_t^\mathsf{T}(H_t P_t^- H_t^\mathsf{T} + R_t)^{-1}(\mathbf{y}_t - H_t\mathbf{m}_t^-), && \text{(B.5b)} \\ P_t &= P_t^- - P_t^- H_t^\mathsf{T}(H_t P_t^- H_t^\mathsf{T} + R_t)^{-1} H_t P_r^-. && \text{(B.5c)} \end{aligned}$$

For the purposes of clarity, (B.3) - (B.5) are written in algorithmic form below, without a time dependence on the parameters A, H, Q, R:

Predictive Mean and Variance

$$\begin{aligned} &\text{(predictive mean)} && \mathbf{m}_t^- = A\mathbf{m}_{t-1} \\ &\text{(predictive variance)} && P_t^- = AP_{t-1}A^\mathsf{T} + Q \end{aligned}$$

Filtered Mean and Variance

$$\begin{aligned} &\text{(marginal variance of } \mathbf{y}_t\text{)} && S_t = HP_t^- H^\mathsf{T} + R \\ &\text{(Kalman gain)} && K_t = P_t^- H^\mathsf{T} S_t^{-1} \\ &\text{(filtered mean)} && \mathbf{m}_t = \mathbf{m}_t^- + K_t(\mathbf{y}_t - H\mathbf{m}_t^-) \\ &\text{(filtered variance)} && P_t = P_t^- - K_t S_t K_t^\mathsf{T} \end{aligned}$$

One can readily observe that in one dimension, the quantity known as the “Kalman gain” matrix is the proportion of the marginal variance of $\mathbf{y}$ ‘explained’ by the covariance of $\mathbf{x}$. Thus intuitively, the Kalman gain determines the amount of $\mathbf{y}_t$ to update the hidden state $\mathbf{x}_t$, and when the covariance $\mathrm{Cov}(\mathbf{x}, \mathbf{y})$ is large compared to the variance of $\mathbf{y}$, we place greater weight on the observation $\mathbf{y}_t$.

The entire procedure forms a prediction based on the state dynamics, and regressing the new observation $\mathbf{y}_t$ onto this model. The Kalman gain matrix is effectively the (simple) regression coefficient (of covariance over variance). Since the distributions are linear-Gaussian with assumed known and constant variance, the posterior variance is independent of the measurement data. This might seem surprising since one might imagine that highly variable measurement data might indicate higher posterior variance than that which is clustered together. This is in general not true for non-linear or non-Gaussian models (Sorenson and Alspach, 1971).

B.2 Smoothing

The smoothing procedure is defined mathematically as:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{y}_{1:T}) &= \int \mathrm{d}\mathbf{x}_{t+1}\;\; p(\mathbf{x}_t, \mathbf{x}_{t+1}|\mathbf{y}_{1:T}) \\ &= \int \mathrm{d}\mathbf{x}_{t+1}\;\; p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) \\ &= \int \mathrm{d}\mathbf{x}_{t+1}\;\; p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) \\ &= \int \mathrm{d}\mathbf{x}_{t+1}\;\; \frac{p(\mathbf{x}_t, \mathbf{x}_{t+1}|\mathbf{y}_{1:t}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} \\ &= \int \mathrm{d}\mathbf{x}_{t+1}\;\; \frac{p(\mathbf{x}_{t+1}|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{y}_{1:t}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} \end{aligned} \tag{B.6}$$

Conditional independence relationships in the latent Markov model have been exploited to simplify these quantities. In (B.6), the numerator consists of the evolution/transition density, the filtered density to time t, and a recursion. Thus we can evaluate this quantity in a backward recursive manner similar to HMMs. Given the distributions described in the filtering step:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(A_{t-1}\mathbf{x}_{t-1}, Q_{t-1}), \tag{B.7a}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(H_t\mathbf{x}_t, R_t), \tag{B.7b}$$

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{m}_t, P_t) \tag{B.7c}$$

and we assume for the purposes of induction the following distribution is known:

$$p(x_{t+1}|y_{1:T}) = \mathcal{N}(\mathbf{m}^s_{t+1}, P^s_{t+1}). \tag{B.7d}$$

Now,

1. Form the joint predictive posterior, using (B.7c), (B.7a) with identity (A.1):

$$p\left(\left.\begin{pmatrix}\mathbf{x}_t \\ \mathbf{x}_{t+1}\end{pmatrix}\right| y_{1:t}\right) = \mathcal{N}\left(\begin{pmatrix}\mathbf{m}_t \\ A_t\mathbf{m}_t\end{pmatrix}, \begin{pmatrix}P_t & P_tA_t^\mathsf{T} \\ A_tP_t & A_tP_tA_t^\mathsf{T} + Q_t\end{pmatrix}\right). \tag{B.8}$$

2. Use identity (A.3) to condition on the future state $\mathbf{x}_{t+1}$:

$$p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}) \triangleq \mathcal{N}(\tilde{\mathbf{m}}_t, \tilde{P}_t) \tag{B.9a}$$

where

$$\tilde{\mathbf{m}}_t = \mathbf{m}_t + P_tA_t^\mathsf{T}(A_tP_tA_t^\mathsf{T} + Q_t)^{-1}(\mathbf{x}_{t+1} - A_t\mathbf{m}_t), \tag{B.9b}$$

$$\tilde{P}_t = P_t - P_tA_t^\mathsf{T}(A_tP_tA_t^\mathsf{T} + Q_t)^{-1}A_tP_t. \tag{B.9c}$$

3. Using the fact that $p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}) \equiv p(\mathbf{x}_t|x_{t+1}, \mathbf{y}_{1:T})$, we incorporate the previous recursion information from (B.7d). Since only the marginal of $\mathbf{x}_t|\mathbf{y}_{1:T}$ is required, we use the result from the law of total expectation/covariance directly:

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) \triangleq \mathcal{N}(\mathbf{m}^s_t, P^s_t) \tag{B.10a}$$

where

$$\begin{aligned}
\mathbf{m}^s_t &= \mathbb{E}_{\mathbf{x}_{t+1}|\mathbf{y}_{1:T}}[\tilde{\mathbf{m}}_t] \\
&= \mathbf{m}_t + P_tA_t^\mathsf{T}(A_tP_tA_t^\mathsf{T} + Q_t)^{-1}(\mathbf{m}^s_{t+1} - A_t\mathbf{m}_t), && \text{(B.10b)} \\
P^s_t &= \mathbb{E}_{\mathbf{x}_{t+1}|\mathbf{y}_{1:T}}[\tilde{P}_t] + \mathrm{Var}_{\mathbf{x}_{t+1}|\mathbf{y}_{1:T}}[\tilde{\mathbf{m}}_t] \\
&= \tilde{P}_t + P_tA_t^\mathsf{T}(A_tP_tA_t^\mathsf{T} + Q_t)^{-1}P^s_{t+1}(A_tP_tA_t^\mathsf{T} + Q_t)^{-1}A_tP_t. && \text{(B.10c)}
\end{aligned}$$

While the complexity here is limited, the number of terms is a little intimidating, and we collect them again in algorithmic form:

Backward Rauch-Tung-Striebel Updates

(forward predicted mean)	$\mathbf{m}^-_{t+1} = A\mathbf{m}_t$
(forward predicted variance)	$P^-_{t+1} = AP_tA^\mathsf{T} + Q$
(gain matrix)	$G_t = P_tA^\mathsf{T}(P^-_{t+1})^{-1}$
(adjust smoothed mean)	$\mathbf{m}^s_t = \mathbf{m}_t + G_t(\mathbf{m}^s_{t+1} - \mathbf{m}^-_{t+1})$
(adjust smoothed variance)	$P^s_t = P_t + G_t(P^s_{t+1} - P^-_{t+1})G_t^\mathsf{T}$

following the exposition in Särkkä, 2013.

It is the position of the conditioning step that determines the structure of these updates. Thus the smoother regresses on the smoothed latent state rather than the observations. Since this is earlier in the update, the additional terms in the variance occur only in the final step and can be written as above. Also because the variance in the emission distribution is not considered during the conditioning, the previous (backward) smoothed estimate can have a larger impact on the recursion.

Appendix C

Sigmoid Learning using BFGS

C.1 Learning the Sigmoid Parameters in the IONLDS

In the IONLDS model (Georgatzis et al., 2016), the model is learned using Expectation Maximisation (EM), and in the M-step, the sigmoid function is optimised using the BFGS algorithm. Since the posterior is derived using the Unscented Transform, we assume existence of the sigma points, $\{\boldsymbol{\chi}_i^t\}_{i=0}^{2n}$ and covariance weights $\{W_i^{(c)}\}_{i=1}^{2n}$ for each smoothed distribution $P(\mathbf{x}_t|\mathbf{y}_{1:T})$. See Kokkala, Solin, and Särkkä [2014] for an exploration of parameter learning for Sigma-Point models. In the interests of clarity, the control inputs have been omitted, but in practice they contribute only 5 extra constant terms in the transition distribution trace.

Define:

$$
\begin{aligned}
Q(\Theta;\Theta^{(\text{old})}) &= \left\langle \log p(\mathbf{x},\mathbf{y}) \right\rangle_{q(\mathbf{x})} \\
&= \log p(\mathbf{x}_1) - \frac{1}{2}\left[\sum_{t=2}^{T}\left\langle \log|2\pi Q| + \operatorname{Tr}\left(Q^{-1}(\mathbf{x}_t - A\mathbf{x}_{t-1})(\mathbf{x}_t - A\mathbf{x}_{t-1})^{\mathsf{T}}\right)\right\rangle_{q(\mathbf{x})}\right] \\
&\quad - \frac{1}{2}\left[\sum_{t=1}^{T}\left\langle \log|2\pi R| + \operatorname{Tr}\left(R^{-1}(\mathbf{y}_t - h(\mathbf{x}_t))(\mathbf{y}_t - h(\mathbf{x}_t))^{\mathsf{T}}\right)\right\rangle_{q(\mathbf{x})}\right] \\
&= \log p(x_1) - \frac{T-1}{2}\log|2\pi Q| - \frac{T}{2}\log|2\pi R| - \frac{T-1}{2}\operatorname{Tr}\left(Q^{-1}(\Sigma - AC^{\mathsf{T}} - CA^{\mathsf{T}} + A\Phi A^{\mathsf{T}})\right) \\
&\quad - \frac{1}{2}\operatorname{Tr}\left(R^{-1}\left(\sum_{t=1}^{T}\sum_{i=0}^{2n} W_i^{(\text{c})}(\mathbf{y}_t - h(\boldsymbol{\chi}_i^t))(\mathbf{y}_t - h(\boldsymbol{\chi}_i^t))^{\mathsf{T}}\right)\right).
\end{aligned}
$$

Here, the upper case letters in the evolution distribution trace are the statistics defined in Kokkala et al., 2014 or Särkkä, 2013 (ch.12). The distribution $q(\mathbf{x})$ is chosen to be the (approximate) posterior distribution $p(\mathbf{x}|\mathbf{y})$ as usual for the EM algorithm.

C.1.1 Derivatives of $Q(\Theta;\Theta^{(\text{old})})$

Parameters A, Q and R can be optimised with relative ease, but parameters in the sigmoid function cannot be optimised analytically, and we perform iterative optimisation instead. The derivatives of $Q(\Theta;\Theta^{(\text{old})})$ can be calculated by the chain rule: first consider the partial derivatives of the generalised logistic function:

$$h(\mathbf{x}) = m + \frac{M-m}{(1+e^{-\gamma C\mathbf{x}})^{1/\nu}}.$$

Clearly this is over-parameterised, and the γ can be absorbed into the C. However, we wish to retain the same model as the original. Some trivial transformations of the derivatives are performed for overflow/underflow reasons.

- $\boldsymbol{m}$ (Minimum of function):
$$\frac{\partial h}{\partial m} = 1 - \frac{1}{(1+e^{-\gamma C\mathbf{x}})^{1/\nu}}.$$
- $\boldsymbol{M}$ (Maximum of function):
$$\frac{\partial h}{\partial M} = \frac{1}{(1+e^{-\gamma C\mathbf{x}})^{1/\nu}}.$$
- $\boldsymbol{\gamma}$ (slope):
$$\begin{aligned}\frac{\partial h}{\partial \gamma} &= (M-m)\frac{-1}{(1+e^{-\gamma C\mathbf{x}})^{2/\nu}}\left(\frac{1}{\nu}(1+e^{-\gamma C\mathbf{x}})^{1/\nu-1}\right)\left(-C\mathbf{x}e^{-\gamma C\mathbf{x}}\right)\\ &= \frac{(M-m)C\mathbf{x}}{\nu(1+\exp\{-\gamma C\mathbf{x}\})^{\frac{1+\nu}{\nu}}}\exp\{-\gamma C\mathbf{x}\}\\ &= \frac{(M-m)C\mathbf{x}}{\nu(1+\exp\{\gamma C\mathbf{x}\})(1+\exp\{-\gamma C\mathbf{x}\})^{\frac{1}{\nu}}}.\end{aligned}$$
- $\boldsymbol{C}$ (latent transformation):
$$\begin{aligned}\frac{\partial h}{\partial C^{\mathsf{T}}} &= (M-m)\frac{-1}{(1+e^{-\gamma C\mathbf{x}})^{2/\nu}}\left(\frac{1}{\nu}(1+e^{-\gamma C\mathbf{x}})^{1/\nu-1}\right)\frac{\partial}{\partial C}\exp\{-\gamma C\mathbf{x}\}\\ &= \frac{(M-m)\gamma}{\nu(1+\exp\{-\gamma C\mathbf{x}\})^{\frac{1+\nu}{\nu}}}\exp\{-\gamma C\mathbf{x}\}\mathbf{x}^{\mathsf{T}}.\end{aligned}$$
- $\boldsymbol{\nu}$ (symmetry of asymptotes): Here we use the following identity:
$$\frac{\partial}{\partial x}z^{-\frac{1}{x}} = (\log z)z^{-\frac{1}{x}}\frac{\partial}{\partial x}(-\frac{1}{x}) = \left(\frac{\log z}{x^2}\right)z^{-\frac{1}{x}}$$
$$\frac{\partial h}{\partial \nu} = \frac{(M-m)\log(1+e^{-\gamma C\mathbf{x}})}{\nu^2(1+e^{-\gamma C\mathbf{x}})^{1/\nu}}.$$

In the IONLDS, each output channel is given its own set of parameters, and so C is a row vector. If C is a matrix, all multiplications and divisions of terms involving $C\mathbf{x}$ should be interpreted as element-wise operations.

Derivative of the expected log joint

The derivative of $Q(\Theta;\Theta^{(\text{old})}$ can now be calculated. With respect to the parameters we are optimising, we have that:

$$Q(\Theta;\Theta^{(\text{old})}) = -\frac{1}{2}\operatorname{Tr}\left(R^{-1}\left(\sum_{t=1}^{T}\sum_{i=0}^{2n}W_i^{(\text{c})}(\mathbf{y}_t - h(\boldsymbol{\chi}_i^t))(\mathbf{y}_t - h(\boldsymbol{\chi}_i^t))^{\mathsf{T}}\right)\right) + \text{const},$$
$$\frac{\partial Q}{\partial \psi} = R^{-1}\left(\sum_{t=1}^{T}\sum_{i=0}^{2n}W_i^{(\text{c})}\frac{\partial h(\boldsymbol{\chi}_i^t)}{\partial \psi}(y - h(\boldsymbol{\chi}_i^t))^{\mathsf{T}}\right),$$

using the partial derivatives of h above.

C.2 Modifications to the IONLDS

There are a couple of issues with fitting IONLDS models. Firstly the calculation of the derivatives encounter substantial issues from finite precision arithmetic. Secondly, very poor local optima may be found from BFGS, and good initialisation strategies must be used.

Numerical issues and expressiveness of sigmoid

Evaluating the gradient of the emission distribution as described in the previous section is prone to numerical imprecision. The most problematic term is $(1 + e^{-\gamma C\mathbf{x}})^{1/\nu}$, where both the exponent $-\gamma C\mathbf{x}$ and ν are $O(100)$. The optimisation tends to find fairly large values of ν due to the lack of a bias term in the generalised sigmoid. The position of the value of z such that $h(z) = m + 0.5(M - m)$ can be altered by changing ν, but ν primarily governs the symmetry in the convergence of the upper and lower asymptotes. A more natural way to translate the sigmoid function is the following:

$$h_2(\mathbf{z}) = m + \frac{M - m}{(1 + e^{-C\mathbf{z}-b})^{1/\nu}}$$

for some bias term b. This alteration may be absorbed into the existing framework by prepending a fixed bias (1) to the state x_t when performing the optimisation, and expanding H accordingly. The parameter ν should be considered for removal from the nonlinearity, since the asymptote asymmetry is a relatively unimportant degree of freedom when compared with finding sensible values for A, C, M, m, b. The numerical stability of the function can then be assured.

Experiments were conducted on the 40 patients to test the performance of removing ν (setting $\nu = 1$), and adding in the bias. Qualitatively, the results indicate that overfitting is easier: this broadly makes sense since the bias increases the modelling capacity of a sigmoid function much more than changing the symmetry of the asymptotes.

Sensitivity to initialisation

Currently the dynamical system parameters are currently initialised using spectral methods as applied to a *linear* dynamical system model, and additional optimisation of the LDS using the EM algorithm. Precisely which parameters were chosen seems to be primarily of an experimental nature than a theoretical one. The parameters of the sigmoid function are chosen from a heuristic initialisiation and a greedy optimisation as well as some of the parameters derived from the PKPD model.

This tactic seems to work well some of the time, however there are some notable failures to find useful optima. It is an undesirable aspect of a model that it fits the data only “on a good day”. We must find better initialisation heuristics which encompass a knowledge of how the model is working. Techniques from global optimisation such as simulated annealing or Bayesian optimisation might also be required to find better values for the sigmoid initialisation.