Robust spatio-temporal latent variable models

Jacqueline Christmas

January 2011
Abstract

Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) are widely-used mathematical models for decomposing multivariate data. They capture spatial relationships between variables, but ignore any temporal relationships that might exist between observations. Probabilistic PCA (PPCA) and Probabilistic CCA (ProbCCA) are versions of these two models that explain the statistical properties of the observed variables as linear mixtures of an alternative, hypothetical set of hidden, or latent, variables and explicitly model noise. Both the noise and the latent variables are assumed to be Gaussian distributed.

This thesis introduces two new models, named PPCA-AR and ProbCCA-AR, that augment PPCA and ProbCCA respectively with autoregressive processes over the latent variables to additionally capture temporal relationships between the observations. To make PPCA-AR and ProbCCA-AR robust to outliers and able to model leptokurtic data, the Gaussian assumptions are replaced with infinite scale mixtures of Gaussians, using the Student-t distribution.

Bayesian inference calculates posterior probability distributions for each of the parameter variables, from which we obtain a measure of confidence in the inference. It avoids the pitfalls associated with the maximum likelihood method: integrating over all possible values of the parameter variables guards against overfitting. For these new models the integrals required for exact Bayesian inference are intractable; instead a method of approximation, the variational Bayesian approach, is used. This enables the use of automatic relevance determination to estimate the model orders.

PPCA-AR and ProbCCA-AR can be viewed as linear dynamical systems, so the forward-backward algorithm, also known as the Baum-Welch algorithm, is used as an efficient method for inferring the posterior distributions of the latent variables. The exact algorithm is tractable because Gaussian assumptions are made regarding the distribution of the latent variables. This thesis introduces a variational Bayesian forward-backward algorithm based on Student-t assumptions.

The new models are demonstrated on synthetic datasets and on real remote sensing and EEG data.
I would very much like to thank four people without whom this work would not have been possible:

**Dr Richard Everson**
for being my supervisor and mentor for the last three years, for inspiring my interest in the subjects of pattern recognition and machine learning, and for doing all this with considerable patience, humour and charm.

**John Hamilton**
for initiating my interest in mathematics, introducing me to computers, and teaching me “proper” maths. *Plus et en vous* indeed.

My parents, **Peter** and **Sissel Christmas**, for their unconditional love and support, for instilling in me an appreciation of learning, and for providing me with the best education possible.

I would also like to thank **Dr Sonia Gallegos** of NASA’s Stennis Space Center for providing the SeaWiFS satellite images and **Dr Aureliu Lavric** of the School of Psychology, University of Exeter, for providing the EEG data that are used throughout this thesis.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of tables</td>
<td>8</td>
</tr>
<tr>
<td>List of figures</td>
<td>9</td>
</tr>
<tr>
<td>Nomenclature and Abbreviations</td>
<td>17</td>
</tr>
<tr>
<td>Publications</td>
<td>18</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>19</td>
</tr>
<tr>
<td>1.1 Real data examples</td>
<td>21</td>
</tr>
<tr>
<td>ONS data</td>
<td>21</td>
</tr>
<tr>
<td>NASA data</td>
<td>22</td>
</tr>
<tr>
<td>EEG data</td>
<td>23</td>
</tr>
<tr>
<td>1.2 Summary of new models</td>
<td>24</td>
</tr>
<tr>
<td>1.3 Aims</td>
<td>25</td>
</tr>
<tr>
<td>1.4 Thesis structure</td>
<td>25</td>
</tr>
<tr>
<td>1.5 Novel contributions</td>
<td>26</td>
</tr>
<tr>
<td>2 Background</td>
<td>27</td>
</tr>
<tr>
<td>2.1 Bayesian probability theory</td>
<td>27</td>
</tr>
<tr>
<td>2.1.1 Bayesian inference</td>
<td>29</td>
</tr>
<tr>
<td>2.1.2 Bayesian approximation techniques</td>
<td>31</td>
</tr>
<tr>
<td>2.1.3 Variational Bayesian approximation</td>
<td>32</td>
</tr>
<tr>
<td>2.2 Latent Variable Models</td>
<td>34</td>
</tr>
<tr>
<td>2.2.1 Principal Component Analysis</td>
<td>36</td>
</tr>
<tr>
<td>2.2.2 Factor Analysis</td>
<td>37</td>
</tr>
<tr>
<td>2.2.3 Probabilistic PCA</td>
<td>38</td>
</tr>
<tr>
<td>2.2.4 Canonical Correlation Analysis</td>
<td>40</td>
</tr>
<tr>
<td>2.2.5 Probabilistic CCA</td>
<td>41</td>
</tr>
<tr>
<td>2.3 Model order selection</td>
<td>41</td>
</tr>
<tr>
<td>2.4 Robustness</td>
<td>44</td>
</tr>
<tr>
<td>2.5 Assessing model quality</td>
<td>47</td>
</tr>
<tr>
<td>2.5.1 Ambiguous transformations</td>
<td>48</td>
</tr>
<tr>
<td>2.5.2 Missing data estimation for PCA-like models</td>
<td>49</td>
</tr>
<tr>
<td>2.5.3 Error measures</td>
<td>51</td>
</tr>
<tr>
<td>2.6 Conclusions</td>
<td>51</td>
</tr>
<tr>
<td>3 Maximum likelihood PPCA with autoregressive latent variables</td>
<td>55</td>
</tr>
</tbody>
</table>
3.1 Autoregression ........................................... 56
3.2 PPCA-AR: a linear dynamical system ................. 57
   Forward-backward algorithm .......................... 58
   EM algorithm: E-step ................................. 59
   EM algorithm: M-step ................................. 60
   Imputing missing values .............................. 62
3.3 Illustration: synthetic data ............................. 64
   3.3.1 Illustration 1: estimating noise precisions ....... 64
   3.3.2 Illustration 2: estimating \( W \) and \( \theta \) ........ 68
   3.3.3 Illustration 3: imputing missing values .......... 68
3.4 Results: real data .................................... 70
3.5 Conclusions ............................................ 71

4 Robust variational PPCA ................................. 74
4.1 Gaussian variational PPCA ............................ 75
4.2 Robust variational PPCA ............................. 75
4.3 New robust variational PPCA ....................... 78
   4.3.1 Priors ............................................ 79
   4.3.2 Factorised variational Bayesian method ........ 80
      Mixing matrix, \( W \) .................................. 81
      Latent variables, \( x_n \) ............................... 82
      ARD parameters, \( \delta \) ............................. 82
      Latent variables, \( z_n \) ............................... 83
      Degrees of freedom, \( d \) ........................... 84
      Latent variables, \( u \) ................................ 84
      Observation noise degrees of freedom, \( v \) ......... 85
      Observation noise precision, \( \lambda \) ............... 85
      Observation mean, \( \bar{t} \) ........................... 86
   4.4 Summary ............................................. 86
4.5 Convergence criterion ................................. 89
4.6 Illustration: synthetic data ............................ 89
   4.6.1 Illustration 1: parameter estimation ............. 90
   4.6.2 Illustration 2: model order estimation ........... 95
   4.6.3 Illustration 3: robustness to outliers ............ 96
4.7 Robust variational FA ................................. 99
4.8 Conclusions ............................................ 100

5 Robust variational autoregression ..................... 102
5.1 Robust variational AR ................................. 102
   5.1.1 Priors ............................................ 103
   5.1.2 Factorised variational Bayesian method .......... 104
      AR coefficients, \( \theta \) ............................... 105
      Excitation noise precision, \( \kappa \) ................. 105
      Latent variables, \( z \) ................................ 106
      Degrees of freedom, \( d \) ........................... 106
List of Tables

Summary of the maximum likelihood PPCA-AR EM algorithm. . . . . . . . . . . 63

Summary of robust, variational PPCA. . . . . . . . . . . . . . . . . . . . . . . . . 88

4.1 Observation noise degrees of freedom $v$ calculated for each of the outlier
test datasets for (top) the model with Student-t observation and latent
variables and (bottom) the model with Student-t observation noise and
Gaussian latent variables. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98

Summary of robust, variational AR. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108

Summary of robust, variational PPCA-AR. . . . . . . . . . . . . . . . . . . . . . 126

Summary of robust, variational ProbCCA-AR. . . . . . . . . . . . . . . . . . . . . 161

7.1 Table showing which of the 8 latent variables, and their corresponding com-
ponents, contributes to each of the 4 observation sets. . . . . . . . . . . . . . . . . . . . . . 177
List of Figures

1.1 Plots showing a Gaussian probability distribution in blue and a heavier-tailed Student-t distribution in red. .............................................. 21
1.2 ONS data: output of 23 UK production industries. .................. 22
1.3 NASA data: example SeaWiFS image. ...................................... 23
1.4 EEG data: example EEG trace and its Normal probability plot. ...... 24

2.1 The class of latent variable models extended or generalised in this thesis. . . 36
2.2 Examples of underfitting and overfitting of a model to the data. .......... 42
2.3 Demonstration of the effect on the maximum likelihood solution for a Gaussian distribution of a small number of outliers. ................. 44
2.4 Graph showing effects of the precision, λ, and degrees of freedom, d, on the shape of the Student-t distribution. The black curve, when d is very large, corresponds to a Gaussian distribution. ....................................... 46
2.5 Demonstration of the effect on the maximum likelihood solution for Gaussian and Student-t distributions of a small number of outliers. .. 46
2.6 Hinton diagrams showing examples of Procrustes transformation matrices. 48
2.7 A summary of models described in this thesis. ............................. 54

3.1 A graphical representation of the PPCA-AR model. ...................... 58
3.2 A graphical representation of the forward-backward algorithm. .......... 59
3.3 Comparison of estimated and actual observation noise precisions. ........ 65
3.4 Comparison of estimated and actual excitation noise precisions. ........ 66
3.5 Comparison of PPCA-AR and PPCA mean squared error of data recon- struction. ................................................................. 66
3.6 Comparison of PPCA-AR mean squared error of data reconstruction with the actual observation noise variance. .............................. 67
3.7 Comparison of estimated and actual AR coefficients. .................... 67
3.8 Hinton diagrams comparing the actual and estimated W matrices. ...... 68
3.9 PPCA-AR: Comparisons of the mean squared error of reconstruction for synthetic data. ......................................................... 69
3.10 PPCA-AR: Comparisons of the mean squared error of reconstruction for the ONS data. ....................................................... 70
3.11 ONS data: comparison of the observation noise precision λ averaged over 30 different sets of missing masks. The dashed lines represent one standard deviation above and below the mean. ............................. 71
3.12 ONS data: comparisons of the reconstructions of missing data using PPCA and PPCA-AR. ....................................................... 72
3.13 ONS data: comparisons of the variances of the reconstructions of missing data using PPCA and PPCA-AR. ........................................... 72

4.1 Graph showing contours of the joint probability density for two independent random variables, $x_1$ and $x_2$, sampled from (a) Gaussian and (b) Student-t distributions. ................................................................. 75

4.2 Graphical model for Bishop’s [1999b] Gaussian variational PPCA. The $\mathbf{m}$ and $\mathbf{V}$ are the mean and covariance for a Gaussian distribution; while $a$ and $b$ are the two parameters for a Gamma density. ............................ 76

4.3 Graphical model for Luttinen et al.’s [2009] and Gai et al.’s [2008] robust, variational PPCA. ............................................................... 77

4.4 Graphical model for the new robust, variational PPCA model. ....... 78

4.5 Boxplot of estimated vs. actual observation noise precision, $\lambda$. The red boxes show the mean and one standard deviation; the blue lines show the full range of estimated values across the 600 synthetic datasets. ................. 90

4.6 Plot of estimated vs. actual observation noise degrees of freedom, $v$. Each point is coloured according to the actual observation noise precision and a small amount of lateral scatter has been introduced for visualisation. The heavy black lines indicate equality between estimates and actuals. ......... 91

4.7 Plot of estimated vs. actual latent variables’ degrees of freedom, $d_q$. Each point is coloured according to the actual observation noise precision and a small amount of lateral scatter has been introduced for visualisation. The heavy black lines indicate equality between estimates and actuals. For each test dataset the $d_q$ have been sorted such that (a) shows the $d_q$ associated with the column of $\mathbf{W}$ that has the greatest magnitude while (d) shows the $d_q$ associated with the column of $\mathbf{W}$ that has the smallest magnitude. Hence (d) displays the $d_q$ associated with the component that is most likely to have been switched off by ARD. .................................................. 92

4.8 Plot of estimated vs. actual mean observation values, $\bar{\mathbf{t}}$. The black line indicates equality. Each point is coloured according to the actual observation noise precision. ................................................................. 93

4.9 Hinton diagrams comparing the estimated and actual mixing matrix, $\mathbf{W}$, where the estimate is good. ......................................................... 94

4.10 Hinton diagrams comparing the estimated and actual mixing matrix, $\mathbf{W}$, where the estimate is bad. ......................................................... 94

4.11 Two different views of the ARD parameters for each of the four components in the synthetic data. In both cases the ARD parameters for each dataset have been sorted by increasing value. In (a) the mean and one standard deviation are shown in red and the full range of values in blue. In (b) each point is coloured according to the actual observation noise precision. ....... 95

4.12 The estimated $\mathbf{W}$ calculated by the two different ARD models. In (a) the true model order of 8 has been found, while in (b) two additional components are switched on. ......................................................... 96
4.13 The estimated $\delta_q$ calculated by the two different ARD models as the iterative process progresses. Large values indicate components that have been switched off by ARD. Note the logarithmic ordinate scale. In (a) the values converge smoothly, while in (b) the model has not converged and has taken a long time to switch off unwarranted components. .......................... 96

4.14 For the datasets with no outliers and the datasets containing each of the four types of outliers, are plotted the mean and one standard deviation of the absolute difference between the estimated and actual values for (a) the mean observation $\bar{t}$ and (b) the observation noise precision $\lambda$. The most robust models are those with a mean value close to zero and a small standard deviation. Note that the most robust model appears to be that with Gaussian observation noise and Student-t latent variables. .................. 97

4.15 The model order for the outlier tests as calculated by robust variational PPCA. The actual model order is marked as a black dotted line. .......... 98

5.1 Graphical model showing the AR model’s parameter variables and their interdependencies. .................. 104

5.2 Gaussian and Student-t autoregression examples. ............. 109

5.3 Comparison of estimated and actual autoregression coefficients and excitation noise precisions. .......................... 110

5.4 Comparison of estimated and actual model order. ............. 111

5.5 Plot of autoregression coefficients for 100 synthetic datasets. ........ 111

5.6 Demonstration of the effect of outliers on Gaussian and Student-t models. . 112

5.7 Comparison of the 80% credibility intervals for the reconstructions of EEG data using AR models with Student-t and Gaussian excitations. ........ 113

5.8 Comparison of AR coefficients for all 58 EEG channels for Student-t and Gaussian excitations. .......................... 114

6.1 Graphical model for the robust variational PPCA-AR model. Those parameters circled in blue have been inherited from the robust variational PPCA model; those circled in red have been inherited from the robust variational AR model. Note that $\hat{\mathbf{x}}_1$ has different dependencies from those shown. . . 119

6.2 Comparison of estimated and actual observation noise precisions, $\lambda$. ..... 130

6.3 Comparison of estimated and actual observation noise degrees of freedom, $v$. 130

6.4 Comparison of estimated and actual latent variables degrees of freedom, $d_q$. 131

6.5 Comparison of estimated and actual mean observation values, $\bar{t}$. ......... 131

6.6 Hinton diagrams comparing the estimated and actual mixing matrix, $\mathbf{W}$, where the estimate is good. .......................... 132

6.7 Hinton diagrams comparing the estimated and actual mixing matrix, $\mathbf{W}$, where the estimate is bad. .......................... 132

6.8 Heat maps for determining the quality of the model’s estimates of the mixing matrix, $\mathbf{W}$. .......................... 134

6.9 Comparison of estimated and actual AR coefficients ($\theta_q$). .......................... 135
List of Figures

6.10 Expectations \pm one standard deviations from the approximate posterior distributions for (a) $W$, (b) $\bar{t}$, (c) $d$, and (d) $\theta_q$ estimated for, on the left, a synthetic dataset with high precision ($\lambda = 100$) and, on the right, a synthetic dataset with lower precision ($\lambda = 1$). In each case the expectation is shown as a red line, with a blue box indicating one standard deviation above and below. .................................................. 136

6.11 Estimation of model order: components in $W$. ........................................ 137

6.12 Box plots of the relative magnitudes of the AR coefficients over all the tests, separated into separate plots for each actual model order. Each diagram shows box plots for $\theta_{q,i}$ for the actual value (in black) and the three different values of $\lambda$ for each model order, from $P = 0$ to $P = 8$. In each case it may be seen that the magnitude of the estimates is significantly greater for coefficients where $i \leq P$, indicating that it is indeed detecting the correct model orders. ................................................................. 139

6.13 Box plots of the standard deviations of the $\theta_{q,i}$ associated with the plots shown in figure 6.12.......................................................... 140

6.14 Imputation of missing data by PPCA-AR, comparing the maximum likelihood model with the robust variational model. .................................................. 141

6.15 Imputation of missing data by PPCA-AR, comparing the maximum likelihood model with the robust variational model. .................................................. 141

6.16 Hinton diagram of the $W$ estimated for the ONS data by robust variational PPCA-AR. .......................................................... 142

6.17 Plot of the $\theta_q$ estimated for the ONS data by robust variational PPCA-AR. 143

6.18 Plots of the degrees of freedom estimated for the ONS data by robust variational PPCA-AR by iteration for (a) the observation noise and (b) the excitation noise. These show that convergence has not yet been reached, but the large values indicate that both noise elements are likely to be approximately Gaussian. .................................................. 143

6.19 Two examples of plausible estimates for the missing data in NASA SeaWiFS images. On the lower image the estimation on the right hand side of the picture seems to be better than that on the left. ...................................... 145

6.20 Two examples of implausible estimates for the missing data in NASA SeaWiFS images. .......................................................... 146

6.21 Number of observed instances for each pixel in the 92 NASA SeaWiFS images. The blue pixels represent approximately 84% missing values, while the red pixels represent approximately 56% missing values. .................................................. 147

6.22 Hinton diagram of the $W$ estimated for the EEG data. Note that components 31, 34, 36, 41-48, 50, 53, 55-57 and 60 have been switched off. . . . . . . 147

6.23 Plots of the three most significant latent variables for the EEG data. . . . . . 148

6.24 Surface plot of the $\theta_q$ for the 44 latent variables that have not been switched off. A value of $P = 20$ was used to train the model. .................................................. 148
7.1 Graphical model for the robust variational ProbCCA-AR model with shared ARD parameters over each $W^{(*)}$. Those parameters circled in red have identical dependencies to the same parameters in the robust variational PPCA-AR model. Note that the first vector of latent variables, $x_1$, has different dependencies from those shown.

7.2 Hinton diagrams of (row 1) actual, (row 2) estimated and (row 3) transformed estimated $W$ and (row 4) the Procrustes transformation matrix for (left) low, (centre) medium and (right) high observation noise. The excitation noise has a heavy-tailed distribution ($d_q = 2$).

7.3 Hinton diagrams of (row 1) actual, (row 2) estimated and (row 3) transformed estimated $W$ and (row 4) the Procrustes transformation matrix for (left) low, (centre) medium and (right) high observation noise. The excitation noise has an approximately Gaussian distribution ($d_q = 100$).

7.4 Plots showing the mean (red) and one standard deviation (blue) of the estimated $\theta_q$ values superimposed on the actual values (black) used to generate the data for (left) low, (centre) medium and (right) high observation noise and excitation noise with (top) a heavy-tailed distribution ($d_q = 2$) and (bottom) an approximately Gaussian distribution ($d_q = 100$).

7.5 Plots of actual latent variable values (in blue) against scaled estimated values (in red, with one standard deviation shown in green) for (a) low, (b) medium and (c) high observation noise and excitation noise with a heavy-tailed distribution ($d_q = 2$).

7.6 Plots of actual latent variable values (in blue) against scaled estimated values (in red, with one standard deviation shown in green) for (a) low, (b) medium and (c) high observation noise and excitation noise with an approximately Gaussian distribution ($d_q = 100$).

7.7 Plots of the median (red) and quartiles (blue) for the components of $W$ estimated by the model where the actual model order is 2.

7.8 Plots of the median (red) and quartiles (blue) for (a) the actual AR coefficients used to generate the data and (b) the estimated values of the elements of $\theta_q$ where the actual model order is 5.

7.9 The $W$ estimated by the model. While ARD has not entirely switched off any components, clearly the second component is more significant than the others.

7.10 The median (in red) and quartile values (in blue) for the components in the estimated $W$, highlighting the significance of the second component.

7.11 The significant component split into the subvectors associated with each observation set and plotted separately. There is a similarity in the shapes of these plots, with higher contributions from the latent signal detected around electrodes 43 and 44 and lower contributions at electrodes 47, 48 and 40.

7.12 The mean (red) and one standard deviation (blue) of the elements in (a) the first component and (b) the significant second component in $W$. 
7.13 The latent signal associated with the significant component. The black dotted line marks the time step at which the trigger for the evoked response was made. ................................................................. 174

7.14 Surface plot of the AR coefficients, \( \theta_q \), associated with each of the 9 latent variables. As for PPCA-AR in chapter 6 the set is remarkably uniform. . . 174

7.15 Heatmap of the observation noise precision matrix, \( B \), estimated by the model. The black lines outline each of the five individual \( B^{(\ast)} \) (for the five sets of observations) that make up the block-diagonal matrix. . . . . . . . . . . . . 174

7.16 Subset of the graphical model for the robust variational ProbCCA-AR model with different ARD parameters over each \( W^{(\ast)} \). The elements in red indicate the change from the previous model. All elements not included are identical to those shown in figure 7.1. . . . . . . . . . . . . . . . . . . . 176

7.17 (a) The actual \( W \) used to generate the synthetic data. (b) The \( W \) estimated by classical ProbCCA-AR. (c) The \( W \) estimated by hybrid ProbCCA-AR. (d) The \( W \) estimated by hybrid ProbCCA-AR, transformed in the direction of the actual \( W \) using the Procrustes transformation. . . . . . . . . . . . . . . . . . . 177

7.18 The transformation matrix calculated by using the Procrustes method to transform the \( W \) estimated by hybrid ProbCCA-AR in the direction of the actual \( W \). It represents differences in sign and permutation of the components, but there is negligible rotation ambiguity. . . . . . . . . . . . 178
Nomenclature and Abbreviations

Acronyms and abbreviations
AR  Autoregression.
CCA  Canonical Correlation Analysis.
EM  Expectation Maximisation.
FA  Factor Analysis.
MAP Maximum A Posteriori.
PCA  Principal Component Analysis.
PPCA  Probabilistic Principal Component Analysis.
PPCA-AR  PPCA with independent AR processes applied to the latent variables.
ProbCCA  Probabilistic Canonical Correlation Analysis.
ProbCCA-AR  ProbCCA with independent AR processes applied to the latent variables.

Matrix/vector/scalar notation
X  A matrix.
y  A column vector.
z  A scalar.
x_j  A column vector; the jth column of matrix X; sometimes written as x_{-,j}.
There is one exception to this rule: the precision matrix B in chapter 7 retains its capitalisation to avoid confusion with other variables.
x_{i,:}  A row vector; the ith row of matrix X (see note regarding matrix B above).
x_{i,j}  A scalar, the element in the ith row, jth column of matrix X (see note regarding matrix B above).

Probability density functions
\( \mathcal{G}(\cdot) \)  Gamma.
\( \mathcal{N}(\cdot) \)  Gaussian or Normal.
\( \mathcal{S}(\cdot) \)  Student-t.
\( \mathcal{W}(\cdot) \)  Wishart.

Greek
\( \alpha, \beta \)  The parameters for a posterior Gamma distribution, \( \mathcal{G}(\alpha, \beta) \).
\( \gamma \)  The ARD prior parameter for the AR coefficients in \( \theta \).
\( \Gamma(\cdot) \)  The gamma function.
\( \Delta \)  The posterior scale matrix of the Wishart distribution for the ProbCCA-
AR precision matrix $\mathbf{B}$.

$\delta$ \hspace{1cm} The ARD prior parameter for the mixing matrix in $\mathbf{W}$.

$\epsilon_n$ \hspace{1cm} The excitation noise associated with the $n$th vector of latent variables.

$\Theta$ \hspace{1cm} The block diagonal matrix made up of the individual $\theta_q$ vectors of AR coefficients (defined in (3.7)).

$\theta$ \hspace{1cm} The AR coefficients for one variable.

$\kappa$ \hspace{1cm} The excitation noise precision.

$\lambda$ \hspace{1cm} The vector of observation noise precisions for FA.

$\lambda$ \hspace{1cm} The observation noise precision for PPCA.

$\mu$ \hspace{1cm} The posterior mean of a Gaussian distribution.

$\rho$ \hspace{1cm} An eigenvalue.

$\Sigma$ \hspace{1cm} The posterior variance of a multivariate Gaussian distribution.

$\sigma^2$ \hspace{1cm} The posterior variance of a univariate Gaussian distribution.

$\Phi$ \hspace{1cm} The block diagonal matrix containing the precisions of the excitation noise for each of the latent variables (defined in (3.8)).

$\chi$ \hspace{1cm} The posterior degrees of freedom of the Wishart distribution for the ProbCCA-AR precision matrix $\mathbf{B}$.

$\psi(\cdot)$ \hspace{1cm} The digamma function.

$\Omega$ \hspace{1cm} The set of all parameters for a given model.

**Roman**

$a, b$ \hspace{1cm} The parameters for a prior Gamma distribution, $\mathcal{G}(a, b)$.

$\mathbf{B}$ \hspace{1cm} Observation noise precision matrix for ProbCCA-AR.

$C$ \hspace{1cm} The number of observation sets in ProbCCA-AR.

$D$ \hspace{1cm} The dimensionality of the observations.

$d$ \hspace{1cm} The degrees of freedom for the excitation noise Student-t distribution.

$e_n$ \hspace{1cm} The observation noise associated with the $n$th observation.

$\mathbf{F}$ \hspace{1cm} A fixed matrix of ones and zeros that extracts a vector latent variables, $\mathbf{x}_n$, from its stacked equivalent, $\tilde{\mathbf{x}}_n$ (see (3.5)).

$g_n$ \hspace{1cm} Where data is missing, those elements of an observation, $t_n$, that are observed.

$h_n$ \hspace{1cm} Where data is missing, those elements of an observation, $t_n$, that are missing.

$M$ \hspace{1cm} The matrix which records which values in the observation matrix $\mathbf{T}$ are missing.

$N$ \hspace{1cm} The number of observations.

$n$ \hspace{1cm} As a subscript: associated with the $n$th observation.

$P$ \hspace{1cm} The number of AR coefficients in $\theta_q$.

$Q$ \hspace{1cm} The dimensionality of the latent variables.

$q$ \hspace{1cm} As a subscript: associated with the $q$th latent variable.

$\mathbf{R}$ \hspace{1cm} An arbitrary rotation matrix.
As a subscript: associated with the $r$th dimension of the observations.

$S$  
The sample covariance matrix.

$T$  
The $D \times N$ matrix whose columns are the ordered observation vectors, $t_n$.

$t_n$  
The $D$-dimensional, $n$th observation vector.

$\bar{t}$  
The $D$-dimensional mean observation.

$u_n$  
The $D$-dimensional vector of latent variables associated with the Student-t distribution for observation noise in the $n$th observation.

$V$  
The variance of a Gaussian prior distribution.

$v$  
The degrees of freedom for the observation noise Student-t distribution.

$W$  
The $D \times Q$ mixing matrix for PPCA, PPCA-AR and ProbCCA-AR.

$X$  
The $Q \times N$ matrix whose columns are the ordered latent variable vectors, $x_n$.

$x_n$  
The $Q$-dimensional vector of latent variables associated with the $n$th observation.

$Y$  
A fixed matrix of ones and zeros used in the maximum likelihood PPCA-AR model as part of the calculation to estimate the AR coefficients, $\theta_q$ (see (3.38)).

$Z$  
A fixed matrix of ones and zeros used in the maximum likelihood PPCA-AR model as part of the calculation to estimate the AR coefficients, $\theta_q$ (see (3.38)).

$z_n$  
The $Q$-dimensional vector of latent variables associated with the Student-t distribution for excitation noise in the $n$th observation.
Publications

Some of the material in chapter 3 has been published in


Some of the material in chapter 5 has been published in

1. Introduction

One way to analyse or learn from a set of data is to construct a mathematical model which somehow represents its essential structure or characteristics, in the hope that this will improve our understanding and perhaps point to some generative system. Many sets of data that we wish to analyse represent a set of observations measured at uniform intervals over a period of time, with each observation consisting of the measurements of a number of different variables. A model for such multivariate time-series data would ideally capture both spatial relationships between variables and temporal relationships between observations.

This thesis describes new models which incorporate temporal structure into well-known spatial models. In fact although they are described in terms of time series, the relationship between observations does not have to be time-related; these models could equally be applied to any set of data where the ordering of the observations is important and the interval between them is uniform; for example, weather measurements recorded at a specific time at one mile intervals along a straight line between London and Edinburgh.

The acquisition of real data introduces different types of errors into the observations. These may be due to imperfections in the measuring devices or processes, or because the data are being measured indirectly and the relationship between the measurement and the underlying data is not well understood, or because there are underlying processes that we are unaware of or not interested in modelling. The new models are designed to accommodate three common types of errors:

- **noise**
  random errors which are hopefully small in size in relation to the actual values of interest

- **outliers**
  unrepresentative values which fall significantly outside the general distribution of the main body of the data

- **missing values**
  values which failed to be recorded, perhaps due to equipment failure or some obstacle moving between the measuring device and the subject of the measurement (for example, clouds obscuring the ground from Earth-sensing satellites)

The aim of this thesis is to create new mathematical models of real datasets that capture the essential information within them while not being distracted by noise or outliers, either to predict the values of the data that are missing (or perhaps future values), or to identify
meaningful structure that might enable us to learn something about the way the data were generated.

There are a large number of standard models, of which variants of four are described in this thesis: Principal Component Analysis (PCA) [Pearson, 1901], Factor Analysis (FA) [Spearman, 1904], Canonical Correlation Analysis (CCA) [Hotelling, 1936] and autoregression (AR) [e.g. Chatfield, 2004]. The first three have been widely used to capture spatial relationships between variables, but they ignore any temporal relationships that might exist between the observations. AR is commonly used to model time-series data.

This thesis describes new models, named PPCA-AR and ProbCCA-AR, which incorporate AR processes into PCA (and by the way FA) and CCA to enable them to model both spatial and temporal relationships between multivariate time-series data. The new models which result from these adaptations are probabilistic in nature, that is they explicitly model noise and lead to determinations of probability distributions for the observations which, in the Bayesian approach, provide measures of uncertainty in the models’ predictions. Standard PCA and CCA are not statistical models, but statistical versions of them exist in Probabilistic PCA (PPCA) [Roweis, 1998; Tipping and Bishop, 1999] and Probabilistic CCA (ProbCCA) [Bach and Jordan, 2005].

PPCA, FA and ProbCCA are examples of linear latent variable models; they explain the statistical properties of the observed variables as linear mixtures of an alternative, hypothetical set of hidden, or latent, variables. The latent variables are usually of lower dimensionality than the observations so that they represent underlying, shared signals. PPCA-AR and ProbCCA-AR view these latent variables as independent AR processes, resulting in models that may be viewed as linear dynamical systems. As such we may make use of the forward-backward algorithm [Rabiner, 1989], also known as the Baum-Welch algorithm [Baum, 1972], as an efficient method for inferring the posterior distributions of the latent variables.

Each of the new models has associated with it a concept of model order, which is a measure of its complexity. The same model may be applied to data using different levels of complexity, giving different results. PPCA-AR and ProbCCA-AR use automatic relevance determination (ARD) [Mackay, 1994; Neal, 1995] to estimate the “correct” model order, a concept that will be described more fully in chapter 2.

The new models incorporate features which enable them to accommodate outliers without affording them undue influence. Many models, including PPCA and ProbCCA, assume that the data and/or noise are distributed according to a Gaussian (Normal) probability distribution. Gaussian models are known to be particularly sensitive to outliers and many real datasets are distributed with tails that are significantly heavier than a Gaussian (that is, they are leptokurtic). The new models replace the Gaussian assumptions with the heavier-tailed Student-t (see figure 1.1) both to accommodate outliers and to model leptokurtic data.

Bayesian inference [Bayes, 1763; Bernardo and Smith, 1994; Denison et al., 2002], which is the basis for PPCA-AR and ProbCCA-AR, depends on integrals which turn out to be
1. Introduction

Gaussian Student−t Gaussian Student−t

Figure 1.1. Plots showing a Gaussian probability distribution in blue and a heavier-tailed Student-t distribution in red.

intractable. Instead, a method of approximation, the variational Bayesian approach, is used. As a result, a new variational Bayesian version of the forward-backward algorithm is introduced to infer the posterior distributions of the (non-Gaussian) latent variables. Through the development of the PPCA-AR model, new variational versions of PPCA and AR are described which use the Student-t distribution to make them robust to outliers and ARD to estimate the model orders and .

The following section describes three real datasets that are used throughout this thesis to illustrate the new models. This is followed by summaries of the aims of the thesis and then the contents of each chapter. Section 1.5 highlights those elements of the thesis that represent novel contributions in this field.

1.1. Real data examples

Figures 1.2 to 1.4 show examples of real, multivariate, time-series data that are used throughout this thesis to demonstrate the new models. The three datasets illustrate the effects of outliers, leptokurtosis and missing values.

ONS data

The UK’s Office of National Statistics (ONS) publishes economic and socio-economic time series data relating to the UK. Figure 1.2 shows the time series release “Detailed Index of Production”, table “A1: Output of the production industries : seasonally adjusted” [Office for National Statistics, 2009], which contains monthly production figures from January 1968 to February 2009 (as at 9th April 2009) for 24 different industries. The 22nd dimension, series “CKZO: IOP: CA_1: Extraction of oil and gas: CVMSA”, has been omitted as figures are not available before January 1976. These figures are index values, with 100 units representing the average monthly figure in 2003 in each case. Looking at the coal industry plot, for example, it may be seen that this is a time series in that measurements taken closely together in time are generally closely related to one another. The
1. Introduction

![UK Industrial Production Output](image1.png)

Figure 1.2. Output of 23 UK production industries recorded monthly from January 1968 to February 2009. Values are indexes, with the average 2003 figure represented as 100 units for each industry.

plot is noisy, rather than smooth, and if the aim is to model the underlying trends then the miners’ strikes of 1972, 1974 and 1984-5 may be regarded as outliers (alternatively, if the aim is to exactly model the output then some sort of switching state model might be more appropriate; these observations are then no longer outliers but switches between the “on strike” and “not on strike” states).

NASA data

The main mission of NASA’s Sea-viewing Wide Field-of-view Sensor (SeaWiFS) platform is to measure the concentration of phytoplankton, and hence chlorophyll, in the surface of the oceans [NASA, 2010b]. This is important to the understanding of the Earth’s carbon cycle. The chlorophyll in plankton absorb certain wavelengths of sunlight causing the light reflected back to SeaWiFS to be reduced at those wavelengths. One of those wavelengths, which is close to the maximum absorption wavelength for chlorophyll, occurs at 443nm and can therefore be used to measure plankton density. Figure 1.3 shows one of the SeaWiFS 443nm images supplied by NASA. There is one image per day for the last 92 days in 2004, containing 113,593 sea pixels in a 300 × 600 picture. Each shows the same stretch of coastline in the northern region of the Gulf of Mexico, around the city of New Orleans. The colours represent different levels of phytoplankton in the surface of the sea; white is land and grey represents areas which are obscured by cloud, i.e. missing data. In the coloured regions there are occasional single pixels that represent values which are significantly higher than the surrounding pixels, i.e. outliers. Single outlier pixels are probably a result of objects in the sea such as ships; the speckled areas around the clouds are probably faint clouds which have not been successfully identified by the cloud identification algorithm.
1. Introduction

Figure 1.3. Example SeaWiFS image showing the northern Gulf of Mexico. White areas are land, colour represents different concentrations of phytoplankton in the sea and grey represents areas where the phytoplankton measurements are missing due to cloud cover.

**EEG data**

Electroencephalography (EEG) is the biomedical technology and science of recording the tiny electrical currents generated by the brain. These recording are made by fixing a number of electrodes to standard positions on the scalp which then pick up aggregations of signals generated by millions of neurons acting together to produce electrical patterns or “brain waves”. Figure 1.4 shows the signal recorded by one electrode and the shows the Normal probability plot (a plot of the sample quantiles of the data versus theoretical quantiles from a normal distribution) for the same signal, showing that it is significantly non-Gaussian in nature (demonstrated by its variation from the black line); with a kurtosis of 3.8 (and negligible skewness) it is leptokurtic\(^1\), i.e. heavier-tailed than Gaussian.

Advances in the use of computers to analyse EEG signals have moved from the general identification of, for example, delta waves (sinusoidal waves in the range 0.5 to 4 Hz which signify deep sleep) to the discovery and measurement of specific event-related potentials or evoked responses. The complexity and minute size of these evoked responses, and their low signal to noise ratio makes it difficult to determine their origin and the pattern of their spread through the brain over time. One method of doing this is to evoke the same response multiple times within one subject, repeat this across multiple test subjects and then use some sort of aggregation or averaging to compare the results with similar tests made for a different evoked response in the hope of finding differences.

\(^{1}\text{A Gaussian distributed variable has a kurtosis of 3.}\)
1. Introduction

Figure 1.4. The plot on the left is a single EEG trace. Its leptokurticity is demonstrated by the Normal probability plot shown on the right, which plots the sample quantiles of the data versus theoretical quantiles from a Gaussian distribution.

1.2. Summary of new models

The new models described in this thesis are PPCA-AR and ProbCCA-AR, which incorporate AR processes into the spatial PPCA and ProbCCA models respectively in order to add temporal modelling. The basic structure of PPCA-AR is first described as a Gaussian maximum likelihood model. Bayesian inference is introduced through new variational PPCA and AR models in which Gaussian assumptions are replaced with the Student-t (for robustness to outliers and to model leptokurtic data) and ARD is used to estimate the model orders. These are then combined into a variational, Student-t version of PPCA-AR, with ARD, which makes use of a new variational forward-backward algorithm. Finally this model is extended to create two different versions of ProbCCA-AR.

FA is a model that is closely related to PPCA. Although it is not the subject of this thesis, straightforward changes to the new variational PPCA and PPCA-AR models provide similar models based on FA. These are described for completeness, but no experimental results are included.
1. Introduction

1.3. Aims

The aims of the work in this thesis are as follows:

1. To incorporate temporal structure into the existing Probabilistic PCA and CCA models using autoregression.

2. To use Bayesian inference to estimate probability distributions for all the model parameters. Exact inference is intractable, so instead variational approximation is used.

3. To make the new models robust to outliers and able to model leptokurtic data.

4. To estimate the model orders for the new models automatically by incorporating automatic relevance determination.

1.4. Thesis structure

The remainder of the thesis is structured as follows:

Chapter 2
Starts with a brief overview of Bayesian probability theory and describes variational Bayesian approximation. Describes the existing models on which this work is based, i.e. (P)PCA, FA and (Prob)CCA. Introduces the concept of model complexity and automatic relevance determination. Demonstrates Gaussian susceptibility to outliers and introduces the Student-t as a robust alternative. Finally, describes some methods involved in assessing models for quality.

Chapter 3
Introduces a new Gaussian model, PPCA-AR, which models each latent variable of PPCA as independent AR processes, and describes an EM algorithm (which incorporates the forward-backward algorithm) for estimating the maximum likelihood model parameters.

Some material in this chapter has been published in [Christmas and Everson, 2010].

Chapter 4
Extends existing robust variational versions of PPCA [Bishop, 1999b; Gai et al., 2008; Luttinen et al., 2009] by changing the way that (i) ARD is used to estimate the model order, thus speeding up convergence, and (ii) the Student-t distribution is used to model the different noise contributions.

Chapter 5
Introduces a new variational version of AR which models the noise using the Student-t distribution, for robustness and to model leptokurtic data, and uses ARD to estimate the model order.

Some material in this chapter has been published in [Christmas and Everson, 2011].
Chapter 6
Extends the PPCA-AR model introduced in chapter 3 in two ways. Firstly, ARD is used to automatically estimate the model orders of the latent AR processes. Secondly, the model is made robust by modelling the AR excitation sequence and observation noise by Student-t distributions. Variational Bayes is used for parameter inference and a new robust, variational version of the forward-backward algorithm.

Chapter 7
Introduces a new model, ProbCCA-AR, which applies independent AR processes to each latent variable of Canonical Correlation Analysis, and describes a variational algorithm for estimating the model parameters. The new model incorporates ARD for model order estimation, the Student-t distribution for robustness, and the robust, variational forward-backward algorithm. A second ProbCCA-AR model is described which represents a hybrid between PPCA-AR and ProbCCA-AR.

Chapter 8
Summarises the models and results described in the thesis and draws conclusions.

Appendix A (page 187) summarises the main properties of the standard probability distributions used in this thesis. Appendix B describes maximum likelihood fitting of Gaussian and Student-t distributions to univariate data; this is used in section 2.4 to illustrate the effect of outliers on these distributions.

1.5. Novel contributions

The following contributions are, to the best of my knowledge, novel to this thesis:

- New latent variable models, PPCA-AR and ProbCCA-AR, which incorporate autoregressive processes over the latent variables in PPCA and ProbCCA in order to capture temporal information in the data. PPCA-AR is described as both a maximum likelihood model (chapter 3) and a robust, variational model (chapter 6); two different robust, variational ProbCCA-AR models are described (chapter 7).

- A new robust, variational PPCA model (chapter 4), extending the work of Bishop [1999b]; Gai et al. [2008]; Luttinen et al. [2009].

- A new robust, variational version of autoregression (chapter 5).

- A new robust, variational version of the forward-backward algorithm (chapter 6, section 6.1.1).
2. Background

This chapter starts by summarising Bayesian probability theory as it applies to the modelling of data such as those described in the introduction. This is followed, in section 2.2, by a survey of the particular models used in this thesis. Each of them includes a discrete parameter which controls its complexity. It is preferable to use the simplest explanation of the underlying structure and hence the lowest possible complexity; section 2.3 discusses techniques for selecting suitable values for this parameter. Section 2.4 introduces robustness and describes the method used in this thesis to make the models robust. Section 2.5 discusses the means of assessing the quality of a model. A summary, in section 2.6, describes how temporal structure is to be included in the models and offers an overview of the structure of the remaining chapters.

2.1. Bayesian probability theory

Classical statisticians were frequentists, that is they defined probability in terms of the long-term frequency of the outcomes of repeatable events. For example, with a fair coin it would be expected that over a large number of tosses 50% would come down heads and 50% tails. It was Bernoulli who first turned the problem on its head by posing such questions as: if after ten tosses of a coin the result is three heads and seven tails, how likely is it that the coin is fair? In response Bayes [1763] (reprinted in [Barnard and Bayes, 1958]) and Laplace [1812] redefined the concept of probability to mean the degree of uncertainty given the available evidence. In their terms a compact probability distribution indicates a high degree of certainty, while a wide spread indicates lack of certainty [Bernardo and Smith, 1994].

Consider two random variables, \( X \) and \( Y \) (no distinction will be made between the variables and sample values of the variables). Using standard notation, \( p(X) \) denotes the probability of the value \( X \) occurring as a sample from its distribution, \( p(X \mid Y) \) denotes the conditional probability of \( X \) given that we already know the value of \( Y \), and \( p(X, Y) \) denotes the probability of \( X \) and \( Y \) together. If \( X \) and \( Y \) are independent then \( p(X, Y) = p(X)p(Y) = p(Y)p(X) \).

Using this notation we may state the two standard rules of probability:

\[
\text{Sum rule: } \quad p(X) = \begin{cases} 
\int p(X, Y) \, dY & \text{if } Y \text{ is continuous} \\
\sum_Y p(X, Y) & \text{if } Y \text{ is discrete}
\end{cases}
\]

\[
\text{Product rule: } \quad p(X, Y) = p(X \mid Y)p(Y)
\]
Combining these with the symmetry that \( p(X, Y) \) is equal to \( p(Y, X) \), we may arrive at Bayes’ Theorem:

\[
p(X | Y) = \frac{p(Y | X) p(X)}{p(Y)}
\]

(2.1)

Instead of \( X \) and \( Y \) we may think of a set of observed data, \( D \), and a model that defines the relationship between \( D \) and a set of parameters, \( \Omega \). The likelihood, \( p(D | \Omega) \), encodes what we know about the relationship between \( \Omega \) and \( D \). In particular it defines \( D \) as a function of \( \Omega \) and thus describes how \( D \) is generated from \( \Omega \). Hence we may say that the likelihood defines a generative model for \( D \) based on \( \Omega \). The prior probability, \( p(\Omega) \), encodes any beliefs we might have about \( \Omega \) without any knowledge of \( D \). The posterior probability, \( p(\Omega | D) \), is the probability \( \Omega \) after knowledge of \( D \) has been gained. The evidence is the normalisation factor that ensures that the posterior is a proper probability distribution, i.e. that the total posterior probability is 1. Assuming, as is likely, that \( p(D) \) is not known, the sum rule may be used to calculate it:

\[
p(D) = \int p(D, \Omega) d\Omega
\]

(2.2)

which is known as marginalising out \( \Omega \). This expression represents the probability of \( D \) having considered all possible values of \( \Omega \).

In the frequentist paradigm the parameters in \( \Omega \) are point values to be determined through some optimisation process; in the Bayesian paradigm they are considered to be random variables for which probability distributions are estimated. To emphasise this difference, the expression “parameter variables” is used throughout this thesis to describe the parameters in a Bayesian model.

There are a number of methods for determining point values for the model parameters. One prominent approach is that of maximum likelihood. This seeks values for the parameters which maximise the likelihood of the data, i.e. \( p(D | \Omega) \). Usually it is more convenient mathematically to maximise the logarithm of the likelihood (referred to as the log likelihood) which is exactly equivalent due to the convexity of the logarithm function. In the context of machine learning the negative log likelihood is regarded as the error function (for example [Bishop, 2006]) and minimising the error necessarily maximises the likelihood. Sometimes maximising the likelihood does not lead to a closed-form, analytic solution; in these cases iterative numerical methods may be used (see, for example, [Fletcher, 1987] and [Nocedal and Wright, 1999]) or, as will be shown in section 2.2.3 and applied in chapter 3, the expectation maximisation (EM) algorithm [Dempster et al., 1977; McLachlan and Krishnan, 1997] may be employed when there are unobserved (latent) variables or where data is missing.

**Maximum a posteriori** (MAP) estimation summarises the posterior distribution of the
model parameters. Here priors are specified for the model parameters, giving

\[ p(\Omega | D) \propto p(D | \Omega) p(\Omega) \]  

(2.3)

for a given set of data. The right-hand side of this expression is then maximised with respect to all the parameters, again resulting in point estimates for each of them. Variants of the EM algorithm or other iterative algorithms can be used to find the parameters’ values. Note that this method makes use of priors, but avoids the difficult evidence calculation. It may be written as a penalised maximum likelihood method, which is made clearer by writing

\[ \log(p(\Omega | D)) = \log(p(D | \Omega)) + \log(p(\Omega)) + \text{const} \]

(2.4)

so that \( \log(p(\Omega | D)) \) is regarded as the log likelihood penalised by the log prior.

In the Bayesian paradigm each of the model parameters in \( \Omega \) are regarded as random variables for which probability distributions must be determined using Bayes’ Theorem:

\[ p(\Omega | D) = \frac{p(D | \Omega) p(\Omega)}{p(D)} \]

(2.5)

Determining posterior distributions for the parameters will be referred to as learning or inferring the Bayesian model.

The difference between the frequentist and Bayesian paradigms is demonstrated by an exercise carried out by Laplace [1812]. Using the new Bayesian approach he calculated that a contemporary’s estimate of the mass of the planet Saturn was likely to have a degree of uncertainty, or error, of less than 1%\(^1\). Such a calculation is not meaningful to frequentists; the mass of a planet is not a random variable which might take different values. In effect that would require multiple instances of Saturn. This logical argument and the fact that the calculations required for the Bayesian technique are often intractable meant that the frequentist approach continued to prevail well into the 20th century (e.g. [Stigler, 1999]).

**2.1.1. Bayesian inference**

Bayesian inference is the process by which Bayes’ Theorem is used to draw conclusions from observed data. In terms of generative modelling it is the process by which a Bayesian model is trained, resulting in posterior distributions for each of the random variables that are the model’s parameters. A fully Bayesian technique avoids the pitfalls associated with the maximum likelihood method: integrating (averaging) over all possible values of the parameter variables guards against overfitting and posterior probability distributions are calculated for each of them, from which we obtain a measure of confidence in the inference.

The first step of the process is to specify the three probabilities on the right-hand side of

---

\(^1\)In fact the estimate of Saturn’s mass being 1/3512th of the mass of the sun is just 0.63% different from NASA’s current estimate [2010a] of 1/3499th.
Bayes’ Theorem (2.5): the likelihood, priors for each parameter variable and the evidence.

The likelihood and priors are often selected from a small set of standard probability distributions, such as the Gaussian (Normal) distribution, which have convenient mathematical properties. Parametric distributions, such as the Gaussian, are defined by a small number of parameters, referred to as the sufficient statistics; in the Gaussian case these are the mean and the variance, which define the location and spread of the values of the variable respectively. One particularly convenient mathematical property is that of conjugacy, where the application of Bayes’ Theorem results in a posterior that has the same functional form (belongs to the same family of distributions) as the prior. In this case the family of the prior distribution is said to be a conjugate family to the likelihood. Conjugacy enables the posterior from one set of observations to become the prior when new samples of the data are observed.

The sufficient statistics of the priors are point values, referred to as hyperparameters in the Bayesian inference procedure. Their values are fixed to encode any beliefs we might have about the shape of the prior distributions before any data has been observed. If the prior knowledge is known to be accurate, then the distribution may be tightly spread around the known location; in terms of a Gaussian, the variance would be very small to indicate the certainty. If there is no prior knowledge at all then an uninformative or vague prior may be selected – one whose spread is so wide that the probability distribution is, to all intents and purposes, flat, indicating a large amount of uncertainty.

While selecting a distribution (e.g. Gaussian) for the likelihood might be driven by mathematical convenience, the relationship between the data and the model’s parameter variables also needs to be decided. This may arise naturally from knowledge about how the data were generated, or it may also be chosen from a standard definition which is flexible enough to adequately explain the data. One family of such definitions is the linear latent variable models, which assume that the data were generated from a linear combination of a smaller number of hypothetical variables, whose values are unknown.

Having defined the likelihood and priors, the evidence may be calculated by marginalising out the parameters from the joint probability distribution:

\[
p(\mathcal{D}) = \int p(\mathcal{D}, \Omega) \, d\Omega = \int p(\mathcal{D} | \Omega) \, p(\Omega) \, d\Omega \tag{2.6}
\]

Finally the posterior distributions for each parameter variable are calculated by combining the likelihood, priors and evidence using Bayes’ Theorem.

Although the procedure sounds simple, in practice performing the integration in (2.6) is often intractable. For this reason the application of the Bayesian method only emerged in the 1930s [Stigler, 1999; Fienberg, 1992], and only really took off in the 1980s after the development of numerical sampling techniques which allowed approximations to be made in the inference process and when computers became powerful enough to perform

---

1 There is some discussion as to whether an uninformative prior is truly uninformative; see, for example, [Carlin and Louis, 1996, section 2.2.3]
the sampling for more than trivial problems in an acceptable time-frame.

2.1.2. Bayesian approximation techniques

If the evidence cannot be calculated exactly, there are a range of methods for approximating either the evidence or the posterior distributions.

If the number of parameters in the model is small, the integration required for the calculation of the evidence may be approximated using a numerical technique such as quadrature (see, for example, [Naylor and Smith, 1982]), but this method is infeasible if the number of parameters is large.

The Laplace Approximation approximates a posterior distribution with a Gaussian centred on the MAP estimate. Under certain regularity conditions and as the number of observations tends to infinity, the posterior distribution approaches a Gaussian. Often the regularity conditions are not met, and the Gaussian approximation does not hold for small datasets.

When the evidence cannot be calculated exactly, it may still be possible to approximate the integral by sampling from it using the Monte Carlo method [Denison et al., 2002]. If the posterior distribution can be determined up to proportionality (in other words if the intractable evidence integral is ignored), then a Markov chain Monte Carlo (MCMC) method, such as the Metropolis-Hastings method [Metropolis et al., 1953; Hastings, 1970] or Gibbs sampling [Geman and Geman, 1984], may be used to approximate the posterior. Green’s [1995] Reversible Jump MCMC (RJMCMC) generalises the Metropolis-Hastings algorithm to allow for the sampling from distributions of varying dimensionality. This technique is useful when the complexity of a given model can vary and we wish to average over the different complexity models (see, for example, [Zhang et al., 2004]). The accuracy of all of these Monte Carlo methods increases as the number of samples increases and the first samples in a set are often discarded as being unrepresentative of the distribution being approximated. Detecting when sufficient samples have been generated is difficult, especially for RJMCMC (Denison et al. [2002] give a number of references). For models with large numbers of parameters, such as will be introduced in this thesis, these sampling methods are computationally very expensive (large numbers of samples are needed to sample the whole parameter space) and are also slow to converge when there are significant dependencies between parameters, as there are in these new models.

Another method for approximating the evidence is known in different areas as the evidence approximation (in machine learning) [Gull, 1989; MacKay, 1992], empirical Bayes (statistics) [Bernardo and Smith, 1994], type II maximum likelihood [Berger, 1985] and generalised maximum likelihood [Wahba, 1975]. If it is not tractable to calculate the evidence by marginalising out all the parameter variables, it is often possible to marginalise out some of them. In which case the tractable portion of the integral is performed and maximum likelihood estimates calculated for the remaining parameters.

In this thesis a variational Bayesian inference technique has been used to approximate the
posterior distributions. This results in posterior distributions for all parameter variables in the model and avoids the computational expense of the sampling methods.

2.1.3. Variational Bayesian approximation

The variational Bayesian inference technique minimises the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951; Cover and Thomas, 1991] between the approximate and actual posterior distributions to determine the optimal hyperparameter values for the approximations; for tutorials see [Jordan et al., 1999], [Lappalainen and Miskin, 2000], and [Bishop, 2006, chapter 10]. Representing the approximation posterior distributions as $q(\cdot)$, variational inference seeks $q(\Omega | D) \approx p(\Omega | D)$, where, as before, $\Omega$ is the set of model parameter variables and $D$ the data. There is nothing inherently approximate about variational methods, but they lend themselves to approximating posterior distributions; the approximations come about because we make assumptions about the factorisation of the approximate posteriors and the families of distributions to which they belong and those assumptions may not hold for the true posteriors.

The evidence may be written as

$$p(D) = \int p(D, \Omega) \, d\Omega \quad (2.7)$$

After straightforward manipulation we obtain

$$\log(p(D)) = \log \left( \int q(\Omega | D) \frac{p(D, \Omega)}{q(\Omega | D)} \, d\Omega \right) \quad (2.8)$$

Since $q(\Omega | D)$ is a probability density function (and hence $\int q(\Omega | D) \, d\Omega = 1$) and log(·) is a convex function, we may apply Jensen’s inequality to find a lower bound for the true log posterior:

$$\log(p(D)) = \log \left( \int q(\Omega | D) \frac{p(D, \Omega)}{q(\Omega | D)} \, d\Omega \right) \geq \int q(\Omega | D) \log \left( \frac{p(D, \Omega)}{q(\Omega | D)} \right) \, d\Omega \quad (2.9)$$

By another simple manipulation the log evidence on the left-hand side of (2.8) may be written as

$$\log(p(D)) = -\text{negative variational free energy} - \text{KL divergence} \quad (2.11)$$

The negative variational free energy, $F(q)$, is the same as the lower bound shown in (2.9). This means that the second term, which is the KL divergence between the approximate posterior $q(\Omega | D)$ and the true posterior $p(\Omega | D)$ (written as $\text{KL}(q \parallel p)$), must necessarily be greater than or equal to zero, being zero only when the two distributions are the same. To make the best possible approximation we must maximise the lower bound and thus
minimise the KL divergence. This is practical if \( p(D, \Omega) \) and \( q(\Omega | D) \) can be factorised into simple terms, leading to the logarithms in \( F(q) \) being split into a sum of multiple manageable terms. Often the definition of the likelihood and the priors associated with a model factorise readily. The key assumption in the variational Bayesian method that makes it tractable is that the approximate posterior distributions are independent of one another so that we may write

\[
q(\Omega | D) = \prod_i q(\Omega_i | D) \tag{2.12}
\]

where \( \Omega_i \) is the \( i \)th parameter variable in \( \Omega \). Once parametrised families for each of the \( q(\Omega_i | D) \) have been chosen (for example, Gaussians), \( F(q) \) can be maximised in the usual fashion by differentiating with respect to each of the parameters of the selected distributions.

The differentiation method requires us to decide in advance which family of distributions each of the \( q(\cdot) \) belongs to. However, an elegant solution is provided by Waterhouse et al. [1995] (see also [Attias, 2000], [Beal and Ghahramani, 2002] and [Beal, 2003]) which exploits the factorisation of the approximate posterior (2.12). It is simpler to calculate and has the added advantage that the exact form of the \( q(\cdot) \) distributions are determined as part of the calculation. For readability let \( Q_i \) represent \( q_i(\Omega_i | D) \) where \( \Omega_i \) is one of the parameter variables (or a group of parameter variables):

\[
F(q) = \int Q \log \left( \frac{p(D, \Omega)}{Q} \right) d\Omega \tag{2.13}
\]

\[
= \int \left( \prod_{i=1}^{G} Q_i \right) \log(p(D, \Omega)) d\Omega_1, \ldots, d\Omega_G - \int \left( \prod_{i=1}^{G} Q_i \right) \left( \sum_{i=1}^{G} \log(Q_i) \right) d\Omega_1, \ldots, d\Omega_G \tag{2.14}
\]

Considering just one of these integrals, say with respect to \( \Omega_j \), and keeping all the other \( \Omega_i \neq j \) (and hence \( Q_i \neq j \)) fixed, this may be rewritten as

\[
F(q) = \int Q_j \left[ \int \log(p(D, \Omega)) \prod_{i \neq j} Q_i d\Omega_{i \neq j} \right] d\Omega_j - \int Q_j \log(Q_j) d\Omega_j + \text{const} \tag{2.15}
\]

where terms not dependent on \( \Omega_j \) have been absorbed into the constant. Notice that the term in square brackets is the expectation of \( \log(p(D, \Omega)) \) with respect to all of the \( Q_j \), except \( i = j \). Let us call this \( \mathbb{E}[\log(p(D, \Omega))]_{i \neq j} \). Comparing (2.15) with (2.11) we see that this is the negative KL divergence between \( Q_j \) and \( \mathbb{E}[\log(p(D, \Omega))]_{i \neq j} \) and hence the maximum value is zero, which is obtained when

\[
\log(Q_j) = \mathbb{E}[\log(p(D, \Omega))]_{i \neq j}. \tag{2.16}
\]

When conjugate priors are selected for each \( \Omega_i \) the approximate posterior turns out to have the same form as the prior [Attias, 2000; Ghahramani and Beal, 2001] and the approximate variational posteriors may be calculated by evaluating (2.16) for each parameter in turn.
2. Background

This does not generally give a closed form solution as the hyperparameters for one posterior distribution may depend on other posteriors, so an iterative process is required whereby the hyperparameters for each posterior are evaluated cyclically until convergence. Ghahramani and Beal [2001] show that this method converges on a local maximum for $F(q)$, thus locally minimising $\text{KL}(q \parallel p)$.

2.2. Latent Variable Models

Considering the example datasets shown in figures 1.2 to 1.4 it is easy to imagine that there are some underlying processes which we cannot measure directly (perhaps because we do not know what they are), but are observing indirectly. For the UK industrial production figures there has been an underlying general trend away from manufacturing towards service industries in the UK. For the NASA SeaWiFS images different types of pollution probably play a part in determining phytoplankton levels. With EEG we can only observe the working of the human brain from outside the body and a single signal generated deep inside the brain might be observed at different strengths by more than one of the electrodes.

Latent variable models are a family of probabilistic models that explicitly model noise and which try to find hidden structure in data. They try to explain the statistical properties of a set of observed variables through the use of an alternative, hypothetical set of hidden, or latent, variables. These latent variables are usually of lower dimensionality than the original observations, sometimes very much lower, which lends them to be used as methods of data compression, where the latent variables are used in preference to the observations. They may also indicate a generative process which gave rise to the data.

Latent variable models assume that correlations between a set of observed variables are as a result of them being generated from the mixing together of a smaller set of variables whose values we are not able to observe. This smaller set of hidden or latent variables has two uses: it may be used as a method of data compression or dimensionality reduction, where the latent variables are used in preference to the higher-dimensioned observed variables, and it may lead to greater understanding of the structure of the original data.

Generally latent variable models for regression analysis take the form

$$\text{observations} = f(\text{latent variables}) + \text{noise} \quad (2.17)$$

where $f(\cdot)$ is some deterministic function. They assume that the observations have some joint probability distribution which is conditioned on the latent variables. A crucial assumption of these models is that any structure in the observations is a direct consequence of structure in the latent variables and that the observations are independent of one another given this conditioning. If the observations are recorded as $T = (t_1, ..., t_N)$, where each of the $t_n$ are $D$-dimensional column vectors, and the $Q$-dimensional latent variables
as \( X = (x_1, ..., x_N) \), then we may write this independence assumption as

\[
p(T \mid X) = \prod_{n=1}^{N} p(t_n \mid X)
\]  

(2.18)

In order to construct a probabilistic model for a particular set of data, we need to infer values for the latent variables given the observations; in particular we want to determine \( p(X \mid T) \), the posterior probability of the latent variables given all the observations. Applying Bayes’ rule we know that

\[
p(X \mid T) = \frac{p(T \mid X) p(X)}{p(T)}
\]  

(2.19)

where

\[
p(T) = \int p(T \mid X) p(X) \, dX
\]  

(2.20)

We do not know what the relationship between the data and the latent variables actually is and we therefore do not have a definition for the true likelihood \( p(T \mid X) \). Looking for the simplest explanation for the data, it makes sense to represent each observation as a linear mixture of the latent variables, rather than using a more complicated, non-linear function for \( f(\cdot) \). This allows (2.17) to be written as the following expression:

\[
t_n = Wx_n + \bar{t} + e_n
\]  

(2.21)

where \( W \) is some \textit{mixing matrix} of dimension \( D \times Q \), \( \bar{t} \) allows the mean observation to be non-zero and \( e_n \) is the residual noise or error term. This definition is not in itself enough to completely specify the model as it does not determine the likelihood, \( p(t_n \mid W, X, \bar{t}, E) \) (where \( E = (e_1, ..., e_N) \)). There are a number of traditional linear mixing latent variable models which all fit the form shown in (2.21). Important ones for this thesis are Principal Component Analysis (PCA), Factor Analysis (FA) and Canonical Correlation Analysis (CCA). Technically PCA and CCA are not latent variable models as they do not model the noise and do not define probability models for the data, however latent variable model versions of them exist in Probabilistic PCA (PPCA) and Probabilistic CCA (ProbCCA, for ease of distinction from PPCA). They all assume that \( p(t_n \mid X) = p(t_n \mid x_n) \) and make Gaussian assumptions for the probability distributions of the latent variables and the noise, but differ in the assumptions they make regarding the form of the covariance of the noise. The Gaussian assumptions lead to a Gaussian form for the likelihood:

\[
p(t_n \mid W, x_n, \bar{t}, e_n) = \mathcal{N}(t_n \mid Wx_n + \bar{t}, \Sigma)
\]  

(2.22)

where \( \Sigma \) is the noise covariance. In these traditional models \( W \) and \( \bar{t} \) are parameters to be estimated. In the Bayesian scheme they become random variables requiring the definition of prior distributions for them.

This thesis is concerned with extending or generalising this particular class of linear mixing latent variable models. Figure 2.1 summarises the relationships and differences between them and in the following sections they are briefly described.
2. Background

Figure 2.1. The class of latent variable models extended or generalised in this thesis; their relationships and the assumptions they make about the form of the observation noise covariance.

2.2.1. Principal Component Analysis

Principal Component Analysis (PCA) [Jolliffe, 2002] is a well-known technique often used for dimensionality reduction, either for feature extraction or data compression. It is also variously known in different contexts as the Karhunen-Loève transform [Karhunen, 1947; Loève, 1945], Empirical Orthogonal Function analysis [Lorenz, 1956], Hotelling transform [Hotelling, 1933] or Proper Orthogonal Decomposition [Kosambi, 1943] (see for example [Berkooz et al., 1993]). In terms of latent variable models, it may be represented by the noiseless expression

$$ t_n = Wx_n + \bar{t} \quad (2.23) $$

where $W$ is a square, orthogonal matrix.

PCA may be used to transform the data or to approximate it. With a square, orthogonal $W$ PCA represents a rotation of $T$ (by $W^T$) onto new axes such that the primary axis (the first column of $W$) is in the direction of maximum variance, the secondary axis (the second column of $W$) is perpendicular to the first and in the direction of next most variance, etc. Thus the columns of $W$ are a new basis for the data and are orthogonal; by convention they are orthonormal (that is $W^T W = I$, where $I$ is the identity matrix).

The $D$ columns of $W$ are termed *components* and are usually ordered so that for any value of $Q$, the first $Q$ columns of $W$ (the *principal components*) capture the maximum possible amount of variance. The components may in themselves capture interesting information about the original observations (see, for example, [Everson et al., 1997]), but often PCA is used to reduce the dimensionality of data; by selecting the $Q$ principal components, where $Q$ is less than $D$ (often very much less), the resulting non-square, orthogonal $W$ may be used to project the data onto a lower-dimensional subspace while retaining the maximum possible amount of variance. The assumption is that directions of high variance
are somehow interesting, while those of low variance represent noise and may be discarded.

The foundations of PCA go back to Pearson [1901], who described the linear orthogonal projection \( x_n = W^T(t_n - \bar{t}) \) that minimises the squared reconstruction error, \( \sum_{n=1}^{N} \| t_n - \hat{t}_n \|^2 \), of the observations, where the reconstruction of observation \( t_n \) is given by \( \hat{t}_n = Wx_n + \bar{t} \). It is Hotelling’s later work [1933] that describes PCA in the more familiar terms of an orthogonal linear projection of the data onto a lower-dimensional subspace.

It can be shown (by minimising the approximation error subject to the constraint that \( W \) is orthonormal) that the components are eigenvectors of the sample covariance matrix

\[
S = \frac{1}{N} \sum_{n=1}^{N} (t_n - \bar{t})(t_n - \bar{t})^T
\]

and that the principal components \((w_1, \ldots, w_Q)\) are those with the largest associated eigenvalues \((\rho_1, \ldots, \rho_Q)\), such that \( Sw_q = \rho_q w_q \). Since \( S \) is symmetric, the eigenvectors are orthogonal and the eigenvalues are all greater than or equal to zero. If \( t_n \) is \( D \)-dimensional, where \( D \leq N \), then \( S \) will generally have \( D \) eigenvectors with non-zero eigenvalues. However, if \( D > N \) (as it is in the NASA phytoplankton images), then the last \( N - D \) eigenvalues will be zero and the eigenvectors, while retaining the orthogonality constraint of the eigenvectors of a covariance matrix, will be arbitrary. In this case the snapshot method [Sirovich, 1987] can be used to efficiently find the eigenvectors by solving the smaller \((N \times N)\) eigenvalue problem \( \frac{1}{N} T^T T v_q = \rho_q v_q \) resulting in \( w_q \propto v_q^T T \).

Since the columns of \( W \) are orthogonal, the projections of observations onto each component are uncorrelated, with \( \rho_q \) being the variance of the projection of \( T \) onto \( w_q \). Each element of \( x_n \) is thus uncorrelated, and the covariance matrix \( \sum_{n=1}^{N} x_n x_n^T / N \) is diagonal with the \( q \)th eigenvalue, \( \rho_q \), as the \( q \)th entry on the diagonal.

### 2.2.2. Factor Analysis

Factor Analysis (FA) is possibly the most commonly-used latent variable model. It was first posited by Spearman [1904], a psychologist, who was interested in the fact that subjects who did well in tests of mental ability tended to do well in other such tests, leading him to suppose that there was some generalised, underlying ability driving the correlation between performances. His idea of one general factor was expanded to the multi-factor case by Thurstone [1931], who was the first to name it factor analysis.

In FA the columns of \( W \) are referred to as the *loadings* and the latent variables as *factors*. The noise term, \( e_n \), is assumed to be Gaussian with zero mean and diagonal precision matrix, \( \text{diag}(\lambda) \), and, conventionally, \( x_n \sim \mathcal{N}(0, I) \), with \( Q < D \). This leads to the conditional distribution for the observations:

\[
p(t_n | W, x_n, \bar{t}, \lambda) = \mathcal{N}(t_n | Wx_n + \bar{t}, \text{diag}(\lambda)^{-1})
\]
and the likelihood for the observations of

$$p(t_n \mid W, \bar{t}, \lambda) = N(t_n \mid \bar{t}, WW^T + \text{diag}(\lambda)^{-1})$$  \hspace{1cm} (2.26)$$

The model parameters, $W$, $\bar{t}$ and $\lambda$, may be estimated using a maximum likelihood method, but unlike PCA there is no single-step algebraic solution, so an iterative method, such as the expectation maximisation (EM) algorithm must be used [Lewis-Beck, 1993].

It may be seen from (2.25) that there is ambiguity in the determination of the latent variables. Consider some arbitrary $Q \times Q$ rotation matrix, $R$, such that $R^T R = I$. Then the expression $WRx_n + \bar{t} + e_n$ leads to a probability distribution for $t_n$ that is indistinguishable from (2.25) and the likelihood is unchanged from (2.26). There are methods for selecting a particular rotation. Varimax [Kaiser, 1958] seeks the rotation that gives the simplest structure to the loadings, by which it means loading values that are either zero or large. Varimax works on the columns of $W$; Quartimax [Neuhaus et al., 1954; Saunders, 1953; Ferguson, 1954] is a similar method which works on the rows of $W$. Both of these methods employ orthogonal rotations. Non-orthogonal (oblique) rotation methods are also available; see, for example, [Tabachnick and Fidell, 1996] and [McDonald, 1985].

### 2.2.3. Probabilistic PCA

Roweis [1998] and Tipping and Bishop [1999] have observed that the subspace defined by the columns of $W$ in FA will not, in general, correspond to that defined by the columns of $W$ in PCA, because of the distinction in FA between variance and covariance. However, as was investigated by Lawley [1953] and Anderson and Rubin [1956], the two are related when the noise covariance is isotropic, that is when the precision is $\lambda I$ rather than $\text{diag}(\lambda)$, where $I$ is the identity matrix. This leads to a formulation of PCA with a noise model as Probabilistic PCA (PPCA):

$$t_n = Wx_n + \bar{t} + e_n$$  \hspace{1cm} (2.27)$$

$$p(t_n \mid x_n) = N(t_n \mid Wx_n + \bar{t}, \lambda^{-1}I)$$  \hspace{1cm} (2.28)$$

$$p(x_n) = N(x_n \mid 0, I)$$  \hspace{1cm} (2.29)$$

which also results in a Gaussian likelihood:

$$p(t_n \mid W, \bar{t}, \lambda) = N(t_n \mid \bar{t}, WW^T + \lambda^{-1}I)$$  \hspace{1cm} (2.30)$$

Tipping and Bishop extend this analysis to show that in the isotropic case the maximum likelihood estimates for the model parameters $W$ and $\lambda$ correspond to those found by PCA, up to a rotation of $W$. They show that, in this case, the columns of $W$ span the same subspace as the principal components. They offer two methods for estimating $W$ and $\lambda$. The first is based on the maximum likelihood EM algorithm for FA by Rubin and Thayer [1982] which alternates the following expectation and maximisation steps until
convergence (the mean observation $\bar{t}$ is found non-iteratively as $\bar{t} = \frac{1}{N} \sum_{n=1}^{N} t_n$):

**E-step:**
Finds the sufficient statistics of the latent variables given the current parameter estimates:

\[
M = W^T W + \lambda^{-1} I \quad (2.31)
\]

\[
\langle x_n \rangle = M^{-1} W^T (t_n - \bar{t}) \quad (2.32)
\]

\[
\langle x_n x_n^T \rangle = \lambda^{-1} M^{-1} + \langle x_n \rangle \langle x_n \rangle^T \quad (2.33)
\]

where $\langle x_n \rangle$ and $\langle x_n x_n^T \rangle$ are the expectations of $x_n$ and $x_n x_n^T$ respectively, conditioned on $t_n$.

**M-step:**
Updates the parameter estimates to maximise the complete data likelihood (that is, $p(T, X | W, \bar{t}, \lambda)$):

\[
W = \left[ \sum_{n=1}^{N} (t_n - \bar{t}) \langle x_n \rangle^T \right] \left[ \sum_{n=1}^{N} \langle x_n x_n^T \rangle \right]^{-1}
\]

\[
\lambda^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \left[ \| t_n - \bar{t} \|^2 - 2 \langle x_n \rangle^T W^T (t_n - \bar{t}) + \text{trace}(W \langle x_n x_n^T \rangle W^T) \right]
\]

In contrast to FA, Tipping and Bishop show that it is also possible to find a single-step algebraic solution for determining $W$ and $\lambda$:

\[
W = U_Q (P_Q - \lambda^{-1} I)^{1/2} R
\]

\[
\lambda^{-1} = \frac{1}{D-Q} \sum_{j=Q+1}^{D} \rho_j
\]

where $U_Q$ are the $Q$ principal eigenvectors of the sample covariance matrix, with their associated eigenvectors as the diagonal matrix $P_Q$. $R$ is an arbitrary orthogonal rotation matrix (this represents the same rotational ambiguity as described for FA) and $\rho_j$ is the eigenvalue associated with the $j$th eigenvector. From (2.37) it may be seen that $\lambda^{-1}$ represents the average amount of variance associated with the lost dimensions. This method also maximises the likelihood and is much faster to compute than the iterative EM algorithm.
2. Background

nth observation, \( \hat{t}_n \), is given by

\[
\hat{t}_n = W(W^TW)^{-1}M\langle x_n \rangle
\]  

(2.38)

The columns of \( W \) may not be orthogonal since

\[
W^TW = R^T(P_Q - \lambda^{-1}I)R
\]  

(2.39)

is not diagonal unless \( R = I \), but they span the same subspace as the principal components \( U_Q \). This rotational ambiguity may be resolved by noting that (2.39) represents an eigenvector decomposition of \( W^TW \) and the columns of \( R \) are the eigenvectors of \( W^TW \), enabling orthogonal principal components to be determined, with each component having the length of their associated eigenvalue, \( \rho_i \), minus \( \lambda^{-1} \).

A complaint levelled at FA [Chatfield and Collins, 1980; Bishop, 2006] is that the loadings discovered in a model of order \( Q \) is not necessarily a subset of the \( Q + 1 \) model; in PCA the first \( Q \) principal components are always a subset of the \( Q + 1 \).

2.2.4. Canonical Correlation Analysis

While FA and (P)PCA aim to find structure within a single set of multivariate observations, Canonical Correlation Analysis (CCA) aims to find structure shared between two sets of multivariate observations. First proposed by Hotelling [1936], it seeks a subspace such that the correlation of the linear projections of the variables onto the subspace is mutually maximised.

Given sets of observations of two random variables, \( T^{(1)} \) and \( T^{(2)} \), then CCA seeks pairs of vectors, \( a_i \) and \( b_i \), such that the linear correlation \( \rho = \text{corr}(a_i^T T^{(1)}, b_i^T T^{(2)}) \) is maximised, subject to the constraint that successive \( a_i \) and \( b_i \) are orthogonal to previous \( a \) and \( b \) vectors respectively. The linear functions \( a_i^T T^{(1)} \) and \( b_i^T T^{(2)} \) (analogous to the projections onto the latent variables in (P)PCA or factors in FA) are referred to as canonical variates.

If \( \Sigma_{11} \) is the covariance of \( T^{(1)} \), \( \Sigma_{22} \) is the covariance of \( T^{(2)} \) and \( \Sigma_{12} = \Sigma_{21} \) is the covariance between the two variables, then CCA seeks to maximise

\[
\rho = \frac{a_i^T \Sigma_{12} b_i}{\sqrt{a_i^T \Sigma_{11} a_i} \sqrt{b_i^T \Sigma_{22} b_i}}
\]  

(2.40)

which leads to

\[
a = \text{eigenvector of } \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}
\]

\[
b = \text{eigenvector of } \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}
\]  

(2.41)

The symmetric nature of covariance matrices guarantees the orthogonality of the eigenvectors.

Horst [1961], Kettenring [1971] and [Hardoon et al., 2003] all generalise CCA to apply to
2. Background

more than two sets of observations.

2.2.5. Probabilistic CCA

Like PCA, CCA does not include a noise term and thus is not strictly a latent variable model. Bach and Jordan [2005] propose a probabilistic model for CCA (ProbCCA, for ease of distinguishing it from PPCA) where the canonical variates are assumed to be exactly correlated and the aim is to find the best linear projections of the variables onto the subspace which retain the maximum variance. The model is constructed in a similar fashion to Tipping and Bishop’s [1999] PPCA. They construct a latent variable model where the two sets of observations, \( \mathbf{T}^{(1)} \) and \( \mathbf{T}^{(2)} \), having dimensions \( D^{(1)} \times N \) and \( D^{(2)} \times N \) respectively, share a common latent variable in \( \mathbf{x}_n \):

\[
\begin{align*}
\mathbf{t}^{(1)}_n &= \mathbf{W}^{(1)} \mathbf{x}_n + \bar{\mathbf{t}}^{(1)} + \mathbf{e}^{(1)}_n \\
\mathbf{t}^{(2)}_n &= \mathbf{W}^{(2)} \mathbf{x}_n + \bar{\mathbf{t}}^{(2)} + \mathbf{e}^{(2)}_n
\end{align*}
\]  

(2.42)

where the error terms, \( \mathbf{e}^{(1)}_n \) and \( \mathbf{e}^{(2)}_n \) are assumed Gaussian distributed with zero means and covariance matrices \( \mathbf{G}^{(1)} \) and \( \mathbf{G}^{(2)} \) respectively. These two expressions may be combined into a single one by stacking the \( \mathbf{t}^{(\ast)}_n \) (where \( \ast \) refers generically to one of the sets of observations), \( \mathbf{W}^{(\ast)} \), \( \bar{\mathbf{t}}^{(\ast)} \) and \( \mathbf{e}^{(\ast)}_n \) terms:

\[
\begin{pmatrix}
\mathbf{t}^{(1)}_n \\
\mathbf{t}^{(2)}_n
\end{pmatrix} =
\begin{pmatrix}
\mathbf{W}^{(1)} \\
\mathbf{W}^{(2)}
\end{pmatrix}
\mathbf{x}_n +
\begin{pmatrix}
\bar{\mathbf{t}}^{(1)} \\
\bar{\mathbf{t}}^{(2)}
\end{pmatrix} +
\begin{pmatrix}
\mathbf{e}^{(1)}_n \\
\mathbf{e}^{(2)}_n
\end{pmatrix}
\]  

(2.43)

\[
\mathbf{t}_n = \mathbf{Wx}_n + \bar{\mathbf{t}} + \mathbf{e}_n
\]  

(2.44)

This is exactly the form of the FA and PPCA latent variable models. However, where FA and PPCA assume that the observed variables are independent conditioned on the latent variables, ProbCCA assumes that the two sets of observations, \( \mathbf{T}^{(1)} \) and \( \mathbf{T}^{(2)} \), are conditionally independent of one another, leading to a Gaussian conditional distribution for \( \mathbf{t}_n \sim \mathcal{N}(\mathbf{t}_n \mid \mathbf{Wx}_n + \bar{\mathbf{t}}, \mathbf{B}^{-1}) \), where \( \mathbf{B} \) is a block diagonal precision matrix with each \( \mathbf{t}^{(\ast)}_n \) having a full \( D^{(\ast)} \times D^{(\ast)} \) precision matrix, \( \mathbf{B}^{(\ast)} = \mathbf{G}^{(\ast)} \mathbf{G}^{(\ast)\top} \).

Bach and Jordan [2005] show that, in a similar way to Tipping and Bishop’s [1999] PPCA, the maximum likelihood solution for (2.42) is related to the linear algebra solution up to a rotation for each set of canonical variates.

2.3. Model order selection

There may be many different models that describe a particular dataset, in which case, applying Occam’s Razor, we would like to choose the one which offers the simplest explanation of the underlying structure. An overly complicated model may fit the given data more precisely, but it may generalise poorly, that is it may offer a much less precise explanation of similar data (perhaps the same variables measured over a different time
2. Background

Figure 2.2. Examples of underfitting and overfitting of a model to the data. The red dots represent observations of some unknown underlying model (a sine wave in this example), with added noise. The blue lines represent hypothetical models which might be used to fit the data. Plot (a) demonstrates underfitting; the line, described by a single parameter, does not capture the underlying shape of the data. Plot (b) demonstrates a model which has captured the essence of the data’s structure. Plot (c) demonstrates overfitting; the model, described by as many parameters as observations, has modelled the data and the noise and will generalise poorly to new data.

For (P)PCA the complexity of the model is controlled by the number of principal components used and thus the model order is $Q$.

For a non-probabilistic model such as PCA it is difficult to determine in a robust and efficient way how many principal components should be used for a given set of data. For a complete set of observations (i.e. where there are no missing values) the best-fitting model will be that which retains all the components. However, such a model will generalise poorly to the estimation of missing values as it is modelling noise which is unique to the observed data. In any case often the objective is to reduce the number of components to arrive at a lower-dimensional description, so there needs to be some method of culling them. Jolliffe [1972] and Beale et al. [1967] describe a number of methods:

1. In the case where data is missing, select those principal components that account for some percentage of the total variation in the data. This is straightforward because the variance associated with the first $Q$ components is the sum of their associated eigenvalues.

2. Again in the case where data is missing, select those principal components whose associated eigenvalues exceed some value, i.e. those that individually account for more than a certain amount of variance.

3. Plot the logarithm of the eigenvalues and select as the cutoff point the position at which the graph “turns”. This is the scree method [Cattell, 1966] and is based on the idea that in the limit of a large amount of data the eigenvalues associated with noise will be flat; with finite data it is hoped that a turning point can be used to
2. Background

identify the noise values.

4. Successively test the null hypothesis that the $Q$th eigenvalue is equal to the $(Q + 1)$th eigenvalue and select the value of $Q$ where the hypothesis is rejected [Daultrey, 1976].

5. Declare some randomly-selected data as being missing. For each value of $Q$ calculate the error associated with the reconstruction of the missing values using the $Q$ principal components and select the value at which this error is minimised, or the smallest value of $Q$ for which the error is lower than some threshold.

There are many other methods suggested in the literature (see, for example, [Jackson, 1993; Férre, 1995; Peres-Neto et al., 2004] and [Jolliffe, 2002, chapter 6]), but they tend to be rather ad hoc. In many of these tests the problem has moved from one of estimating how many principal components to use to what value a threshold should be set to and no principled criterion is available. They are also data-specific, requiring multiple tests for each set of data to determine the best number to use.

For a probabilistic model with a maximum likelihood algorithm it is possible to extend the variety of methods that can be used. A simple method is to calculate the likelihood of the data for each value of $Q$ and select the $Q$ that maximises it. On its own this will always select every component, so some method of penalising overly-complicated models (i.e. those with large $Q$) needs to be used; for example, Akaike’s Information Criterion (AIC) [Akaike, 1974], Bayesian Information Criterion (BIC) [Schwarz, 1978], also called the Schwarz criterion, or Minimum Description Length (MDL) [Rissanen, 1978]. Clearly this requires the model to be trained multiple times, once for each $Q$ to be tested. An even more computationally expensive method is to calculate the complete data log likelihood at each iteration of the EM algorithm, penalised as before, and at each iteration select the $Q$ that maximises it. It is easy to see that the $Q$ selected in the early iterations can unnecessarily limit the range of values for later iterations. Minka [2000] places priors over all the variables and then maximises a Laplace approximation to the evidence $p(T | Q)$. Despite promising results on low-dimensional, synthetic data, it was found that the resulting calculations do not take account of zero-value eigenvalues arising when the number of dimensions exceeds the number of observations, as for the NASA data. This eventuality results in singularities in the calculations. The problem is solved by limiting the calculations to non-zero eigenvalues (see [Hoyle, 2008]).

Each of these methods can be computationally expensive, especially if it is decided to select any subset of $Q$ components rather than the $Q$ principal components. In a series of tests on real data these methods tended to select almost all the principal components where a significantly sparser selection would have been preferable, but they do represent steps towards the automation of model order selection.

In a fully Bayesian scheme the model orders would be incorporated as parameters for which probability distributions are estimated. However, their discrete nature makes them difficult to integrate into the scheme. An alternative method, developed by Mackay [1994] and Neal [1995], is automatic relevance determination (ARD). In brief, a Gaussian prior is placed over the magnitude of each individual component which encodes the belief that the
2. Background

Figure 2.3. Demonstration of the effect on the maximum likelihood solution for a Gaussian distribution of a small number of outliers. Plot (a) shows a histogram of 30 samples taken from a Gaussian distribution with mean zero and standard deviation of 0.5. Overlaid in green is the Gaussian maximum likelihood estimate of the distribution. Plot (b) shows the same, but with the addition of three outliers. Notice how both the mean and the variance have been adversely affected.

magnitude of the component should be zero unless there is evidence in the data causing it to be non-zero. This method, employed, for example, by Bishop [1999b] to estimate the model order in a variational PPCA model, is used in chapters 4–7 of this thesis and is described in more detail there.

2.4. Robustness

Many real sets of data contain occasional values that are unrepresentative of the data as a whole, or data that “deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [Hawkins, 1980]. In the NASA dataset, for example, there are occasional single pixels which seem to indicate a high phytoplankton level when all the surrounding pixels indicate low levels. These may be caused by the presence of a solid object floating on the surface of the sea, such as a ship, or by a measurement or instrumentation error. We do not want these outlier values to have an adverse effect on the probability distributions estimated for our models as they are probably not truly representative of the variables we wish to measure. That is, we wish to make our models robust, in the sense defined by Huber [1972]: “insensitivity against small deviations in assumptions”. In the case of outliers their magnitude might be considered to represent large deviations, but the numbers of outliers is generally small with respect to the total number of observed values.

Gaussian models are known to be particularly susceptible to outliers. Figure 2.3 demonstrates the effect on the maximum likelihood solution for a Gaussian distribution of a small number of outliers. Plot (a) shows a histogram of 30 samples taken from a Gaussian distribution with mean zero and standard deviation of 0.5. Overlaid in green is the Gaussian maximum likelihood estimate of the distribution (see appendix B). Plot (b) shows the same, but with the addition of three outliers. Notice how both the mean and the vari-
ance have been substantially adversely affected. The Gaussian latent variable models, in which deviations from the mean are measured by the square of the deviation, are similarly susceptible to outlying observations.

There are many different methods of dealing with outliers [Stigler, 1973; Hampel, 1973; Huber, 1972; Barnett and Lewis, 1994]: there is a whole field devoted to identifying and removing them from the data before further processing, for example (see, for example, [Hawkins, 1980] and [Anscombe, 1960]). Rather than removing them, robust modelling seeks to accomodate them within the probabilistic model. The EEG data, for example, is leptokurtic in nature and removing what a Gaussian model would consider to be outliers would damage the data. One method of accomodating the outliers and enabling the modelling of leptokurtic datasets is to replace the Gaussian assumptions of the models so far discussed with mixtures of Gaussians in which one component models the majority of the observations while a second, wider component models the outliers.

The benefit to robustness of mixtures of Gaussians was, according to Stigler [1973], first recognised by Newcomb [1886]. An additional benefit is that these mixtures are able to model distributions that have heavier tails than Gaussians, such as the EEG data. With finite mixtures of Gaussians the tails still decay exponentially, like \(\exp(-c^2)\), and the variance is always finite. We may allow for very slowly decaying tails and possibly infinite variance by basing the models on infinite mixtures of Gaussians, using the Student-t distribution.

The probability density function of a univariate Student-t distribution, denoted as \(S(\cdot)\), is defined as:

\[
p(x \mid \mu, \lambda, d) = S(x \mid \mu, \lambda, d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \left( \frac{\lambda}{\pi d} \right)^{1/2} \left( 1 + \frac{\lambda}{d}(x - \mu)^2 \right)^{-(d+1)/2} \tag{2.46}
\]

where \(\mu\) is the mean, \(\lambda\) is known as the precision, though it is not the inverse of the variance, and \(d\) as the degrees of freedom. \(\Gamma(\cdot)\) denotes the Gamma function. As \(d \to \infty\) the Student-t distribution becomes a Gaussian. As \(d\) decreases the tails become heavier until, when \(d \leq 2\), the variance is infinite. When \(d = 1\) the distribution is equivalent to the Cauchy density and the tails decay like \(-x^2\). Figure 2.4 shows the effects of the precision and degrees of freedom on the shape of the Student-t distribution.

As alluded to above, the Student-t distribution may also be regarded as a scale mixture of Gaussians. With \(G(\tau \mid a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau)\) denoting a Gamma distribution, the Student-t may be represented as an infinite mixture of Gaussians sharing a common mean and with variances scaled by a Gamma distribution:

\[
S(x \mid \mu, \lambda, d) = \int_0^\infty \mathcal{N}(x \mid \mu, (\lambda z)^{-1}) G(z \mid d/2, d/2) \, dz. \tag{2.47}
\]

It thus generalises the finite mixture of Gaussians (with a common mean) sometimes used

\[\text{See appendix A for the main properties of the standard probability distributions referred to in this thesis and some of their key statistics.}\]
2. Background

Figure 2.4. Graph showing effects of the precision, $\lambda$, and degrees of freedom, $d$, on the shape of the Student-t distribution. The black curve, when $d$ is very large, corresponds to a Gaussian distribution.

Figure 2.5. Demonstration of the effect on the maximum likelihood solution for Gaussian and Student-t distributions of a small number of outliers. Plot (a) shows a histogram of 30 samples taken from a Gaussian distribution with mean zero and standard deviation of 0.5 (the same data as shown in figure 2.3). Overlaid in green is the Gaussian maximum likelihood estimate of the distribution. Overlaid in red is the Student-t maximum likelihood estimate of the distribution. Plot (b) shows the same, but with the addition of three outliers. Notice how both the mean and the variance have been adversely affected in the Gaussian case, but the Student-t distribution is largely unaffected.

to model outliers.

Repeating the outlier experiment for the maximum likelihood solution of a Student-t distribution (which may be determined using an EM algorithm; see appendix B), it may be seen from figure 2.5 that it is more robust than the Gaussian in that it is much less affected by the addition of outliers, because the heavy tails are able to accommodate observations far from the mean.

Lange et al. [1989] replace Gaussian assumptions with the Student-t for robust estimation of a variety of real datasets. They show that the parameterisation of the Student-t distribution may be used to control the down-weighting of outliers to give more robust statistical inference in a number of different models, including linear and non-linear regression. Not surprisingly they comment that this method is not well suited to modelling platykurtic (shorter-tailed than Gaussian) or asymmetric error distributions, for which the
Gaussian is also not appropriate. They also suggest that it is not well suited to modelling multivariate data where each of the variables have different degrees of freedom (though they note that this may be mitigated by modelling the different variables using differently parametrised Student-t distributions) or data which contains extreme outliers. In the models described in this thesis, some multivariate variables are modelled using different Student-t distributions for each variable and it will be shown, through illustrations using synthetic data, that schemes based on the Student-t are able to model very heavy-tailed data (with degrees of freedom less than 2, which have, in effect, infinite variance and so observations that might be classed as extreme outliers) effectively, though the real datasets to which these models are applied are not so heavy-tailed. In this instance Lange et al.'s may be related to models which do not learn the degrees of freedom parameter.

The Student-t method for accommodating outliers has been chosen for the models in this thesis because the representation of the distribution as a Gamma-scaled mixture of Gaussians is computationally attractive. In an exact Bayesian scheme it leads to intractable integrals, but it is easily incorporated into the variational Bayesian approximation scheme.

2.5. Assessing model quality

Having constructed models to represent multivariate time-series data and calculated approximate posterior distributions for each of the parameters, there needs to be some method for determining how well they capture the underlying structure of the data. In this thesis two principal methods for evaluating models are used: estimation of known parameters in synthetic datasets and imputation of missing data.

Synthetic datasets which have been constructed to conform to the model being tested, and for which the model order(s) and the value of each of the model parameters is known, allow direct comparisons of the estimated posterior distributions with the actual parameter values. A good posterior distribution should have an expected value close to the actual and a small variance. The form of the latent variable models means that the mixing matrix \( W \) can only be determined up to a rotation of the actual; this is discussed further in section 2.5.1.

Varying the precision of the additive noise terms allows an analysis of the robustness of the model to varying signal to noise ratios. Duplicating a synthetic dataset and adding outliers to the copy allows analysis of the robustness of the model to the presence of outliers.

A further measure of the ability of the model to capture the underlying structure of data is through the imputation of missing values, where the actual missing values are known. To evaluate a model in this way, a certain proportion of the synthetic data values is declared as missing. The model is trained and then used to reconstruct the complete observations, including the missing values. This process is described more fully in section 2.5.2. The error between the missing elements of the reconstructed observations and the original observations is an indication of its ability to impute missing values. Overfitting has previously been mentioned; measuring the error between the non-missing elements of
2. Background

(a) 
(b) 
(c) 
(d) 

Figure 2.6. Hinton diagrams showing examples of Procrustes transformation matrices for (a) no transformation, (b) permutation, (c) permutation and sign and (d) rotation.

the reconstructed observations and the original observations and comparing it with the missing data error shows to what extent this is a problem. Section 2.5.3 discusses how the error is measured.

2.5.1. Ambiguous transformations

As described in sections 2.2.2, 2.2.3 and 2.2.5, some of the models in this thesis, such as PPCA, exhibit certain transformational ambiguities with regards to the mixing matrix, $W$, and the latent variables, through the $Wx_n$ term of the latent variable model. Consider an arbitrary rotation matrix, $R$, such that $RR^T = I_Q$, where $I_Q$ is the $Q \times Q$ identity matrix. Clearly $Wx_n = (WR)(R^Tx_n)$ and $R^Tx_n \sim \mathcal{N}(0, I_Q)$ if $x_n \sim \mathcal{N}(0, I_Q)$. Thus any estimation of $W$ can only be determined up to some rotation of the original and, as pointed out in sections 2.2.2, 2.2.3 and 2.2.5, the likelihood for PPCA, FA and ProbCCA are invariant under this rotation. In this thesis I shall distinguish between three types of rotation, referred to as sign, permutation and rotation ambiguities. A rotation which results in some subset of the columns of $W$ being rotated into exactly the opposite direction (i.e. the vector being multiplied by the value $-1$) will be referred to as a sign ambiguity. Another distinct rotation causes the columns of $W$ to be reordered, resulting in a permutation ambiguity. Rotation ambiguity refers to any other type of rotation.

These ambiguities are likely to lead to an expected value for the estimated $W$ (denoted by $W_{est}$) which is not directly comparable with that of the actual $W$ (denoted by $W_{act}$) used to generate the data. Using the Procrustes method [Schönemann, 1966], $W_{est}$ may be transformed in the direction of $W_{act}$ by finding the matrix $R$ that minimises $\| W_{act} - RW_{est} \|$. This results in two matrices: a version of $W_{est}$ which has been transformed for sign, permutation and rotation to be as close as possible to $W_{act}$ and the matrix which caused the transformation, $R$. The latter may be used to apply appropriate transformations to the latent variables.

In order to apply the Procrustes technique, both matrices must have the same dimensionality. In the models described in this thesis $W_{est}$ may contain more columns than $W_{act}$, for example where the model order is overestimated. If this is the case then those columns of $W_{est}$ that have the highest absolute linear correlation with the columns of $W_{act}$ have been selected.

Inspection of the form of the Procrustes transformation matrix $R$ indicates which of the
ambiguities are present. Figure 2.6 illustrates example of the various possible cases. If none then the matrix will be the identity. If only permutation ambiguity exists then the transformation matrix is some permutation of identity matrix. If only directional ambiguity exists then the transformation matrix is diagonal with +1 and −1 values on the main diagonal. In both these cases each row and each column of the transformation matrix contains just one non-zero value. Rotational ambiguity results in a transformation matrix where the rows and columns do not contain just one non-zero value.

### 2.5.2. Missing data estimation for PCA-like models

One measure of the ability of the model to capture the underlying structure of data is through the imputation of missing values, where the actual missing values are known. A synthetic dataset is generated and a certain proportion of its values is declared as missing. The model is learned and then used to reconstruct the complete observations, including the missing values. The error, as a function of the difference between the imputed and actual values, may be used to measure the efficacy of the model.

There are many different methods for imputing missing values using PCA-like models (see, for example, [Jolliffe, 2002] and [Little and Rubin, 1987]). The method selected for this thesis, and applied to models based on (P)PCA, is an iterative algorithm similar to that demonstrated by Everson and Sirovich [1995]. To start with the missing data values are set to the temporal mean, calculated as

\[
\bar{t} = \frac{1}{N} \sum_{n=1}^{N} t_n \tag{2.48}
\]

resulting in a completed dataset, \(\hat{T}\). The process then iterates through the following steps until convergence:

(a) Perform PCA on the completed data, \(\hat{T}\), to get estimates for \(W\) and \(X\).

(b) Calculate \(\bar{t}\) as the mean observation, as per (2.48).

(c) Calculate the reconstruction of each observation as \(\hat{t}_n = Wx_n + \bar{t}\).

(d) Replace the missing values in \(T\) with the corresponding values in \(\hat{T}\).

(e) Set \(\hat{T} = T\).

This method represents a sort of poor man’s EM algorithm; missing values are replaced by the current estimates and then the parameters are recalculated (if the missing value estimates were expectations then this would be a proper EM algorithm). The model order, \(Q\), needs to be carefully selected to prevent overfitting to the observed data at the expense of poor generalisation to the missing values. This method includes the current estimates for the missing data in the calculation of the latent variables in step 1.

A similar algorithm, more like a proper EM algorithm, may be used for filling in missing
values using PPCA. To start with \( W \) and \( \lambda \) are initialised with random values and the missing data values are set to the temporal mean (2.48), as before. The process then iterates through the following steps until convergence:

(a) Calculate the expected values \( \langle x_n|t_n \rangle \) and \( \langle x_n x_n^T|t_n \rangle \) from (2.31)-(2.33):

\[
M = W^T W + \lambda^{-1} I \tag{2.49}
\]

\[
\langle x_n \rangle = M^{-1} W^T (t_n - \bar{t}) \tag{2.50}
\]

\[
\langle x_n x_n^T \rangle = \lambda^{-1} M^{-1} + \langle x_n \rangle \langle x_n \rangle^T \tag{2.51}
\]

using the current estimates for \( W \) and \( \lambda \).

(b) Calculate new estimates for \( W \) and \( \lambda \) from (2.34)-(2.35):

\[
W = \left[ \sum_{n=1}^{N} (t_n - \bar{t}) \langle x_n \rangle^T \right] \left[ \sum_{n=1}^{N} \langle x_n x_n^T \rangle \right]^{-1} \tag{2.52}
\]

\[
\lambda^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \left[ \| t_n - \bar{t} \|^2 - 2 \langle x_n \rangle^T W^T (t_n - \bar{t}) + \text{trace}(W \langle x_n x_n^T \rangle W^T) \right] \tag{2.53}
\]

(c) Calculate the reconstruction, \( \hat{T} \), of \( T \) from (2.38):

\[
\hat{t}_n = W(W^T W)^{-1} M \langle x_n \rangle \tag{2.54}
\]

(d) Replace the missing values in \( T \) with the corresponding values in \( \hat{T} \).

(e) Set \( \hat{T} = T \).

This method will be referred to as PPCA-t. Like the poor man’s EM algorithm described at the beginning of this section, it includes the current estimates for the missing data in the calculation of the latent variables. However, the only real evidence we have for the model are those values in \( T \) that are not missing, and it would be better for the estimation of the latent variables in \( X \) to be based only on them. To achieve this, each observation, \( t_n \), is considered to be split into an observed part, \( g_n \), and a missing part, \( h_n \), with, for each observation, \( W \) being similarly split:

\[
t_n = \begin{pmatrix} g_n \\ h_n \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} W_{g_n} \\ W_{h_n} \end{pmatrix} \tag{2.55}
\]

In the PPCA Expectation step, (2.49)-(2.51), \( t_n \) is replaced by \( g_n \), and \( \langle x_n|t_n \rangle \) and \( \langle x_n x_n^T|t_n \rangle \) are replaced by \( \langle x_n|g_n \rangle \) and \( \langle x_n x_n^T|g_n \rangle \) respectively. The Maximisation step remains unchanged. This method will be referred to as PPCA-g.

Note that while the estimation of \( X \) in PPCA-t is a single calculation, for PPCA-g each vector in \( X \) must be separately calculated, which makes the process notably slower to execute.
2. Background

From (2.30) the covariance of $t_n$ is given by

$$WW^T + \lambda^{-1}I = \begin{pmatrix} W_{gn}W_{gn}^T + \lambda^{-1}I & W_{gn}W_{hn}^T \\ W_{hn}^T & W_{hn}W_{hn}^T + \lambda^{-1}I \end{pmatrix}. \tag{2.56}$$

Hence (e.g. [Petersen and Pedersen, 2008, section 8.1.3]) the covariance of missing data, $h_n$, given the observed data, $g_n$, is

$$C_{h_n} = (W_{hn}W_{hn}^T + \lambda^{-1}I) - W_{hn}W_{gn}^T(W_{gn}W_{gn}^T + \lambda^{-1}I)^{-1}W_{gn}W_{hn}^T. \tag{2.57}$$

The variance for the $i$th missing data value in $h_n$ is given by the $ii$th element of the covariance matrix $C_{h_n}$. The calculation of the variances does not need to be performed for each iteration, just once when the model has been learned.

2.5.3. Error measures

Some measure is required to assess and compare the performance of the models in the imputation of missing values. In the Gaussian observation noise case the mean squared error between the reconstructions and the actuals, calculated as

$$E_2 = \frac{1}{N} \sum_{n=1}^{N} |t_n - \hat{t}_n|^2, \tag{2.58}$$

is a natural choice as it is derived from the Gaussian error function (the negative log likelihood).

Where data are heavier-tailed than Gaussian the mean squared error gives undue weight to values in the tails of the distribution; here the mean absolute error

$$E_1 = \frac{1}{N} \sum_{n=1}^{N} |t_n - \hat{t}_n|, \tag{2.59}$$

which is derived from the error function for the heavier-tailed Laplacian distribution, may be more appropriate.

2.6. Conclusions

Chapter 1 introduces three real sets of multivariate, time-series data whose underlying structure is of interest from the point of view of regression analysis. The ONS data is relatively low-dimensional (23 dimensions, 494 observations) and complete, with some obvious outliers. The NASA data consists of high-dimensional satellite images (113,593 dimensions, 92 observations) where a significant proportion of the data is missing due to cloud cover. While the ONS and NASA datasets each contain a single set of observations, the EEG data consists of multiple sets of data which are assumed to share some underlying signal and which are strongly leptokurtic.
The current chapter starts with a brief summary of Bayesian probability theory as it applies to the parametrised models described in this thesis. Fully Bayesian models result in posterior probability distributions for each parameter variable, with the advantage that this provides a measure of certainty in the estimates, but the integrals required for exact Bayesian inference are often intractable. Section 2.1.2 describes a number of approximation techniques from which the factorised variational method has been selected.

Section 2.2 describes a number of related models, (P)PCA, FA and (Prob)CCA, which assume that the data were generated by a noisy linear mixing of a lower-dimensional set of latent variables. One difficulty with these models is how to determine the dimensionality of the latent variables, i.e. the model order. A number of schemes are described in section 2.3; one which fits in well with the Bayesian scheme is automatic relevance determination (ARD). This places a prior over the magnitude of each mixing component, constraining the magnitude to be zero where there is no evidence in the data for it. Thus the model order is automatically estimated as part of the Bayesian inference procedure.

The probabilistic models thus far described (PPCA, FA and ProbCCA) all make Gaussian assumptions regarding their parameter variables. Section 2.4 demonstrates how the maximum likelihood estimation of a Gaussian distribution is considerably disturbed by the presence of outliers in the data. It also demonstrates how the much heavier tails of the Student-t distribution are able to accommodate the outliers, leading to significantly less disturbance. Thus replacing Gaussian assumptions with the Student-t in the latent variable models provides greater robustness to outliers, with the added benefit of enabling them to model leptokurtic data.

Finally, section 2.5 describes methods for assessing the quality of new models. The two principal approaches used here are estimation of known parameters in synthetic datasets and the imputation of missing data. For the PCA-like models the variables in the latent variable mixing term is only able to be estimated to within a rotation of the originals; section 2.5.1 describes how they are compared using Procrustes transformations.

(P)PCA, FA and (Prob)CCA model spatial information, but they ignore one important aspect of the data that we are trying to capture: their temporal nature. Reordering the observations makes no difference to these models other than to reorder the latent variables. For the related Independent Component Analysis model, Pearlmutter and Parra [1997] propose a context-sensitive generalisation to capture temporal structure in the sources. Autoregressive (AR) processes with non-Gaussian additive noise (they use a logistic noise distribution) are used to model each source. Most of the new models introduced in this thesis are based on PPCA and ProbCCA. In particular, two new models are proposed that aim to capture both spatial and temporal structure by applying AR to the latent variables of PPCA and ProbCCA. They are defined as Bayesian models so that posterior probability distributions are obtained for each parameter variable, giving some measure of the uncertainty in the estimation of their values given the data from which they were calculated. Exact Bayesian inference is intractable, so a variational approximation technique is used. The new models are made robust to outliers and able to model leptokurtic data through the assumption of the noise being Student-t distributed. The Bayesian approach
supports the incorporation of ARD to estimate the models’ orders, so that a parsimonious explanation of the data is achieved.

The thesis introduces two new models based on PPCA. Similar models based on FA are described because their derivation from the PPCA-based models is straightforward, but there is no discussion of them and no experimental results are included for them.

Figure 2.7 summarises all the models described in this thesis and their interrelationships. Each node represents a named model, with the assumption regarding the distribution of the noise (either Gaussian or Student-t) shown in small italics underneath. A “v” prefix to the model name denotes a variational Bayesian model; the others are either algebraic or maximum likelihood solutions. Shaded nodes represent the novel models that will be described in the following chapters. The section in which each model is primarily discussed are shown as a small red number to the right of the node.

This chapter has described PCA and PPCA; chapter 3 starts by describing AR and then introduces PPCA-AR, a maximum likelihood formulation of PPCA incorporating AR processes over the latent variables. In order to concentrate on the basis of this model, the noise is assumed Gaussian distributed, the observations are assumed to have a zero mean and the model order is fixed. The journey towards the variational version of PPCA-AR starts in chapter 4 where variation PPCA is described and a new robust, variational PPCA model is introduced (and a similar model for FA derived) with ARD for automatic model order estimation. Chapter 5 introduces a new robust, variational version of AR, also including ARD for model order estimation. The work from these three chapters is then brought together in chapter 6, where a robust, variational version of PPCA-AR, with ARD and support for a non-zero mean observation, is described. Chapter 7 extends this model to create robust, variational ProbCCA-AR, with ARD, which uses the same methods for ProbCCA as were previously applied to PPCA.
Figure 2.7. A summary of models described in this thesis. Each node represents a named model, with the assumption regarding the distribution of the noise (either Gaussian or Student-t) shown in small italics underneath. A “v” prefix to the model name denotes a variational Bayesian model; the others are either algebraic or maximum likelihood solutions. Shaded nodes represent the novel models that will be described in the following chapters. The section in which each model is primarily discussed are shown as a small red number to the right of the node.
3. Maximum likelihood PPCA with autoregressive latent variables

Probabilistic Principal Component Analysis (PPCA) decomposes a multivariate observation into a weighted sum of orthogonal principal components. This has previously been expressed as

\[ t_n = Wx_n + \bar{t} + e_n \]  

(3.1)

where \( t_n \) is the \( D \)-dimensional vector representing the \( n \)th observation, \( x_n \) the associated \( Q \)-dimensional vector of latent variables is assumed to be distributed as \( \mathcal{N}(x_n \mid 0, I) \), \( W \) is the mixing matrix, and \( e_n \) is noise which is assumed to be \( \mathcal{N}(e_n \mid 0, \lambda^{-1}I) \). For simplification, in this chapter it will be assumed, without loss of generality, that the mean observation \( \bar{t} \) is 0. The decomposition may be rewritten as

\[ t_n = \sum_q x_{q,n}w_q + e_n \]  

(3.2)

where the \( w_q \) are the principal components and \( x_{q,n} \) are the corresponding zero mean, unit variance, Gaussian-distributed latent variables. Where the observations are temporally ordered, the PPCA decomposition results in spatial variables, the principal components \( w_q \), and associated temporal latent variables, \( x_{q,n} \).

Despite this apparent spatio-temporal decomposition, there is in fact no temporal coupling built into PPCA, or indeed into PCA or Factor Analysis (FA); the only effect of reordering the observations is to reorder the latent variables. In this chapter PPCA is augmented with a temporal model in the latent space, by coupling the latent variables in time with an autoregressive (AR) model (e.g. Chatfield [2004]) with coefficients \( \theta_{q,p} \):

\[ x_{q,n} = \sum_{p=1}^{P} \theta_{q,p}x_{q,n-p} + \epsilon_{q,n} \]  

(3.3)

where the excitation noise, or innovations sequence, \( \epsilon_{q,n} \), is assumed Gaussian distributed, \( \mathcal{N}(\epsilon_{q,n} \mid 0, \kappa^{-1}) \). The Gaussian excitation noise means that, as for the standard PPCA model, each latent variable is Gaussian distributed, so this new model may be viewed as a generalisation of PPCA.

The new model, referred to as PPCA-AR, is constructed as a linear dynamical system. The inference is by maximum likelihood and the expectation maximisation (EM) algorithm is used, incorporating a forward-backward algorithm to estimate the distribution of the

\(^*\)Some of the material in this chapter has been published as [Christmas and Everson, 2010].
latent variables in the E-step. It will be shown that, for synthetic data, it is able to make
good estimates of the actual parameter values, and that the additional temporal model
is advantageous when used to impute missing values for both synthetic and real data
compared with two non-temporal PPCA methods.

3.1. Autoregression

With the latent variables collected into a single $Q \times N$ matrix $X = (x_1, \ldots, x_N)$, let us
consider first the AR model for a single row of $X$, which is denoted by $x$. AR models each
of the $N$ values in $x$ with the following linear expression of order $P$:

$$x_n = \sum_{p=1}^{P} \theta_p x_{n-p} + \epsilon_n$$  \hspace{1cm} (3.4)

where $\theta = (\theta_1, \ldots, \theta_P)^T$ is a vector of scalar coefficients that are specific to this $x$, $\epsilon_n$ is the
Gaussian excitation noise associated with this $x_n$ and $n > P$. It is possible to transform
this order $P$ model into a first order Markov model by defining $\tilde{x}_n$ and $\tilde{\epsilon}_n$ as the lag vectors

$$\tilde{x}_n = \begin{pmatrix} x_{n-P+1} \\ x_{n-P+2} \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} \text{ and } \tilde{\epsilon}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \epsilon_n \end{pmatrix}$$  \hspace{1cm} (3.5)

so that

$$\tilde{x}_n = \Theta \tilde{x}_{n-1} + \tilde{\epsilon}_n$$  \hspace{1cm} (3.6)

with $\Theta$ as the $P \times P$ matrix

$$\Theta = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \cdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \theta_P & \theta_{P-1} & \theta_2 & \theta_1 \end{pmatrix}$$  \hspace{1cm} (3.7)

$\epsilon_n$ is Gaussian distributed with mean 0 and covariance given by the $P \times P$ matrix

$$\Phi = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \cdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \kappa^{-1} \end{pmatrix}$$  \hspace{1cm} (3.8)
The first \( P - 1 \) elements of the main diagonal of \( \Phi \) are expected to be zero because the first \( P \) elements of \( x_n \) are exactly determined by \( \Theta \tilde{x}_{n-1} \) and they therefore have no variance.

This first order model results in a Gaussian probability distribution for \( \tilde{x}_n \) conditioned on \( \tilde{x}_{n-1} \) as follows

\[
p(\tilde{x}_n | \tilde{x}_{n-1}) = N(\tilde{x}_n | \Theta \tilde{x}_{n-1}, \Phi) \tag{3.9}
\]

By stacking the \( \tilde{x}_n \) and \( \tilde{\epsilon}_n \) for each row of \( X \), and constructing composite \( \Theta \) and \( \Phi \) as block-diagonal matrices of each row’s individual \( \Theta \) and \( \Phi \) matrices respectively, the AR expressions for all rows of \( X \) may be written in the form of (3.6), giving a conditional distribution for the latent variables in the form of (3.9).

The use of AR to independently model each row of the latent variables makes no requirement that the model order should be the same in each case; each row of \( X \) is able to have different values of \( P \).

Note that the \( N(0, I) \) constraint placed on the latent variables in PPCA has been discarded and the scale of the latent variables is determined by the variance of the innovations sequence.

### 3.2. PPCA-AR: a linear dynamical system

We may combine (3.6) with a straightforward adjustment to the PPCA latent variable equation (3.1) to give the two expressions that define PPCA-AR as a linear dynamical system:

\[
\tilde{x}_n = \Theta \tilde{x}_{n-1} + \tilde{\epsilon}_n \tag{3.10}
\]
\[
t_n = \tilde{W} \tilde{x}_n + e_n \tag{3.11}
\]

where \( \tilde{W} = WF \) and \( F \) is the matrix of ones and zeros that extracts \( x_n \) from \( \tilde{x}_n \). From (3.9) and an appropriately adjusted (3.1) we get the Gaussian transition and observation probabilities

\[
p(\tilde{x}_n | \tilde{x}_{n-1}) = N(\tilde{x}_n | \Theta \tilde{x}_{n-1}, \Phi) \tag{3.12}
\]
\[
p(t_n | \tilde{x}_n) = N(t_n | \tilde{W} \tilde{x}_n, \lambda^{-1} I). \tag{3.13}
\]

Figure 3.1 is a graphical representation of the PPCA-AR model, which makes clear the recursive dependency between the probabilities for each \( \tilde{x}_n \) and the similarity with the Kalman filter [Kalman and Bucy, 1961], which is also a linear Gaussian state space model. In order to complete the definition an assumption must be made regarding the distribution of the first latent variable:

\[
p(\tilde{x}_1) = N(\tilde{x}_1 | m_0, V_0) \tag{3.14}
\]

which corresponds to a prior specification of the \( P \) “observations” immediately before the
3. Maximum likelihood PPCA with autoregressive latent variables

Figure 3.1. A graphical representation of the PPCA-AR model which makes clear its definition as a linear dynamical system and the recursive dependencies between latent variables.

actual observations.

As a linear dynamical system we may apply the EM algorithm defined by Ghahramani and Hinton [1996] to estimate maximum likelihood values for the parameters of this model, \( \Omega = \{ W, \lambda, \Theta, \Phi, m_0, V_0 \} \), together with the latent variables, \( X \). This makes use of the forward-backward algorithm, which is described first.

Forward-backward algorithm

In the expectation step of the EM algorithm the probability distribution for the latent variables conditioned on the observations is calculated. In cases where each instance of the latent variables is statistically independent (and Gaussian), as it is in PPCA, this calculation is straightforward. In this case though each instance is dependent on the previous instance in the ordering, as is shown in figure 3.1, which means that there is some dependence between each instance and every other instance. The forward-backward algorithm [Rabiner, 1989], also known as the Baum-Welch algorithm [Baum, 1972], is an efficient method for inferring the posterior distribution of each \( \tilde{x}_n \). Described here is the most widely-known variant, the alpha-beta algorithm (for a tutorial see [Bishop, 2006, chapter 13]).

If \( \hat{\alpha}(\tilde{x}_n) = p(\tilde{x}_n | t_1, \ldots, t_n) \), that is the probability of the current instance of the latent variables given the current observation and all previous observations, and \( \hat{\beta}(\tilde{x}_n) = p(\tilde{x}_n | t_{n+1}, \ldots, t_N) \), the probability of the current instance of the latent variables given all future observations, then the posterior probability \( p(\tilde{x}_n | T) \) is the product of these two terms: \( p(\tilde{x}_n | T) = \hat{\alpha}(\tilde{x}_n) \hat{\beta}(\tilde{x}_n) \). These \( \hat{\alpha} \) and \( \hat{\beta} \) probabilities are defined recursively:

\[
\hat{\alpha}(\tilde{x}_n) = p(t_n | \tilde{x}_n) \int \hat{\alpha}(\tilde{x}_{n-1}) p(\tilde{x}_n | \tilde{x}_{n-1}) d\tilde{x}_{n-1} \tag{3.15}
\]

\[
\hat{\beta}(\tilde{x}_n) = \int \hat{\beta}(\tilde{x}_{n+1}) p(t_{n+1} | \tilde{x}_{n+1}) p(\tilde{x}_{n+1} | \tilde{x}_n) d\tilde{x}_{n+1} \tag{3.16}
\]

Figure 3.2 shows a graphical representation of these probabilities.

Also required for the EM algorithm is the joint distribution \( p(\tilde{x}_{n-1}, \tilde{x}_n | T) \), which is proportional to \( \hat{\alpha}(\tilde{x}_{n-1}) p(t_n | \tilde{x}_n) p(\tilde{x}_n | \tilde{x}_{n-1}) \hat{\beta}(\tilde{x}_n) \).

The algorithm is performed in two steps. In the forward sweep, known as the Kalman filter
\[ \hat{\alpha}(\tilde{x}_n) \propto p(\tilde{x}_n | t_1, \ldots, t_n) \]
\[ = p(t_n | \tilde{x}_n) \int \hat{\alpha}(\tilde{x}_{n-1}) p(\tilde{x}_n | \tilde{x}_{n-1}) d\tilde{x}_{n-1} \]

\[ \hat{\beta}(\tilde{x}_n) \propto p(\tilde{x}_n | t_{n+1}, \ldots, t_N) \]
\[ = \int \hat{\beta}(\tilde{x}_{n+1}) p(t_{n+1} | \tilde{x}_{n+1}) p(\tilde{x}_{n+1} | \tilde{x}_n) d\tilde{x}_{n+1} \]

Figure 3.2. A graphical representation of the forward-backward algorithm.

[Kalman, 1960; Zarchan and Musoff, 2005], \( \hat{\alpha}(\tilde{x}_n) \) is calculated for each \( n \) from 1 to \( N \) (the number of observations). The backward sweep, known as the Kalman smoother, or Rauch-Tung-Striebel equations [Rauch et al., 1965], takes the results of the forward sweep and combines them with the calculation for \( \hat{\beta}(\tilde{x}_n) \) for each \( n \) from \( N \) back down to 1.

**EM algorithm: E-step**

Following the forward-backward algorithm the probability distributions for the latent variables are calculated in two steps. After the forward sweep we arrive at \( p(\tilde{x}_n | t_1, \ldots, t_n) = \mathcal{N}(\tilde{x}_n | \mu_1, \Sigma_1) \) for each \( n \), where

\[ P_0 = \Sigma_0 \]
\[ \mu_1 = \mu_0 + K_1(t_1 - \tilde{W}\mu_0) \]

and

\[ P_{n-1} = \Theta \Sigma_{n-1} \Theta^T + \Phi \]
\[ K_n = P_{n-1} W^T (\tilde{W}P_{n-1}W^T + \lambda^{-1}I)^{-1} \]
\[ \mu_n = \Theta \mu_{n-1} + K_n(t_n - \tilde{W}\Theta\mu_{n-1}) \]
\[ V_n = (I - K_n\tilde{W})P_{n-1}. \]

where \( K_n \) is known as the Kalman gain matrix.

If \( T \) is the matrix, \( (t_1, \ldots, t_N) \), of ordered observations, then after both the forward and backward sweeps have been completed we obtain the posterior distribution for each latent variable \( p(\tilde{x}_n | T) = \mathcal{N}(\tilde{x}_n | \tilde{\mu}_N, \tilde{\Sigma}_N) \) where

\[ \tilde{\mu}_N = \mu_N \]
\[ \tilde{\Sigma}_N = \Sigma_N \]
and
\[ J_n = \Sigma_n \Theta^T P_n^{-1} \]  \hspace{1cm} (3.25)
\[ \hat{\mu}_n = \mu_n + J_n(\hat{\mu}_{n+1} - \Theta \mu_n) \]  \hspace{1cm} (3.26)
\[ \hat{\Sigma}_n = \Sigma_n + J_n(\hat{\Sigma}_{n+1} - P_n)J_n^T. \]  \hspace{1cm} (3.27)

Inspecting (3.21) and (3.26), it may be seen that the expectation of \( \tilde{x}_n \) is made up of three elements: the prediction from the previous time-step (\( \Theta \mu_{n-1} \)), an adjustment proportional to the predicted error of the reconstruction of the current observation (\( t_n - \tilde{W}\Theta \mu_{n-1} \)) and an adjustment proportional to the predicted error of the next time-step (\( \tilde{\mu}_{n+1} - \Theta \mu_n \)). The adjustments are weighted according to the relative magnitudes of the predicted precisions, so that more weight is given to the element which has greater certainty.

The results of this E-step are the following expectations:
\[ \langle \tilde{x}_n | T \rangle = \tilde{\mu}_n \]  \hspace{1cm} (3.28)
\[ \langle \tilde{x}_n \tilde{x}_n^T | T \rangle = J_{n-1} \hat{\Sigma}_n + \hat{\mu}_n \hat{\mu}_n^T \]  \hspace{1cm} (3.29)
\[ \langle \tilde{x}_n \tilde{x}_n^T | T \rangle = \hat{\Sigma}_n + \hat{\mu}_n \hat{\mu}_n^T. \]  \hspace{1cm} (3.30)

**EM algorithm: M-step**

From (3.12), (3.13) and (3.14) the complete data log likelihood to be maximised is
\[ p(T, X | \Omega) = -\frac{1}{2} \log |\Sigma_0| - \frac{1}{2}(\tilde{x}_1 - \mu_0)^T \Sigma_0^{-1}(\tilde{x}_1 - \mu_0) \]
\[ -\frac{N-1}{2} \log |\Phi| - \frac{N}{2} \log |\lambda^{-1}I| \]
\[ -\frac{1}{2} \sum_{n=2}^{N} (\tilde{x}_n - \Theta \tilde{x}_{n-1})^T \Phi^{-1} (\tilde{x}_n - \Theta \tilde{x}_{n-1}) \]
\[-\frac{1}{2} \sum_{n=1}^{N} \lambda (t_n - \tilde{W}\tilde{x}_n)^T (t_n - \tilde{W}\tilde{x}_n). \]  \hspace{1cm} (3.31)

By the usual method of differentiation with respect to each of the model parameters in \( \Omega \) we obtain update expressions as follows (the expectation dependencies on \( T \) have been omitted to aid readability):

\[ \mu_0 = \langle \tilde{x}_1 \rangle \]  \hspace{1cm} (3.32)
\[ \Sigma_0 = \langle \tilde{x}_1 \tilde{x}_1^T \rangle + \langle \tilde{x}_1 \rangle \langle \tilde{x}_1 \rangle^T \]  \hspace{1cm} (3.33)
\[ \Theta = \left[ \sum_{n=2}^{N} \langle \tilde{x}_n \tilde{x}_{n-1}^T \rangle \right] \left[ \sum_{n=2}^{N} \langle \tilde{x}_{n-1} \tilde{x}_{n-1}^T \rangle \right]^{-1} \]  \hspace{1cm} (3.34)
\[ \Phi = \frac{1}{N-1} \sum_{n=2}^{N} \left[ \langle \tilde{x}_n \tilde{x}_n^T \rangle - \Theta \langle \tilde{x}_{n-1} \tilde{x}_{n-1}^T \rangle - \langle \tilde{x}_n \tilde{x}_{n-1}^T \rangle \Theta^T - \Theta \langle \tilde{x}_{n-1} \tilde{x}_{n-1}^T \rangle \Theta^T \right] \]  \hspace{1cm} (3.35)
\[ \tilde{W} = \left[ \sum_{n=1}^{N} \langle t_n \tilde{x}_n \rangle \tilde{F}^T \right] \left[ \sum_{n=1}^{N} \tilde{F}^T \langle \tilde{x}_n \tilde{x}_n^T \rangle \tilde{F} \right]^{-1} \]  \hspace{1cm} (3.36)
\[ \lambda^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \text{trace} \left( t_n t_n^T - \tilde{W}(\tilde{x}_n) t_n^T - t_n (\tilde{x}_n)^T \tilde{W}^T + \tilde{W}(\tilde{x}_n \tilde{x}_n^T) \tilde{W}^T \right). \] 

(3.37)

with \text{trace}(\cdot) denoting the trace operator. As might be expected, the expressions for the expectations in (3.28)-(3.30) differ considerably from those for PPCA in (2.31)-(2.33), but the definitions for \( W \) and \( \lambda^{-1} \) ((3.36)-(3.37) and (2.34)-(2.35)) show a close similarity.

However, the expressions for \( \Theta \) and \( \Phi \) do not respect the constraints imposed by the required structures of these matrices (see (3.7) and (3.8)). By considering each of the rows of the latent variables in turn it may be shown that these constraints can be respected. Rather than introducing a new subscript, the matrices and vectors in the following explanation pertain only to a single row of \( X \).

The matrix \( \Theta \) may be written as a function of the vector \( \theta \) using the Kronecker product:

\[ \Theta = Z(\theta \otimes I_P) + Y \] 

(3.38)

where \( Z \) and \( Y \) are matrices containing only ones and zeros and \( I_P \) is the \( P \times P \) identity matrix. The log likelihood (3.31) may then be rewritten as a function of \( \theta \) rather than \( \Theta \) and the differentiation performed accordingly. The result is

\[ \theta = 2(\mathbf{H}\mathbf{H}^T)^{-1} \left[ \sum_{n=2}^{N} (\tilde{x}_n^{T} - \tilde{x}_{n-1}^{T} \mathbf{Y}^{T}) \Phi^{-1}Z(I_p \otimes \tilde{x}_{n-1}) \right]^{T} \] 

(3.39)

where \( \mathbf{H} = \sum_{n=2}^{N} (I_p \otimes \tilde{x}_{n-1})^{T} \Phi^{-1}Z(I_p \otimes \tilde{x}_{n-1}) \). Substituting \( \theta \) into expression (3.38) generates \( \Theta \) in the correct form.

For the excitation noise variance the log likelihood for one row of the latent variables may be rewritten as a function of \( \kappa^{-1} \) rather than \( \Phi \); differentiation with respect to \( \kappa^{-1} \) yields

\[ \kappa^{-1} = \frac{1}{N-1} \sum_{n=2}^{N} \left( \langle x_n^2 \rangle - 2 \langle x_n \tilde{x}_{n-1}^T \rangle \theta + \theta^T \langle \tilde{x}_{n-1} \tilde{x}_{n-1}^T \rangle \theta \right) \] 

(3.40)

which may then be inserted into \( \Phi \).

In the observation model, \( t_n = Wx_n + e_n \), it is easy to see that an increase in the scale of \( W \) may be offset by a corresponding decrease in scale of \( x_n \) without affecting the overall result. PCA fixes the relationship between \( W \) and \( x_n \) by making the vectors in \( W \) orthonormal, so that \( W^T W = I \). PPCA fixes the scale of the \( x_n \) by specifying a prior for it of \( \mathcal{N}(\mathbf{0}, \mathbf{I}) \). As a result the components in \( W \) are no longer orthonormal; they are still orthogonal as, from (2.36), \( W^T W = P_q - \lambda^{-1} I \) is a diagonal matrix, but it is unlikely to be the identity. PPCA may be rewritten, without loss of generality, to enforce orthonormality on the \( W \) and remove the \( \mathcal{N}(\mathbf{0}, \mathbf{I}) \) constraint on \( x_n \) as follows. Taking \( A \) as the diagonal matrix \( (P_q - \lambda^{-1} I)^{1/2} \), the reconstruction of observation \( t_n \) is, from (2.38)

\[ \hat{t}_n = U_q(A + (\lambda A)^{-1})^{-1} \langle x_n \rangle \] 

(3.41)

We may define \( W = U_q \) (now orthonormal) and combine the diagonal matrix \( A + (\lambda A)^{-1} \).
with the $\langle x_n \rangle$ term, giving a model where the scale of $W$ is fixed and scale information resides in the $x_n$.

In order to constrain $W$ to be orthonormal, such that $W^T W = I$, Lagrange multipliers were included in the maximisation of the complete data log likelihood with respect to $W$, giving a new update statement for $W$ (replacing (3.36)):

$$W = B (B^T B)^{-\frac{1}{2}}$$  \[(3.42)\]

where $B = \sum_{n=1}^{N} t_n \langle \tilde{x}_n \rangle^T F^T$.

The overall algorithm is summarised on page 63.

The algorithm is initialised with the results of PPCA, giving starting points for $W$, $X$ and $\lambda$, and then an AR model is learned for each row of $X$ to prime the $\theta_q$ and $\kappa_q$. It is deemed to have converged when the complete data log likelihood (3.31) has converged.

So far the problem of selecting the model orders, $P$ and $Q$, has not been addressed. One criterion might be to use maximum description length (see section 2.3 on page 41 for a discussion of other possible methods). However, in chapter 6 automatic relevance determination is used to automatically estimate values for $P$ and $Q$ based on evidence in the data.

**Imputing missing values**

Tipping and Bishop [1999] have shown that PPCA gives the optimal (among linear mixing models) reconstruction of the observations in the mean squared error sense, so it would not be surprising if PPCA-AR, which enforces additional constraints, performs less well in this respect. However, these additional constraints provide more information about the underlying nature of the observations and prove advantageous when the model is used to impute missing data.

With PPCA-AR, as for PPCA-g (see section 2.5.2), where data is missing the values for each $\tilde{x}_n$ are estimated based only on those elements of $T$ which are observed. This is achieved by replacing $t_n$ with $g_n$ in the E-step (3.28)-(3.30). Once the likelihood has been maximised, the probability distribution for $t_n$ is given as

$$p(\hat{t}_n \mid T) = \mathcal{N}(t_n \mid \hat{W} \hat{\mu}_n, \hat{W} A_n \hat{W}^T + \lambda^{-1} I)$$  \[(3.43)\]

where $A_n = \tilde{\Sigma}_n + \hat{\mu}_n \hat{\mu}_n^T$. The covariance of $t_n$ is

$$\hat{W} A_n \hat{W}^T + \lambda^{-1} I = \begin{pmatrix} \hat{W} g_n A_n \hat{W}^T_{g_n} + \lambda^{-1} I & \hat{W} g_n A_n \hat{W}^T_{h_n} \\ \hat{W} h_n A_n \hat{W}^T_{g_n} & \hat{W} h_n A_n \hat{W}^T_{h_n} + \lambda^{-1} I \end{pmatrix}.$$  \[(3.44)\]

Hence [Petersen and Pedersen, 2008, section 8.1.3] the covariance of the missing data,
### Summary of the PPCA-AR EM algorithm

**E-step**

Forward sweep

\[ P_0 = \Sigma_0 \]  
(3.45)

\[ \mu_1 = \mu_0 + K_1(t_1 - \tilde{W}\mu_0) \]  
(3.46)

\[ P_{n-1} = \Theta \Sigma_{n-1} \Theta^T + \Phi \]  
(3.47)

\[ K_n = P_{n-1}\tilde{W}^T(\tilde{W}P_{n-1}\tilde{W}^T + \lambda^{-1}I)^{-1} \]  
(3.48)

\[ \mu_n = \Theta \mu_{n-1} + K_n(t_n - \tilde{W}\Theta\mu_{n-1}) \]  
(3.49)

\[ \Sigma_n = (I - K_n\tilde{W})P_{n-1}. \]  
(3.50)

Backward sweep

\[ \bar{\mu}_N = \mu_N \]  
(3.51)

\[ \bar{\Sigma}_N = \Sigma_N \]  
(3.52)

\[ J_n = \Sigma_n \Theta^T P_{n-1} \]  
(3.53)

\[ \bar{\mu}_n = \mu_n + J_n(\bar{\mu}_{n+1} - \Theta \mu_n) \]  
(3.54)

\[ \bar{\Sigma}_n = \Sigma_n + J_n(\Sigma_{n+1} - P_n)J_n^T. \]  
(3.55)

Resulting expectations

\[ \langle \bar{x}_n|T \rangle = \bar{\mu}_n \]  
(3.56)

\[ \langle \bar{x}_n, \bar{x}_n^T|T \rangle = J_{n-1} \bar{\Sigma}_n + \bar{\mu}_n \bar{\mu}_n^T \]  
(3.57)

\[ \langle \bar{x}_n, \bar{x}_n^T|T \rangle = \bar{\Sigma}_n + \bar{\mu}_n \bar{\mu}_n^T \]  
(3.58)

**M-step**

\[ \mu_0 = \langle \bar{x}_1 \rangle \]  
(3.59)

\[ \Sigma_0 = \langle \bar{x}_1 \bar{x}_1^T \rangle + \langle \bar{x}_1 \rangle \langle \bar{x}_1 \rangle^T \]  
(3.60)

\[ \theta_q = 2(HH^T)^{-1} \left[ \sum_{n=2}^N \bar{x}_{q,n}^T - \bar{x}_{q,n-1}^T Y^T \right] \Phi^{-1} Z(I_p \otimes \bar{x}_{q,n-1})^T \]  
(3.61)

where \( H = \sum_{n=2}^N (I_p \otimes \bar{x}_{n-1})^T Z^T \Phi^{-1} Z(I_p \otimes \bar{x}_{n-1}) \)

\[ \kappa_q^{-1} = \frac{1}{N-1} \sum_{n=2}^N \left( \langle x_{q,n}^2 \rangle - 2\langle x_{q,n} \bar{x}_{q,n-1}^T \rangle \theta_q + \theta_q^T \langle x_{q,n-1} \bar{x}_{q,n-1}^T \rangle \theta_q \right) \]  
(3.62)

\[ W = B (B^T B)^{-rac{1}{2}} \]  
where \( B = \sum_{n=1}^N t_n (\bar{x}_n)^T F^T \)  
(3.63)

\[ \lambda^{-1} = \frac{1}{ND} \sum_{n=1}^N \text{trace} \left[ t_n t_n^T - \tilde{W}(\bar{x}_n)t_n^T - t_n(\bar{x}_n)^T \tilde{W}^T + \tilde{W}(\bar{x}_n) \bar{x}_n^T \tilde{W}^T \right] \]  
(3.64)

Subscript \( n \) denotes the \( n \)th observation and subscript \( q \) denotes belonging to the \( q \)th latent variable. The variables’ distributions were calculated in the order shown above. In the M step the conditioning of the expectations on \( T \) is omitted for readability.
3. Maximum likelihood PPCA with autoregressive latent variables

\[
C_{h_n} = \tilde{W}_{h_n} A_n \tilde{W}_{h_n}^T + \lambda^{-1} I - \tilde{W}_{h_n} A_n \tilde{W}_{g_n}^T (\tilde{W}_{g_n} A_n \tilde{W}_{g_n}^T + \lambda^{-1} I)^{-1} \tilde{W}_{g_n} A_n \tilde{W}_{h_n}^T \tag{3.65}
\]

The variance for the \(i\)th missing data value in \(h_n\) is given by the \(ii\)th element of the covariance matrix \(C_{h_n}\); this is not calculated for each iteration, just once when the iterative process has converged.

### 3.3. Illustration: synthetic data

It is necessary to demonstrate that PPCA-AR is able to estimate model parameters effectively, which means constructing synthetic data which conforms to the model’s assumptions and for which the parameters are known. The first illustration demonstrates the estimation of the observation and excitation noise precisions, \(\lambda\) and \(\kappa_q\) (for the \(q\)th latent variable) respectively. The second illustration demonstrates the estimation of the more structural parameters: \(W\) and \(\theta_q\). The third illustration demonstrates that the temporal modelling included in PPCA-AR is advantageous when imputing missing values.

A stationary AR signal of order \(P\), for the \(q\)th latent variable, may be generated by careful construction of the vector of coefficients, \(\theta_q\). This may be achieved as follows. If \(P\) is even, \(P/2\) complex conjugate pairs \(\eta_i, \bar{\eta}_i\) lying within the unit circle in the complex plane are drawn randomly (uniformly with respect to area); if \(P\) is odd, a single \(\eta_P\) is drawn on the real axis and \([P/2]\) conjugate pairs are drawn for the remainder. The \(\eta_i\) and \(\bar{\eta}_i\) are used as the roots of the auxiliary polynomial and \(\theta_q\) is the vector whose elements are its coefficients. The data sequence is initialised by generating random values for the first \(P\) elements of \(x_q\): \(x_{q,n}\), for \(n\) from \(P + 1\) to \(2N\), is calculated from (3.3), where a random sample of excitation noise, \(\epsilon_{q,n}\), is selected from the Gaussian distribution with zero mean and precision \(\kappa_q\). Finally the first \(N\) elements of \(x_q\) are deleted to ensure that the whole sample conforms to the autoregression model.

The synthetic data is generated according to the expression \(t_n = Wx_n + e_n\), where \(W\) is some \(D \times Q\) matrix and \(e_n\) is a \(D\)-dimensional vector whose elements are randomly drawn from a Gaussian with zero mean and precision \(\lambda\). The latent variables are each generated independently from AR processes of order \(P\), using a different, randomly-generated \(\theta\) (as above) and potentially different precisions, \(\kappa_q\), for each variable.

#### 3.3.1. Illustration 1: estimating noise precisions

Datasets were generated with \(D = 3\), \(N = 200\), \(Q = 2\) and \(P = 5\) and with every combination of \(\lambda\) and \(\kappa_q\) (the same value for both latent variables) in the set \(\{1000, 100, 10, 1, 0.1\}\).

The elements of \(W\) were randomly sampled from a Gaussian distribution with zero mean and unit variance and different \(\theta_q\) values were generated for each latent variable. This was repeated 100 times, resulting in 2500 synthetic test datasets. PPCA-AR was trained
against each of them using the known $Q$ and $P$ values, until convergence. Figure 3.3 shows that there is a close correlation between the model’s estimates of $\lambda$ and the actual values used to generate the data, even where the actual $\lambda$ value is significantly larger than the actual $\kappa_q$ values, i.e. when the observation noise variance is significantly smaller than the excitation noise variance. Figure 3.4 shows that the model’s estimates of the $\kappa_q$ values for each row are related to the actual values, but they are heavily influenced by the observation noise and tend to be under-estimated.

For each of these same 2500 datasets the mean squared error between the model’s reconstruction of the data and the actual data was calculated as a measure of accuracy. Figure 3.5 shows how the PPCA-AR accuracy compares with that for PPCA. In every case the accuracy achieved by the standard PPCA model is better than that for PPCA-AR. This is to be expected since PPCA is known to provide the best fit of the data in the mean squared error sense and ignores the temporal structure of the data, while the PPCA-AR model includes the extra temporal constraints. Figure 3.6 shows that the accuracy is closely related to the magnitude of the observation noise; unsurprisingly, the greater the observation noise, the greater the error of reconstruction.
3. Maximum likelihood PPCA with autoregressive latent variables

Figure 3.4. Comparison of estimated excitation noise precision $\kappa_q$ with actual values for each of 2500 trials. Top and bottom panels correspond to the two rows of $X$. Grey dots are estimated values; solid black lines are actual values; grey dotted line is the actual observation noise precision, $\lambda$.

Figure 3.5. Comparison of PPCA-AR mean squared error of data reconstruction with the equivalent PPCA error. The PPCA-AR error is always greater than that for PPCA (the black line indicates equality).
3. Maximum likelihood PPCA with autoregressive latent variables

Figure 3.6. Comparison of PPCA-AR mean squared error of data reconstruction with the actual observation noise variance, $\lambda^{-1}$.

Figure 3.7. Comparison of estimated $\theta$ values (crosses) with actuals (circles) for the 5 rows of $X$. 
3.3.2. Illustration 2: estimating $W$ and $\theta$

In PPCA-AR the columns of $W$ are constrained to be orthonormal to one another, so the estimates of $W$ (and hence $X$ and the $\theta_q$ vectors) can only accurately predict the actuals if the synthetic data are generated with that constraint. Even in this case the order and sign of the columns of $W$ and the corresponding rows of $X$ may differ between estimates and actuals without affecting the reconstruction of the data. To ensure that a proper comparison is made the linear correlation coefficients between the estimated and actual rows of $X$ are inspected.

A dataset was generated with $D = 12$, $N = 500$, $Q = 6$, $P = 10$ and $\lambda = 10$. For each row of $X$, $\kappa_q = 2.5$ and $\theta_q$ was different in each case. PPCA-AR was trained against each of them using the known $Q$ and $P$ values, until convergence. Figure 3.7 compares the model’s estimates of $\theta_q$ with the actual values. Figure 3.8 shows Hinton diagrams for the actual and estimated $W$ matrices. In both cases the correspondence is good, though this has been found to be dependent on the relative sizes of the noise variances.

3.3.3. Illustration 3: imputing missing values

In order to register where elements of the observations are missing, a mask matrix, $M$, of the same dimensionality as $T$, is created. Each element, $m_{i,j}$, is an indicator variable which takes the value 1 if the corresponding $t_{i,j}$ is observed or 0 if it is missing. Initially all the values are set to 1. A gap may be introduced into the data by uniformly randomly selecting a row and column of $M$ as the centre, sampling from a Poisson distribution of mean 3 to determine the length of the gap and then setting the appropriate values in that row of $M$ to zero, that is, the data is missing in gaps with temporal extents. Gaps were generated independently for each row of $M$ to ensure that approximately the same proportion of values were missing for each observed variable. Different $M$ matrices were generated for 0% to 90% missing at 5% intervals, with the gaps being cumulative, meaning...
that the 20% missing \( M \) was the 15% missing \( M \) with 5% more values missing. These percentages are approximate as the algorithm added gaps until the required percentage was just exceeded.

A new dataset was generated with parameters \( D = 12, N = 200, Q = 6, P = 10, \lambda = 10 \) and \( \kappa_q = 10 \). PPCA-t, PPCA-g and PPCA-AR, were each trained against this dataset for 2000 iterations for each of the missing patterns. This process was performed against 30 different synthetic datasets, using a different set of missing masks each time. PPCA-g had a tendency to oscillate; this was prevented by smoothing the update of \( W \) by setting it equal to the convex combination \( 0.75W_{\text{old}} + 0.25W_{\text{new}} \). Neal and Hinton [1998] have shown that this does not prevent the convergence of the EM algorithm. The three models were compared by measuring the mean squared error of the reconstruction of the missing values. The results are shown in figure 3.9, where the mean errors are plotted against the proportions of missing data, along with one standard deviation above and below the mean.

Figure 3.9 shows that both PPCA-t and PPCA-g display similar accuracy of predictions for the missing values up to about 55% missing, and that PPCA-AR demonstrates significantly better predictions of missing values up to 65% missing, again with generally increasing uncertainty as more data is missing. Above 65% missing it seems that PPCA-t performs better than PPCA-AR. This is due to the fact that as the gaps in the data become larger and more frequent, there is no longer any evidence in the data for the autoregressive process of order 10 and the advantage of PPCA-AR’s temporal modelling is lost.
3. Maximum likelihood PPCA with autoregressive latent variables

Figure 3.10. ONS data: comparison of the mean squared error of reconstruction of the missing data, averaged over 30 different sets of missing masks. The dashed lines represent one standard deviation above and below the mean.

3.4. Results: real data

The three models, PPCA-t, PPCA-g and PPCA-AR, were applied to the ONS data described in section 1.1, a dataset with $D = 23$ and $N = 494$. PPCA-AR occasionally produced a large spike in the first two or three iterations in the calculation of the observation noise variance, so for computational convenience all values were divided by 100 to ensure that the variance for any dimension was less than 10.

Using the same method as for synthetic data, gaps were made in the ONS data, with missing proportions from 0% to 90% in increments of 5%. An arbitrary model order of $Q = 5$ was selected for all three models, and $P = 20$ for PPCA-AR. The models were each trained against the data for 2000 iterations. This process was repeated using 30 different sets of missing masks. The results, shown in figure 3.10, demonstrate that PPCA-AR is significantly better at predicting the missing values than PPCA for up to 45% missing values.

Figure 3.11 shows the observation noise precision $\lambda$ calculated over the same 30 sets of missing masks for the three models. The PPCA-AR model’s estimate is more robust to the proportion of missing data than either PPCA-t or PPCA-g; the line is almost horizontal up to about 55% missing data.

It is instructive to compare the reconstructions made by PPCA-AR and PPCA. Figure 3.12 contains three plots: the top two show the reconstructions for dimension 22 (that with the largest variance) and dimension 3 (that with the smallest variance) with approximately
3. Maximum likelihood PPCA with autoregressive latent variables

Figure 3.11. ONS data: comparison of the observation noise precision $\lambda$ averaged over 30 different sets of missing masks. The dashed lines represent one standard deviation above and below the mean.

50% of the values missing over the whole dataset. The gray bars indicate where data is missing for that dimension. The bottom plot is a histogram showing how many values are missing for each observation. Regarding dimension 22, the horizontal line which PPCA-t and PPCA-g seem to favour where data is missing corresponds to the mean value of the observed data for this dimension.

Figure 3.13 shows how the variances associated with the reconstructions differ between PPCA-AR and PPCA. As well as calculating more accurate estimates for the missing values, PPCA-AR is much more confident in those estimates.

3.5. Conclusions

Although PPCA, for a particular model order, gives the optimal mean squared reconstruction of data, the PPCA-AR model, which additionally models temporal structure in the data, provides better reconstruction of missing values. As might be expected, this additional temporal constraint makes it less good at modelling the observed values. Where a whole observation is missing, the autoregressive processes within PPCA-AR are able to estimate values for the observation dependent on the observations before and after it in time, while PPCA is only able to fill in the missing data with a mean observation.

On synthetic data, PPCA-AR is able to estimate the observation noise precision, $\lambda$, effectively. The quality of excitation noise precision, $\kappa_q$, estimation is dependent on $\lambda$ being greater than $\kappa$. $W$ is well estimated to within a sign, permutation and/or rotation of the original, dependent on the relative sizes of the noise precisions. Estimates of $\theta$ can be good, but the sign, permutation and/or rotation of $W$ often leads to AR coefficients that do not mimic the originals. The variance of the reconstruction of the missing observation values is smaller for PPCA-AR than for PPCA, but it should be noted that in both cases it is being under-estimated as there are no contributions of variation from the model parameters.
3. Maximum likelihood PPCA with autoregressive latent variables

Figure 3.12. ONS data: Reconstruction of missing data by PPCA-g (light-blue), PPCA-t (dark blue) and PPCA-AR (red) for those dimensions, 22 and 3, that have the greatest and smallest variances respectively. Vertical grey stripes indicate where values are missing and the black line shows the actual values. The histogram (bottom) shows the total number of values missing across all dimensions.

Figure 3.13. ONS data, dimension 1. Comparison of standard deviation error bars for missing data by PPCA-g (light-blue), PPCA-t (dark blue) and PPCA-AR (red), with the actual values shown in black. 40% of data values are missing.
The benefit of PPCA-AR has been demonstrated on the real ONS data where the improvement in estimation of missing values is marked for variables having notably different variances, even with as much as 50% of the data values missing.

Further work is required to determine why the PPCA-g model, which is using a more principled method of filling in the missing values, performs poorly compared with PPCA-t, especially since PPCA-AR uses the same method but does not seem to be badly affected. I surmise that the autoregressive processes in PPCA-AR are having a beneficial effect in controlling the variability of the latent variables.

This chapter has introduced PPCA-AR as a maximum likelihood model and demonstrated that it has merit for the imputation of missing data where the data form a multivariate time-series. It results in point estimates for $W, \theta, \lambda$ and $\kappa_q$, so there is no measure of uncertainty for these parameters, and it does not address the problem of determining the model orders, i.e. the number of components in $W, Q$, and the number of AR coefficients in $\theta, P$. It would also be beneficial to introduce some means of enabling the mean observation to be non-zero. The noise assumptions in the model are Gaussian, so it will be adversely affected by outliers and not able to model leptokurtic data.

These issues will be addressed in the following chapters. Chapters 4 and 5 start by introducing robust, variational Bayesian versions of PPCA and AR separately, both making use of automatic relevance determination [Mackay, 1994] to estimate the model orders. These are then brought together in chapter 6 to form a robust, variational Bayesian version of PPCA-AR with automatic relevance determination.
4. Robust variational PPCA

Probabilistic Principal Component Analysis (PPCA), proposed by Tipping and Bishop [1999], decomposes multivariate observations using the latent variable model expressed in the previous chapters as

\[ t_n = Wx_n + \bar{t} + e_n \]  

With \( N \) observations in total, \( t_n \) is the \( n \)th \( D \)-dimensional observation, \( x_n \) is the \( n \)th \( Q \)-dimensional vector of latent variables which is assumed to be Gaussian distributed, \( \mathcal{N}(x_n \mid 0, I) \), \( W \) is the mixing matrix, \( \bar{t} \) enables the observations to have a non-zero mean, and \( e_n \) is a noise term, also assumed to be Gaussian distributed with isotropic covariance, \( \mathcal{N}(e_n \mid 0, \lambda^{-1}I) \). Tipping and Bishop provide both an algebraic solution and an EM algorithm for determining the maximum likelihood solution.

Bishop [1999a] describes a Bayesian treatment for PPCA which incorporates automatic relevance determination (ARD) (see section 2.3) to estimate the model order by controlling the magnitude of the components in \( W \). As with most non-trivial Bayesian models, the integrals for exact Bayesian inference prove to be intractable, so Bishop applies a type-II maximum likelihood approach to the estimation of the ARD parameters, and determines maximum likelihood point estimates for other parameters.

Zhang et al. [2004] also describe a Bayesian treatment for PPCA. They use a reversible jump Markov Chain Monte Carlo (RJ-MCMC) approach [Green, 1995] to approximate the posterior distributions of the model parameters, including the model order, using Gibbs sampling. Bishop [1999a] refers to an unpublished paper that describes a similar scheme.

The type-II maximum likelihood approach does not result in probability distributions for all the model parameters, and RJ-MCMC can be computationally expensive, even for simple problems. Instead, in this chapter a factorized variational approximation approach is described, incorporating ARD as described by Bishop [1999a] for the model order estimation. This approach allows the Gaussian assumptions regarding the observation noise and latent variables to be generalised to an infinite mixture of Gaussians having the same mean and different precisions by using Student-t distributions instead. This generalisation makes the model more robust to outliers.

The chapter begins by reviewing Gaussian variational PPCA models and then progresses to replace the Gaussian assumptions with Student-t distributions to provide a new robust variational PPCA model.
4. Robust variational PPCA

Bishop [1999b] introduces variational PPCA as a Gaussian model, with ARD used to estimate the model order by controlling the magnitudes of the components in $W$. Each component (column of $W$), $w_q$, is assigned the Gaussian ARD prior $N(w_q | 0, \delta_q^{-1}I)$. The precision, $\delta_q$, controls the magnitude of $w_q$, constraining it to have values close to zero, thereby effectively switching it off, when $\delta_q$ is large. The $\delta_q$ is only small when there is sufficient evidence in the data for the magnitude of $w_q$ to be greater than zero. A common Gamma prior, $G(\delta_q | a_\delta, b_\delta)$, is placed over each of the precisions. If $a_\delta$ and $b_\delta$ are chosen to have the same value then this is effectively a Student-t prior over the columns of $W$, c.f. (2.47). Tipping [2001] presents a nice graphical illustration (see figure 4.1) that the joint distribution, $p(x_1, x_2)$, of two Student-t densities concentrates probability mass close to zero values of $x_1$ and $x_2$ rather than in regions where both $x_1$ and $x_2$ are non-zero, thus encouraging sparse solutions.

As is the case for PPCA, a zero mean, unit variance Gaussian prior is selected for each latent variable value, $x_{q,n}$. In the absence of any other information, a very vague zero mean Gaussian prior is selected for $\bar{t}$ and a Gamma prior is used for the noise precision $\lambda$. Figure 4.2 shows a graphical representation of the parameter dependencies for this model.

4.2. Robust variational PPCA

Luttinen et al. [2009] incorporate robustness into this variational PPCA model by assuming that outliers in the data are as a result of outliers in the observation noise. They model the latent variables as for standard PPCA and the observation noise as independent Student-t
4. Robust variational PPCA

Figure 4.2. Graphical model for Bishop’s [1999b] Gaussian variational PPCA. The \( \mathbf{m} \) and \( \mathbf{V} \) are the mean and covariance for a Gaussian distribution; while \( a \) and \( b \) are the two parameters for a Gamma density.

For each variable:

\[
\begin{align*}
p(\mathbf{x}_n) &= \mathcal{N}(\mathbf{x}_n | \mathbf{0}, \mathbf{I}) \\
p(\mathbf{e}_{r,n}) &= \mathcal{S}(\mathbf{e}_{r,n} | 0, \lambda_r, v_r)
\end{align*}
\]

(4.2) \hspace{2cm} (4.3)

Using (2.47) they replace each of these Student-t distributions with a Gaussian conditioned on a new latent variable, \( u_{r,n} \), for which posterior distributions must be inferred as part of the process, and a Gamma prior over the \( u_{r,n} \):

\[
\begin{align*}
p(\mathbf{e}_{r,n} | 0, \lambda, u_{r,n}) &= \mathcal{N}(\mathbf{e}_{r,n} | 0, (\lambda u_{r,n})^{-1}) \\
p(u_{r,n} | v_r) &= \mathcal{G}(u_{r,n} | \frac{v_r}{2}, \frac{v_r}{2})
\end{align*}
\]

(4.4) \hspace{2cm} (4.5)

Figure 4.3a shows a graphical representation of the parameter dependencies for this model. This is in fact more like robust variational Factor Analysis (see section 4.7), since the diagonal precision matrix for the observation noise in \( \mathbf{e}_n \) has potentially different values along the diagonal.

Gai et al. [2008] incorporate robustness into the variational PPCA model by assuming that outliers are as a result of outliers in the latent variables and that the heavy-tailed nature of some data is as a result of a heavy-tailed observation noise distribution. They model \( \mathbf{x}_n \) and \( \mathbf{e}_n \) as samples from Student-t distributions with a common value for the degrees of freedom:

\[
\begin{align*}
p(\mathbf{x}_n) &= \mathcal{S}(\mathbf{x}_n | \mathbf{0}, \mathbf{I}, v) \\
p(\mathbf{e}_n) &= \mathcal{S}(\mathbf{e}_n | \mathbf{0}, \lambda^{-1} \mathbf{I}, v)
\end{align*}
\]

(4.6) \hspace{2cm} (4.7)
Using (2.47) each of these Student-t distributions is replaced by a Gaussian conditioned on the latent variable, $u_{i,n}$, and a Gamma prior over the $u_{i,n}$, $\mathcal{G}(u_{i,j} \mid \frac{\nu}{2}, \frac{\nu}{2})$. This leads to the following priors (using their abuse of notation for the latent variables):

$$p(x_{q,n} \mid u_{q,n}) = \mathcal{N}(x_{q,n} \mid 0, u_{q,n}^{-1}) \quad (4.8)$$
$$p(e_{r,n} \mid \lambda, u_{r,n}) = \mathcal{N}(e_{r,n} \mid 0, (\lambda u_{r,n})^{-1}) \quad (4.9)$$
$$p(u_{i,n} \mid v) = \mathcal{G}(u_{i,n} \mid \frac{\nu}{2}, \frac{\nu}{2}) \quad (4.10)$$

They further assume that the degrees of freedom of the Student-t, $\nu$, has a fixed value that is already known, though they suggest that it might be included in the estimation process. Figure 4.3b shows a graphical representation of the parameter dependencies for this model.

It is not obvious that the observation noise and latent variables should share a distribution; the degrees of freedom might be quite different, for example the noise may be Gaussian while the latent variables account for both outliers and the leptokurtic nature of the data. It is also unlikely that the degrees of freedom will be known a priori; it would be preferable to infer posterior distributions as part of the process.

If the latent variables were independent then they may be drawn from Student-t distributions with different degrees of freedom (though they are still constrained to have zero mean and unit precision). Where ARD switches off a component in $W$, the corresponding latent variable may also be switched off or it may be left with arbitrary values. If all the latent variables share a distribution then these switched off variables will be included in
4. Robust variational PPCA

Figure 4.4. Graphical model for the new robust, variational PPCA model.

the estimation process for that distribution; if their values are constrained to zero then, with zero variance and hence infinite precision, the effect on the estimate of the posterior distribution of the latent variables may be significant.

4.3. New robust variational PPCA

The issues with the Gaussian variational PPCA model are avoided by the new robust, variational PPCA model proposed here. It assumes separate Student-t priors for the observation noise, $e_n$, and for each of the latent variables, $x_n$, and incorporates the degrees of freedom variables ($v$ for the observation noise and $d_q$ for the $q$th latent variable) into the estimation process.

In the robust models described previously in this chapter the ARD variable, $\delta_q$, constrains the magnitude of the $q$th column of $W$. Where there is no evidence in the data for that particular component, the magnitude is constrained towards zero. In the $Wx_n$ term of this latent variable model, reducing the magnitude of one column in $W$ can be offset by an equivalent increase in the magnitude of the corresponding latent variable (row of $X$, where $X = (x_1, \ldots, x_N)$). Where ARD switches off a component, the corresponding latent variable is essentially arbitrary. To counter this, and to ensure that the latent variable is also switched off, the same ARD variable, $\delta_q$ is incorporated into the precision of the latent variables.

As before, the vector $\bar{t}$ represents the mean observation. Two new sets of latent variables, in $u_n$ and $z_n$, support the expression of the Student-t distributions as scale mixture of Gaussians, as per (2.47). Figure 4.4 shows a graphical representation of the parameter dependencies for this new model.
4.3.1. Priors

In their generalised component analysis, Tipping and Lawrence [2005] model the noise as Student-t distributed and decompose it into a scale mixture of Gaussians as follows:

\[
S(e_{r,n} | 0, \lambda, v) = \int \mathcal{N}(e_{r,n} | 0, \beta^{-1}_{r,n}) \mathcal{G}(\beta_{r,n} | a, b) \, d\beta_{r,n} \tag{4.11}
\]

where \( v = 2a \) and \( \lambda = \sqrt{a/b} \). However, since \( a \) and \( b \) are not independent there are no conjugate priors for them that fit into the chosen formulation. Hence they are considered to be parameters rather than variables and point values are estimated for them outside the variational framework. An alternative method for representing the scale mixture of Gaussians for the Student-t (and the one selected by Luttinen et al. [2009] and Gai et al. [2008]) introduces a new latent variable, \( u_{r,n} \):

\[
S(e_{r,n} | 0, \lambda, v) = \int \mathcal{N}(e_{r,n} | 0, (\lambda u_{r,n})^{-1}) \mathcal{G}(u_{r,n} | v/2, v/2) \, du_{r,n} \tag{4.12}
\]

For this formulation there are suitable conjugate priors for both \( \lambda \) and \( v \) and posterior distributions may be calculated for them as part of the variational procedure.

With the observation noise assigned the Student-t prior \( S(e_n | 0, \lambda^{-1}I, v) \), the likelihood for each observation, derived from (4.1), is \( p(t_n | W, x_n, \bar{t}, \lambda, v) = S(t_n | WX_n + \bar{t}, \lambda, v) \). By introducing a new set of latent variables, \( u_n \), the Student-t may be expressed, as before, as a scale mixture of Gaussians with a Gamma distribution over the degrees of freedom:

\[
p(t_n | W, x_n, \bar{t}, u_n) = \mathcal{N}(t_n | WX_n + \bar{t}, (\lambda \text{diag}(u_n))^{-1}) \tag{4.13}
\]

\[
p(u_{r,n} | v) = \mathcal{G}(u_{r,n} | v/2, v/2) \tag{4.14}
\]

\[
p(v) = \mathcal{G}(v | a_v, b_v) \tag{4.15}
\]

Applying ARD priors over the columns of \( W \), with precision of component \( q \) being controlled by the ARD variable \( \delta_q \) gives, for each row of \( W \):

\[
p(w_{r,:} | \delta) = \mathcal{N}(w_{r,:} | 0, \text{diag}(\delta)^{-1}) \tag{4.16}
\]

\[
p(\delta_q) = \mathcal{G}(\delta_q | a_\delta, b_\delta) \tag{4.17}
\]

where \( w_{r,:} \) denotes the \( r \)th row of \( W \). The distributions for \( W \) are expressed in terms of \( w_{r,:} \), rather than the more obvious \( p(w_q) = \mathcal{N}(w_q | 0, \delta_q^{-1}I) \) for later mathematical convenience.

For the latent variables, each of the \( Q \) variables is modelled by a separate Student-t distribution, unrelated to the observation noise: \( S(x_{q,n} | 0, \kappa_q, d_q) \). As for PPCA there needs to be some fixing of the scaling between \( W \) and the \( x_n \). In the Gaussian model this was achieved by setting the precision of the latent variables to 1. Applying the same restriction here, each \( \kappa_q \) is set to the value 1. Note, however, that the variance of the Student-t is given by \( d_q \lambda/(d_q-2) \), so the variance of the latent variables is not fixed, but varies.
4. Robust variational PPCA

with $d_q$.

As before, each latent variable Student-t distribution may be expressed as a scale mixture of Gaussians, by associating a new latent variable, $z_{q,n}$, with each $x_{q,n}$, again with a Gamma prior for the degrees of freedom parameters. With the ARD variable incorporated into the precision of the latent variables, the resulting priors related to the latent variable Student-t distribution are as follows:

$$p(x_n | \delta, z_n) = \mathcal{N}(x_n | 0, (\text{diag}(\delta) \text{diag}(z_n))^{-1}) \quad (4.18)$$
$$p(z_{q,n} | d_q) = \mathcal{G}(z_{q,n} | \frac{d_q + d_q}{2}, \frac{d_q}{2}) \quad (4.19)$$
$$p(d_q) = \mathcal{G}(d_q | a_d, b_d) \quad (4.20)$$

The observation mean, $\bar{t}$, and observation noise precision, $\lambda$, are assigned the following priors:

$$p(\bar{t}) = \mathcal{N} (\bar{t} | m_\bar{t}, V_\bar{t}) \quad (4.21)$$
$$p(\lambda) = \mathcal{G} (\lambda | a_\lambda, b_\lambda) \quad (4.22)$$

In the absence of a priori knowledge regarding these probability distributions we use uninformative priors, setting $a_\delta = b_\delta = a_v = b_v = a_d = b_d = a_\lambda = b_\lambda = 10^{-3}$, $m_\bar{t} = 0$ and $V_\bar{t} = 10^3 I$.

4.3.2. Factorised variational Bayesian method

The joint distribution, $p(T, \Omega)$, of the data, $T$, and all the model parameters, $\Omega = \{W, X, \delta, U, Z, \bar{t}, \lambda, v, d\}$, may be factorised as

$$p(T | W, X, \bar{t}, U, \lambda) p(W | \delta) p(X | \delta, Z) p(\delta) p(U | v) p(v) p(Z | d) p(d) p(\bar{t}) p(\lambda). \quad (4.23)$$

To apply the factorised variational Bayesian technique described in chapter 2, the approximate joint posterior distribution, $q(\Omega | T)$, is also grouped into factors following the prior factorisation:

$$q(W | T) q(X | T) q(\delta | T) q(U | T) q(v | T) q(Z | T) q(d | T) q(\bar{t} | T) q(\lambda | T). \quad (4.24)$$

For each of these groups in turn, the factorised variational Bayesian method is used obtain the approximate posterior distributions.
4. Robust variational PPCA

Mixing matrix, $W$

Starting from the factorisation of the joint distribution in (4.23), the approximate posterior for the components that maximises the lower bound, $F(q)$, is

$$\log(q(W | T)) = E_{\mathbf{w}, \mathbf{r}, \mathbf{i}} \left[ \log(p(T | W, X, \bar{t}, U, \lambda) p(W | \delta) p(X | \delta, Z) \right. \left. p(\delta) p(U | v) p(v) p(Z | d) p(d) p(\bar{t}) p(\lambda) \right]$$

where $E_{\alpha}[b]$ denotes the expectation of $b$ taken with respect to the approximate posteriors of all variables except $a$. Expanding this and moving all terms not dependent on $W$ into a single constant term gives

$$\log(q(W | T)) = E_{\mathbf{w}, \mathbf{r}, \mathbf{i}} \left[ \log(p(T | W, X, \bar{t}, U, \lambda)) + \log(p(W | \delta)) \right] + \text{const} \quad (4.25)$$

Again, expanding and moving all terms not dependent on $W$ into the constant term:

$$\log(q(W | T)) = -\frac{1}{2} E_{\mathbf{w}, \mathbf{r}, \mathbf{i}} \left[ \sum_{r=1}^{D} \left( \mathbf{w}_r : \left( \text{diag}(\delta) + \lambda \sum_{n=1}^{N} \mathbf{u}_{r,n} \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{w}_r^T \right) \right. \left. - 2 \left( \lambda \sum_{n=1}^{N} \mathbf{u}_{r,n} (t_{r,n} - \bar{t}_r) \mathbf{x}_n \right) \mathbf{w}_r^T \right] + \text{const} \quad (4.27)$$

$$= -\frac{1}{2} \sum_{r=1}^{D} \left\{ \mathbf{w}_r : \left( \text{diag}(\langle \delta \rangle) + \langle \lambda \rangle \sum_{n=1}^{N} \langle \mathbf{u}_{r,n} \rangle \langle \mathbf{x}_n \mathbf{x}_n^T \rangle \right) \mathbf{w}_r^T \right. \left. - 2 \langle \lambda \rangle \sum_{n=1}^{N} \langle \mathbf{u}_{r,n} \rangle (t_{r,n} - \langle \bar{t}_r \rangle) \langle \mathbf{x}_n \rangle \mathbf{x}_n^T \right\} + \text{const} \quad (4.28)$$

where $\langle f(a, b) \rangle$ denotes the expectation of $f(a, b)$ taken with respect to the approximate posterior distributions of $a$ and $b$.

Since (4.28) is quadratic in $\mathbf{w}_r :$, it can be seen that $q(\mathbf{w}_r : | T)$ is a Gaussian and we have

$$q(\mathbf{w}_r : | T) = \mathcal{N}(\mathbf{w}_r : | \mathbf{\mu}_{\mathbf{w}_r}, \mathbf{\Sigma}_{\mathbf{w}_r}) \quad (4.29)$$

where

$$\mathbf{\Sigma}_{\mathbf{w}_r} = \left( \text{diag}(\langle \delta \rangle) + \langle \lambda \rangle \sum_{n=1}^{N} \langle \mathbf{u}_{r,n} \rangle \langle \mathbf{x}_n \mathbf{x}_n^T \rangle \right)^{-1} \quad (4.30)$$

$$\mathbf{\mu}_{\mathbf{w}_r} = \mathbf{\Sigma}_{\mathbf{w}_r} \left( \langle \lambda \rangle \sum_{n=1}^{N} \langle \mathbf{u}_{r,n} \rangle \langle \mathbf{x}_n \rangle (t_{r,n} - \langle \bar{t}_r \rangle) \right) \quad (4.31)$$

Each $\mathbf{\mu}_{\mathbf{w}_r}$ makes up the $r$th row of the posterior expectation of $W$. In the following calculations it will be seen that the posterior distributions of the other parameter variables...
may be expressed in terms of \( \langle W \rangle, \langle w_{r,:} \rangle \) and \( \langle w^T_{r,:}w_{r,:} \rangle \).

### Latent variables, \( x_n \)

Applying the same procedure for each of the \( x_n \):

\[
\log(q(x_n | T)) = \mathbb{E}_{x_n} [\log(p(t_n | W, x_n, \hat{t}, u_n, \lambda)) + \log(p(x_n | \delta, z_n))] + \text{const} \quad (4.32)
\]

\[
= \mathbb{E}_{x_n} [\log(\mathcal{N}(t_n | WX_n + \hat{t}, (\lambda \text{diag}(u_n))^{-1})) + \log(\mathcal{N}(x_n | 0, (\text{diag}(\delta) \text{diag}(z_n))^{-1}))] + \text{const} \quad (4.33)
\]

Expanding and moving all terms not dependent on \( x_n \) into the constant term, and taking expectations:

\[
\log(q(x_n | T)) = -\frac{1}{2} \left\{ x_n^T \left( \text{diag}(\langle \delta \rangle) \text{diag}(\langle z_n \rangle) + \langle \lambda \rangle \sum_{r=1}^D \langle u_{r,n} \rangle \langle w^T_{r,:}w_{r,:) \rangle \right) x_n -2x_n^T \left( \langle \lambda \rangle \sum_{r=1}^D \langle u_{r,n} \rangle \langle w_{r,:) \rangle^T (t_{r,n} - \langle \mu_r \rangle) \right) \right\} + \text{const} \quad (4.34)
\]

This may be recognised as the logarithm of a Gaussian distribution:

\[
q(x_n | T) = \mathcal{N}(x_n | \mu_{x_n}, \Sigma_{x_n}) \quad (4.35)
\]

where

\[
\Sigma_{x_n} = \left( \text{diag}(\langle \delta \rangle) \text{diag}(\langle z_n \rangle) + \langle \lambda \rangle \sum_{r=1}^D \langle u_{r,n} \rangle \langle w^T_{r,:}w_{r,:) \rangle \right)^{-1} \quad (4.36)
\]

\[
\mu_{x_n} = \Sigma_{x_n} \langle \lambda \rangle \sum_{r=1}^D \langle u_{r,n} \rangle \langle w_{r,:) \rangle^T (t_{r,n} - \langle \mu_r \rangle) \quad (4.37)
\]

### ARD parameters, \( \delta \)

With the same ARD parameters applied to both the columns of \( W \) and the corresponding rows of \( X \), we have:

\[
\log(q(\delta | T)) = \mathbb{E}_{\delta} [\log(p(W | \delta)) + \log(p(X | \delta)) + \log(p(\delta))] + \text{const} \quad (4.38)
\]

\[
= \mathbb{E}_{\delta} \left[ \sum_{r=1}^D \log(\mathcal{N}(w_{r,:} | 0, \text{diag}(\delta)^{-1})) + \sum_{n=1}^N \sum_{q=1}^Q \log(\mathcal{N}(x_{q,n} | 0, (\delta_q z_{q,n})^{-1})) \right. \\
\left. + \sum_{q=1}^Q \log(\mathcal{G}(\delta_q | a_\delta, b_\delta)) \right] + \text{const} \quad (4.39)
\]

\[
= \mathbb{E}_{\delta} \left[ \frac{1}{2} \log | \text{diag}(\delta) | - \frac{1}{2} \sum_{r=1}^D \text{tr}(w_{r,:} \text{diag}(\delta) w_{r,:}^T) + \frac{1}{2} \sum_{n=1}^N \sum_{q=1}^Q (\log(\delta_q) - x_{q,n}^2 z_{q,n} \delta_q) \right]
\]
+ \sum_{q=1}^{Q} \left( (a_\delta - 1) \log(\delta_q) - b_\delta \delta_q \right) \right] + \text{const} \quad (4.40)

This gives rise to the following expression for each \( \delta_q \):

\[
\log(q(\delta_q | \mathbf{T})) = \left( a_\delta + \frac{D}{2} - 1 \right) \log(\delta_q) - \left( b_\delta + \frac{1}{2} \sum_{r=1}^{D} \langle w^2_{r,q} \rangle + \frac{1}{2} \sum_{n=1}^{N} \langle x^2_{q,n} \rangle \right) \delta_q + \text{const} \quad (4.41)
\]

This may be recognised as the logarithm of a Gamma distribution:

\[q(\delta_q | \mathbf{T}) = \mathcal{G}(\delta_q | \alpha_\delta, \beta_\delta) \quad (4.42)\]

where

\[\alpha_\delta = a_\delta + \frac{D}{2} \quad (4.43)\]
\[\beta_\delta = b_\delta + \frac{1}{2} \sum_{r=1}^{D} \langle w^2_{r,q} \rangle + \frac{1}{2} \sum_{n=1}^{N} \langle x^2_{q,n} \rangle \quad (4.44)\]

**Latent variables, \( z_{q,n} \)**

Locating a joint distribution for \( z_{q,n} \) is not analytically tractable, so, since the \( z_{q,n} \) are conditionally independent, we may examine each of them individually. With the constant term dropped from now on to aid readability, the following is obtained:

\[
\log(q(z_{q,n} | \mathbf{T})) = \mathbb{E}_{z_{q,n}} \left[ \log(p(x_{q,n} | \delta_q, z_{q,n})) + \log(p(z_{q,n} | d_q)) \right]
\]

\[= \mathbb{E}_{z_{q,n}} \left[ \log \left( \mathcal{N}(x_{q,n} | 0, (\delta_q z_{q,n})^{-1}) \right) + \log(\mathcal{G}(z_{q,n} | \frac{d_q}{2}, \frac{d_q}{2})) \right] \quad (4.46)\]

\[= \left( \frac{\langle d_q \rangle}{2} + 1 \right) \log(z_{q,n}) - \left( \frac{\langle d_q \rangle}{2} + \frac{1}{2} \langle x^2_{q,n} \rangle \right) z_{q,n} \quad (4.47)\]

The result is the following Gamma distribution for each \( z_{q,n} \):

\[q(z_{q,n} | \mathbf{T}) = \mathcal{G}(z_{q,n} | \alpha_{z_q}, \beta_{z_{q,n}}) \quad (4.48)\]

where

\[\alpha_{z_q} = \frac{\langle d_q \rangle}{2} + 1 \quad (4.49)\]

is common to all the \( z_{q,n} \) and

\[\beta_{z_{q,n}} = \frac{1}{2} (\langle d_q \rangle + \langle x^2_{q,n} \rangle) \quad (4.50)\]

As \( d_q \) becomes large, the \( z_{q,n} \) tend to 1 and the distribution of the \( q \)th latent variable tends to the Gaussian \( \mathcal{N}(x_{q,n} | 0, \delta_q^{-1}) \).
4. Robust variational PPCA

Degrees of freedom, $d$

With $d_q$ denoting the degrees of freedom for row $q$ of the latent variables, for $q(d_q \mid T)$ we obtain:

$$
\log(q(d_q \mid T)) = -N \log(\Gamma(\frac{d_q}{2})) + N \frac{d_q}{2} \log(\frac{d_q}{2})

+ \frac{d_q}{2} \sum_{n=1}^{N} (\langle \log(z_{q,n}) \rangle - \langle z_{q,n} \rangle) + (a_d - 1) \log(d_q) - b_d d_q
$$

(4.51)

which does not correspond to any standard distribution, but using Stirling’s approximation\footnote{Stirling’s first order approximation for $\log(\Gamma(a))$ is $(a - \frac{1}{2}) \log(a) - a.$} for $\log(\Gamma(\frac{d_q}{2}))$ we get

$$
\log(q(d_q \mid T)) = -N \left[ -\left( \frac{d_q}{2} - \frac{1}{2} \right) \log(\frac{d_q}{2}) + \frac{d_q}{2} \right] + N \frac{d_q}{2} \log(\frac{d_q}{2})

+ \frac{1}{2} \sum_{n=1}^{N} (\langle \log(z_{q,n}) \rangle - \langle z_{q,n} \rangle) + (a_d - 1) \log(d_q) - b_d d_q

(4.52)

$$

$$
= (a_d + N - 1) \log(d_q)

- \left( b_d - \frac{N}{2} - \frac{1}{2} \sum_{n=1}^{N} (\langle \log(z_{q,n}) \rangle - \langle z_{q,n} \rangle) \right) d_q
$$

(4.53)

which may be recognised as the log of a Gamma distribution. Therefore

$$
q(d_q \mid T) = \mathcal{G}(d_q \mid \alpha_d, \beta_{d_q})
$$

(4.54)

where

$$
\alpha_d = a_d + \frac{N}{2}
$$

(4.55)

is common to all the $d_q$ and

$$
\beta_{d_q} = b_d - \frac{1}{2} \left( N + \sum_{n=1}^{N} (\langle \log(z_{q,n}) \rangle - \langle z_{q,n} \rangle) \right)
$$

(4.56)

While Stirling’s approximation becomes very good for large values of $d_q$, as we show later it is useful even for moderately small $d_q$.

Latent variables, $u$

Locating a joint distribution for $u_{r,n}$ is not analytically tractable, so, since the $u_{r,n}$ are conditionally independent, we may examine each of them individually. The following is obtained:

$$
\log(q(u_{r,n} \mid T)) = E_{\setminus u_{r,n}} \left[ \log(p(t_{r,n} \mid w_{r,:}, x_n, t_r, u_{r,n}, \lambda)) + \log(p(u_{r,n} \mid v)) \right]

(4.57)

= E_{\setminus u_{r,n}} \left[ \log(\mathcal{N}(t_{r,n} \mid w_{r,:} x_n + t_r, (\lambda u_{r,n})^{-1})) \right]

+ \sum_{r=1}^{D} \log\left( \mathcal{G}(u_{r,n} \mid 0, \frac{v}{2}, \frac{v}{2}) \right)
$$

(4.58)
\[
= \sum_{r=1}^{D} \left\{ \left( \frac{\langle v \rangle + 1}{2} - 1 \right) \log(u_{r,n}) - \left( \frac{\langle v \rangle}{2} + \frac{1}{2} \langle \lambda \rangle (\langle t_{r,n} - w_{r,:}x_n - \bar{t}_r \rangle)^2 \right) u_{r,n} \right\}
\]

This results in the following Gamma distribution for each \( u_{r,n} \):

\[
q(u_{r,n} \mid T) = \mathcal{G}(u_{r,n} \mid \alpha_u, \beta_{u_{r,n}}) \tag{4.60}
\]

where

\[
\alpha_u = \frac{\langle v \rangle + 1}{2} \tag{4.61}
\]

\[
\beta_{u_{r,n}} = \frac{1}{2} \left( \langle v \rangle + \langle (t_{r,n} - w_{r,:}x_n - \bar{t}_r)^2 \rangle \right) \tag{4.62}
\]

with

\[
\langle (t_{r,n} - w_{r,:}x_n - \bar{t}_r)^2 \rangle = t_{r,n}^2 + \text{trace}((w_{r,:}^T w_{r,:}) (x_n x_n^T)) + \langle t_{r}^2 \rangle - 2t_{r,n} \langle (w_{r,:} x_n) + \langle t_r \rangle \rangle + 2 \langle w_{r,:} \rangle \langle x_n \rangle \langle \bar{t}_r \rangle \tag{4.63}
\]

As \( v \) becomes large, the \( u_{r,n} \) tend to 1 and the distribution of the observation noise tends to the Gaussian \( \mathcal{N}(e_n \mid 0, \lambda^{-1}I) \).

**Observation noise degrees of freedom, \( v \)**

Applying a similar method to that for the \( d_q \) described previously, the following Gamma distribution is obtained:

\[
q(v \mid T) = \mathcal{G}(v \mid \alpha_v, \beta_v) \tag{4.64}
\]

where

\[
\alpha_v = a_v + \frac{ND}{2} \tag{4.65}
\]

\[
\beta_v = b_v - \frac{1}{2} \left( ND + \sum_{n=1}^{N} \sum_{r=1}^{D} \left( \langle \log(u_{r,n}) \rangle - \langle u_{r,n} \rangle \right) \right) \tag{4.66}
\]

**Observation noise precision, \( \lambda \)**

For \( q(\lambda \mid T) \) we get:

\[
\log(q(\lambda \mid T)) = \log \left( p(\lambda) \prod_{n=1}^{N} p(t_n \mid W, x_n, t, \lambda, u_n) \right) \tag{4.67}
\]

The result is the following Gamma distribution:

\[
q(\lambda \mid T) = \mathcal{G}(\lambda \mid \alpha_{\lambda}, \beta_{\lambda}) \tag{4.68}
\]
where
\[
\alpha_\lambda = a_\lambda + \frac{ND}{2} \tag{4.69}
\]
\[
\beta_\lambda = b_\lambda + \frac{1}{2} \sum_{n=1}^{N} \left( (t_n - \mathbf{W}x_n - \bar{t})^T \text{diag}(u_n)(t_n - \mathbf{W}x_n - \bar{t}) \right) \tag{4.70}
\]

with
\[
\langle (t_n - \mathbf{W}x_n - \bar{t})^T \text{diag}(u_n)(t_n - \mathbf{W}x_n - \bar{t}) \rangle = \sum_{r=1}^{D} (u_{r,n}) \left( t_{r,n}^2 + \text{trace}(\langle \mathbf{w}_{r:r}^T \mathbf{w}_{r:r} \rangle \langle \mathbf{x}_n \mathbf{x}_n^T \rangle) + \langle \bar{f}_r^2 \rangle \right.
\]
\[
-2t_{r,n} (\langle \mathbf{w}_{r:r} \rangle \langle \mathbf{x}_n \rangle + \langle \bar{f}_r \rangle) + 2 \langle \mathbf{w}_{r:r} \rangle \langle \mathbf{x}_n \rangle \langle \bar{f}_r \rangle \right) \tag{4.71}
\]

**Observation mean, \( \bar{t} \)**

For \( q(\bar{t} \mid T) \) we get:
\[
\log(q(\bar{t} \mid T)) = \log \left( p(\bar{t}) \prod_{n=1}^{N} p(t_n \mid \mathbf{W}, \mathbf{x}_n, \bar{t}, \lambda, \mathbf{u}_n) \right) \tag{4.72}
\]

The result is the following Gaussian distribution:
\[
q(\bar{t} \mid T) = \mathcal{N}(\bar{t}, \mu_\bar{t}, \Sigma_\bar{t}) \tag{4.73}
\]

where
\[
\Sigma_\bar{t} = \left( \mathbf{V}_\bar{t}^{-1} + \langle \lambda \rangle \sum_{n=1}^{N} \text{diag}(\langle u_n \rangle) \right)^{-1} \tag{4.74}
\]
\[
\mu_\bar{t} = \Sigma_\bar{t} \langle \lambda \rangle \sum_{n=1}^{N} \text{diag}(\langle u_n \rangle) (t_n - \langle \mathbf{W} \rangle \langle \mathbf{x}_n \rangle) \tag{4.75}
\]

### 4.4. Summary

A summary of the approximate posterior distributions for each of the model parameters is shown on page 88. Each approximate posterior distribution is dependent on the expected values of one or more of the others, so a closed-form algebraic solution cannot be obtained. We may arrive at a solution by initialising the required expectations (from maximum likelihood Gaussian PPCA for \( \mathbf{W}, \mathbf{X} \) and \( \lambda \) and setting each \( u_{r,n}, z_{q,n} \) and \( \delta_q \) to 1, and \( v \\) and each \( d_q \) to 1000) and then iteratively updating the estimate for each hyperparameter based on the current estimates of the values on which it depends, until convergence. The required current expectations are obtained using the standard expressions (see appendix...
A): and, from (4.71),

\[
(t_n - Wx_n - \bar{t})^T \text{diag}(u_n)(t_n - Wx_n - \bar{t})
\]

\[
= \sum_{r=1}^{D} (u_{r,n})^2 (t_{r,n}^2 + \text{trace}(w_{r,n}^T w_{r,n}) + \langle \bar{t}_r^2 \rangle)
\]

\[
-2t_{r,n}(\langle w_{r,n} \rangle \langle x_n \rangle + \langle \bar{t}_r \rangle) + 2\langle w_{r,n} \rangle \langle x_n \rangle \langle \bar{t}_r \rangle
\]

(4.76)

Having formulated this model, straightforward adjustments can be made to change the assumptions about the distributions of the noise. If all the elements in the latent variables take the value 1, then the Student-t distribution for the observation noise, which may be expressed as \(\int \mathcal{N}(e_{r,n} | 0, (\lambda u_{r,n})^{-1}) \mathcal{G}(u_{r,n} | \frac{v}{2}, \frac{n}{2}) du_{r,n}\), becomes just \(\mathcal{N}(e_{r,n} | 0, \lambda^{-1})\) and the observation noise assumption is now Gaussian. Similarly, setting all the elements in the latent variables for the latent variable noise (in the \(z_n\)) to 1 converts the latent variable noise assumption to Gaussian. Thus an implementation of this model is easily able to support the four possible Student-t/Gaussian combinations of noise assumptions.

\[
\langle w_{r,:} \rangle = \mu_{w_r}
\]

(4.77)

\[
\langle w_{r,:}^T w_{r,:} \rangle = \Sigma_{w_r} + \mu_{w_r}^T \mu_{w_r}
\]

(4.78)

\[
\langle u_{r,q}^2 \rangle = \langle w_{r,:}^T w_{r,:} \rangle_{q,q}
\]

(4.79)

\[
\langle \delta_q \rangle = \alpha_\delta / \beta_\delta
\]

(4.80)

\[
\langle \lambda \rangle = \alpha_\lambda / \beta_\lambda
\]

(4.81)

\[
\langle u_{r,n} \rangle = \alpha_u / \beta_{u_{r,n}}
\]

(4.82)

\[
\langle \log(u_{r,n}) \rangle = \psi(\alpha_u) - \log(\beta_{u_{r,n}})
\]

(4.83)

\[
\langle x_n \rangle = \mu_{x_n}
\]

(4.84)

\[
\langle x_n x_n^T \rangle = \Sigma_{x_n} + \mu_{x_n}^T \mu_{x_n}
\]

(4.85)

\[
\langle x_{q,n}^2 \rangle = \langle x_n x_n^T \rangle_{q,q}
\]

(4.86)

\[
\langle \bar{t} \rangle = \mu_{\bar{t}}
\]

(4.87)

\[
\langle \bar{t} \bar{t}^T \rangle = \Sigma_{\bar{t}} + \mu_{\bar{t}} \mu_{\bar{t}}^T
\]

(4.88)

\[
\langle \bar{t}_r^2 \rangle = \langle \bar{t} \bar{t}^T \rangle_{r,r}
\]

(4.89)

\[
\langle z_{q,n} \rangle = \alpha_z / \beta_{z_{q,n}}
\]

(4.90)

\[
\langle \log(z_{q,n}) \rangle = \psi(\alpha_z) - \log(\beta_{z_{q,n}})
\]

(4.91)

\[
\langle v \rangle = \alpha_v / \beta_v
\]

(4.92)

\[
\langle d_q \rangle = \alpha_d / \beta_{d_q}
\]

(4.93)

where \(\psi(\cdot)\) is the digamma function. Also required are, from (4.63),

\[
\langle (t_r - w_{r,:} x_n - \bar{t}_r)^2 \rangle = t_{r,n}^2 + \text{trace}(w_{r,:}^T w_{r,:}) + \langle \bar{t}_r^2 \rangle
\]

\[
-2t_{r,n}(\langle w_{r,:} \rangle \langle x_n \rangle + \langle \bar{t}_r \rangle) + 2\langle w_{r,:} \rangle \langle x_n \rangle \langle \bar{t}_r \rangle
\]

(4.94)
4. Robust variational PPCA

Summary of robust, variational PPCA

\[ q(x_n | T) = \mathcal{N}(x_n | \mu_{x_n}, \Sigma_{x_n}) \] (4.95)

\[ \Sigma_{x_n}^{-1} = \text{diag}(\langle \delta \rangle) \text{diag}(\langle z_n \rangle) + \langle \lambda \rangle \sum_{r=1}^{D} \langle u_{r,n} \rangle (w_{r,1}^T, w_{r,1}) \] (4.96)

\[ \mu_{x_n} = \Sigma_{x_n} \langle \lambda \rangle \sum_{r=1}^{D} \langle u_{r,n} \rangle (w_{r,1})^T (t_{r,n} - \langle \mu_r \rangle) \] (4.97)

\[ q(u_{r,n} | T) = \mathcal{G}(u_{r,n} | \alpha_u, \beta_{u_{r,n}}) \] (4.98)

\[ \alpha_u = ((v) + 1)/2 \] (4.99)

\[ \beta_{u_{r,n}} = \frac{1}{2} \left( (v) + ((t_{r,n} - w_{r,1}x_n - \bar{t}_r)^2) \right) \] (4.100)

\[ q(z_{q,n} | T) = \mathcal{G}(z_{q,n} | \alpha_z, \beta_{z_{q,n}}) \] (4.101)

\[ \alpha_z = ((q) + 1)/2 \] (4.102)

\[ \beta_{z_{q,n}} = \frac{1}{2} ((q) + (x_{q,n}^2)) \] (4.103)

\[ q(w_{r,1} | T) = \mathcal{N}(w_{r,1} | \mu_{w_{r,1}}, \Sigma_{w_{r,1}}) \] (4.104)

\[ \Sigma_{w_{r,1}}^{-1} = \text{diag}(\langle \delta \rangle) + \langle \lambda \rangle \sum_{n=1}^{N} \langle u_{r,n} \rangle (x_n x_n^T) \] (4.105)

\[ \mu_{w_{r,1}} = \Sigma_{w_{r,1}} \langle \lambda \rangle \sum_{n=1}^{N} \langle u_{r,n} \rangle (x_n^T) (t_{r,n} - \langle \bar{t}_r \rangle) \] (4.106)

\[ q(\delta_q | T) = \mathcal{G}(\delta_q | \alpha_{\delta}, \beta_{\delta_q}) \] (4.107)

\[ \alpha_{\delta} = a_{\delta} + D/2 \] (4.108)

\[ \beta_{\delta_q} = b_{\delta} + \frac{1}{2} \sum_{r=1}^{D} \langle w_{r,q}^2 \rangle + \frac{1}{2} \sum_{n=1}^{N} \langle z_{q,n} \rangle (x_{q,n}^2) \] (4.109)

\[ q(\bar{t} | T) = \mathcal{N}(\bar{t} | \mu_{\bar{t}}, \Sigma_{\bar{t}}) \] (4.110)

\[ \Sigma_{\bar{t}}^{-1} = V_{\bar{t}}^{-1} + \langle \lambda \rangle \sum_{n=1}^{N} \text{diag}(\langle u_n \rangle) \] (4.111)

\[ \mu_{\bar{t}} = \Sigma_{\bar{t}} \langle \lambda \rangle \sum_{n=1}^{N} \text{diag}(\langle u_n \rangle) (t_n - \langle W \rangle (x_n)) \] (4.112)

\[ q(\lambda | T) = \mathcal{G}(\lambda | \alpha_{\lambda}, \beta_{\lambda}) \] (4.113)

\[ \alpha_{\lambda} = a_{\lambda} + ND/2 \] (4.114)

\[ \beta_{\lambda} = b_{\lambda} + \frac{1}{2} \sum_{n=1}^{N} (\langle t_n - Wx_n - \bar{t} \rangle^T \text{diag}(u_n) (t_n - Wx_n - \bar{t})) \] (4.115)

\[ q(v | T) = \mathcal{G}(v | \alpha_v, \beta_v) \] (4.116)

\[ \alpha_v = a_v + ND/2 \] (4.117)

\[ \beta_v = b_v - \frac{1}{2} \left( ND + \sum_{n=1}^{N} \sum_{r=1}^{D} (\langle \log(u_{r,n}) \rangle - \langle u_{r,n} \rangle) \right) \] (4.118)

\[ q(d_q | T) = \mathcal{G}(d_q | \alpha_d, \beta_{d_q}) \] (4.119)

\[ \alpha_d = a_d + N/2 \] (4.120)

\[ \beta_{d_q} = b_d - \frac{1}{2} \left( N + \sum_{n=1}^{N} (\langle \log(z_{q,n}) \rangle - \langle z_{q,n} \rangle) \right) \] (4.121)

The variables’ distributions were calculated in the order shown above.
A further simple change may be made so that ARD only controls the variance of the components in \( \mathbf{W} \) and not the latent variables: the \( \text{diag}(\langle \delta \rangle) \text{diag}(\langle \mathbf{z}_n \rangle) \) term in the calculation of \( \mathbf{\Sigma}_{x_n} \) (4.36) may be replaced with \( \text{diag}(\langle \mathbf{z}_n \rangle) \) and the \( \frac{1}{2} \sum_{n=1}^{N} \langle z_{q,n} \rangle^2 \langle x_{q,n}^2 \rangle \) term removed altogether from the calculation of \( b_{d_q} \) (4.44). This enables a direct comparison to be made between these two ARD formulations to determine the effect of applying ARD to the latent variables.

### 4.5. Convergence criterion

Convergence of the variational algorithm may be monitored either by calculating the lower bound at each iteration, or by checking whether the parameter estimates form a consistent set. The lower bound is expensive to calculate and, particularly for the later models in this thesis, added expense in terms of computation time needs to be avoided. Therefore convergence is monitored through the parameter estimates. We have observed that the parameters close to the observations in the graphical model (see figure 4.4) converge fastest and so convergence is determined by monitoring the degrees of freedom variables, \( v \) and \( d_q \).

### 4.6. Illustration: synthetic data

In order to demonstrate the performance of robust variational PPCA, synthetic datasets were generated which are known to conform to the model being tested. Once values for the dimensionality variables had been selected, i.e. the number of samples, \( N \), the dimensionality of the observations in \( \mathbf{t}_n \), \( D \), and the dimensionality of the latent variables in \( \mathbf{x}_n \), \( Q \) which is less than \( D \), and values for other model parameters, \( \mu \), \( \lambda \), \( d \) and \( v \), had been selected, the synthetic data was constructed according to (4.1).

\( \mathbf{W} \) was generated as a random, orthonormal \( D \times Q \) matrix, each \( \mathbf{x}_n \) was sampled from the Student-t distribution \( S(\mathbf{x}_n \mid \mathbf{0}, \mathbf{I}, \text{diag}(\mathbf{d})) \), each \( \mathbf{e}_n \) was sampled from the Student-t distribution \( S(\mathbf{e}_n \mid \mathbf{0}, \lambda \mathbf{I}, v) \), and each observation \( \mathbf{t}_n \) was then calculated using (4.1). Selecting large values for the degrees of freedom causes the associated Student-t distribution to become more Gaussian-like.

Using synthetic datasets generated in this way, the first illustration demonstrates that the model estimates the the actual parameter variables effectively, the second illustration demonstrates the estimation of the model order (i.e. the number of components in \( \mathbf{W} \), or \( Q \)), and the third illustration shows that the new model is more robust to outliers than the fully Gaussian model.
4. Robust variational PPCA

![Boxplot of estimated vs. actual observation noise precision, \( \lambda \). The red boxes show the mean and one standard deviation; the blue lines show the full range of estimated values across the 600 synthetic datasets.](image)

**4.6.1. Illustration 1: parameter estimation**

With \( N = 200, D = 5 \) and \( Q = 3 \), datasets were generated with every combination of \( \lambda \) in the set \{0.01, 0.1, 1, 10, 100, 1000\}, \( d \) in the range of integers between 1 and 10 (the same value for each latent variable) and \( v \) in the range of integers between 1 and 10, with different, randomly-generated \( W \) and \( \bar{t} \) for each dataset. The new robust variational PPCA model was trained for 2000 iterations against each of the resulting 600 datasets with the model order set to the known value of \( Q \).

Figure 4.5 compares the estimated observation noise precision, \( \lambda \), with the actual values used to generate the datasets. There is a tendency to underestimate large precisions and overestimate small ones, but the relationship between estimated and actual is approximately linear.

Figure 4.6 compares the estimated and actual observation noise degrees of freedom, \( v \). The points have been coloured according to the the actual \( \lambda \) value used to generate the data and a small amount of lateral scatter introduced for visualisation. The heavy black lines indicate equality between estimates and actuals. For small values of actual \( v \), where the variance of the distribution is effectively infinite (\( v \leq 2 \)) there is a reasonably good correspondence, with a tendency to over-estimate \( v \). For larger values of \( v \) the results depend very much on the actual noise precision: where there is high precision, i.e. small noise, the estimate tends to indicate a Gaussian distribution, while for low precision the estimate tends to indicate a more heavy-tailed distribution than was actually used. With a small, finite set of data, a few values far from the mean can be accounted for either by increasing the degrees of freedom (making the distribution heavier-tailed) or by decreasing the precision (making the distribution wider); it is not clear what causes either of these alternatives to be adopted.

Looking at figure 4.6 there seems to be an upper limit to the posterior expectation of \( v \),
Figure 4.6. Plot of estimated vs. actual observation noise degrees of freedom, $v$. Each point is coloured according to the actual observation noise precision and a small amount of lateral scatter has been introduced for visualisation. The heavy black lines indicate equality between estimates and actuals.

given by $\alpha_v / \beta_v$ where, from (4.65) and (4.66),

$$\alpha_v = a_v + \frac{ND}{2}$$

and

$$\beta_v = b_v - \frac{1}{2} \left( ND + \sum_{n=1}^{N} \sum_{r=1}^{D} (\langle \log(u_{r,n}) \rangle - \langle u_{r,n} \rangle) \right)$$

The largest value for $\langle v \rangle$ is achieved when $\beta_v$ is smallest, i.e. when the summation term in (4.123) is maximised. This occurs when all the $u_{r,n}$ are 1 giving the summation the total value of $-ND$. So the smallest possible value for $\beta_v$ is $b_v$. Hence the upper limit for $\langle v \rangle$ is $\frac{a_v + ND/2}{b_v}$.

The comparison of estimated and actual latent variables’ degrees of freedom, $d$, in figure 4.7 is similarly coloured according to the the actual $\lambda$ values and a small amount of lateral scatter has been introduced to aid visualisation. For each test dataset the $d_q$ have been sorted such that (a) shows the $d_q$ associated with the column of $W$ that has the greatest magnitude while (d) shows the $d_q$ associated with the column of $W$ that has the smallest magnitude. Hence (d) displays the $d_q$ associated with the component and latent variable that is most likely to have been switched off by ARD (remembering that the true model order is 3); in nearly every case the estimated $d_q$ is large, though there are clearly some datasets for which the model order is being overestimated. For those latent variables that are not switched off, for small values of actual $d_q$, where the variance of the distribution is effectively infinite ($d_q \leq 2$) there is a reasonably good correspondence, with a tendency to over-estimate the value. For larger values of $d_q$ the results depend very much on the actual noise precision: where there is high precision the estimate has a tendency to be close to the actual, while for low precision the estimate has a tendency to indicate a more heavy-tailed distribution than was actually used.

As previously stated, with a small, finite set of data, a few values far from the mean can be accounted for either by increasing the degrees of freedom (making the distribution
heavier-tailed) or by decreasing the precision (making the distribution wider). Measuring the linear correlation between the error in the estimate of $\lambda$ and the error in the estimates of each of the $d_q$ (ordered as per figure 4.7) for the test datasets with large noise (actual $\lambda$ is 0.01) gives values of -0.38, -0.60, -0.68 and -0.66 respectively, showing that the model is balancing these two options.

Figure 4.8 shows a plot of estimated against actual values for the mean observation, $\bar{t}$. There is generally a good, linear correlation between the two, with most of the estimated values for low noise datasets (in red) being closer to the actuals than those from high noise datasets (in blue). However, there are a small number of low noise values which are significantly different from the actuals.

The results for the estimates of $W$ are similarly better where there is low noise and worse where there is high noise. As described in section 2.5.1, the model is only able to estimate the components in $W$ up to a sign, permutation and rotation, so the Procrustes transformation is used to transform the estimate in the direction of the actual value.
Figure 4.8. Plot of estimated vs. actual mean observation values, $\bar{t}$. The black line indicates equality. Each point is coloured according to the actual observation noise precision.

Figure 4.9 shows an example of a $W$ where the model’s estimate is close to the actual. The figure contains four Hinton diagrams: (a) the actual $W$ used to generate the data; (b) the estimate produced by the model; (c) the estimate produced by the model after a Procrustes transformation has been applied; (d) the Procrustes transformation matrix.

The correlations between each column of the actual $W$ and the corresponding column of the transformed estimated matrix are good. The shape of the Procrustes transformation matrix indicates that virtually no rotation has been required; each row and column contains only one sizeable value. The large values are on the main diagonal, so no permutation has been involved, and the large squares are not all white, so some sign change has also been involved, as can be seen by comparing figures 4.9(a) and (b). The reason why the rotational ambiguity seems to have been “cured” may be surmised by inspecting the shape of the joint probability densities of two independent random variables (see figure 4.1). Where the variables are Gaussian the density is circular and there is no cause for preferring one direction over another. However, where the variables are Student-t regions of higher density extend along the axes, enabling the model to “lock into” the correct orientation. This will only happen where there are sufficient samples and the degrees of freedom of the Student-t distribution is small enough for this effect to be significant.

Figure 4.10 shows an example of a $W$ where the model’s estimate is not close to the actual. The figure contains four Hinton diagrams, as before. The estimated matrix obviously bears no resemblance to the actual, either before the transformation or after it. This does not mean that the model is unable to represent the data, just that it has found a different
4. Robust variational PPCA

Figure 4.9. Hinton diagrams comparing the estimated and actual mixing matrix, \( W \), where the estimate is good. (a) the actual \( W \) used to generate the data; (b) the estimate produced by the model; (c) the estimate produced by the model after a Procrustes transformation has been applied; (d) the Procrustes transformation matrix.

Figure 4.10. Hinton diagrams comparing the estimated and actual mixing matrix, \( W \), where the estimate is bad. (a) the actual \( W \) used to generate the data; (b) the estimate produced by the model; (c) the estimate produced by the model after a Procrustes transformation has been applied; (d) the Procrustes transformation matrix.
4. Robust variational PPCA

Figure 4.11. Two different views of the ARD parameters for each of the four components in the synthetic data. In both cases the ARD parameters for each dataset have been sorted by increasing value. In (a) the mean and one standard deviation are shown in red and the full range of values in blue. In (b) each point is coloured according to the actual observation noise precision.

representation of it from the actual structure used to generate it.

4.6.2. Illustration 2: model order estimation

The previous experiments were repeated, but instead of setting the model order to its known value ($Q$), it was set to the maximum allowable value ($D - 1$) to demonstrate the ability of ARD to estimate the model order. Figure 4.11 shows two different plots of the ARD parameter values $\delta_q$ over all the datasets. Within each dataset the $\delta_q$ have been sorted into increasing order to take account of the fact that the order of the columns of $W$ may be different from the actuals (due to the potential permutation ambiguity). Plot (a) shows the mean, one standard deviation and the full range of values for each of the 600 datasets. There is a significant tendency for three components to be switched on (the true model order) and the fourth to be switched off. Plot (b) shows each value coloured, as before, according to the amount of observation noise. It is noticeable that the fourth component is tending to get switched on for the high noise datasets and switched off for those with low noise, indicating that the extra component is being used to model the noise.

It was suggested in section 4.4 that a slight modification to the algorithm may be made to remove the dependence of the latent variables on the ARD parameter variables $\delta_q$. This enables a direct comparison to be made between the model where ARD controls the precisions of both $W$ and $x_n$, with a model where ARD controls only the precisions of $W$. A synthetic dataset was generated with $N = 200$ observations, $D = 15$ observation dimensions and $Q = 8$ latent variable dimensions (the true model order). The observation noise was generated with precision $\lambda = 100$ and degrees of freedom $v = 2$ and the latent variables with precision 1 and degrees of freedom $d_q = 2$. Thus both the observation noise and latent variables are significantly heavy-tailed. Figure 4.12 shows the $W$ estimated by each model after 5000 iterations. The model with ARD over both $W$ and $x_n$ has estimated the model order to be the true model order of 8, while the model with ARD over only $W$
4. Robust variational PPCA

![Figure 4.12](image1.png)

(a) ARD over $\mathbf{W}$ and $\mathbf{x}_n$

(b) ARD over $\mathbf{W}$ only

Figure 4.12. The estimated $\mathbf{W}$ calculated by the two different ARD models. In (a) the true model order of 8 has been found, while in (b) two additional components are switched on.

![Figure 4.13](image2.png)

(a) ARD over $\mathbf{W}$ and $\mathbf{x}_n$

(b) ARD over $\mathbf{W}$ only

Figure 4.13. The estimated $\delta_q$ calculated by the two different ARD models as the iterative process progresses. Large values indicate components that have been switched off by ARD. Note the logarithmic ordinate scale. In (a) the values converge smoothly, while in (b) the model has not converged and has taken a long time to switch off unwarranted components.

has switched on two further components.

Figure 4.13 shows plots of the values of the ARD variables $\delta_q$ as the iterative process progresses. It is clear that the “$\mathbf{W}$ and $\mathbf{x}_n$” model has converged smoothly. The “only $\mathbf{W}$” model has not converged and also takes much longer to switch off unwarranted components.

The “$\mathbf{W}$ and $\mathbf{x}_n$” model is slightly slower in performing each iteration since the calculations for the $\delta_q$ must include terms dependent on the latent variables whereas in the “$\mathbf{W}$ only” model the calculations are based only on the components of $\mathbf{W}$. However this effect is more than offset by the speed of convergence of the former.

4.6.3. Illustration 3: robustness to outliers

The new model, with its Student-t distributions over the observation noise and the latent variables, should be robust to outliers that occur in the $e_{r,n}$ and $x_{q,n}$ variables. This
Figure 4.14. For the datasets with no outliers and the datasets containing each of the four types of outliers, are plotted the mean and one standard deviation of the absolute difference between the estimated and actual values for (a) the mean observation $\bar{t}$ and (b) the observation noise precision $\lambda$. The most robust models are those with a mean value close to zero and a small standard deviation. Note that the most robust model appears to be that with Gaussian observation noise and Student-t latent variables.

Robustness is illustrated by generating synthetic datasets which contain one of four different types of outlier:

**type 1:** a single element of an observation is an outlier due to the corresponding observation noise element $e_{r,n}$ being an outlier,

**type 2:** a whole observation is an outlier due to the observation noise $e_n$ being an outlier,

**type 3:** a whole observation is an outlier due to one element in the corresponding latent variable vector $x_{q,n}$ being an outlier, and

**type 4:** a whole observation is an outlier due to the corresponding vector of latent variable $x_n$ being an outlier.

Ten datasets were generated with Gaussian distributed observation noise and latent variables. Each of these datasets was then copied four times and three outliers of the four types (one type per dataset) were added. The magnitude of the outliers was five times the magnitude of the largest value in the dataset. Against each of these 50 datasets, four different models were trained:

**GG:** Gaussian observation noise and Gaussian latent variables,

**GS:** Gaussian observation noise and Student-t latent variables (see also an alternative formulation by [Tipping and Lawrence, 2005]),

**SG:** Student-t observation noise and Gaussian latent variables, and

**SS:** Student-t observation noise and Student-t latent variables.
Figure 4.15. The model order for the outlier tests as calculated by robust variational PPCA. The actual model order is marked as a black dotted line.

SS model

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>no outliers</td>
<td>446</td>
<td>421</td>
<td>211</td>
<td>344</td>
<td>392</td>
<td>267</td>
<td>440</td>
<td>116</td>
<td>296</td>
<td>371</td>
</tr>
<tr>
<td>type 1 outliers</td>
<td>399</td>
<td>369</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>435</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>type 2 outliers</td>
<td>513</td>
<td>475</td>
<td>428</td>
<td>442</td>
<td>490</td>
<td>482</td>
<td>508</td>
<td>181</td>
<td>467</td>
<td>480</td>
</tr>
<tr>
<td>type 3 outliers</td>
<td>506</td>
<td>487</td>
<td>436</td>
<td>452</td>
<td>449</td>
<td>476</td>
<td>475</td>
<td>427</td>
<td>467</td>
<td>444</td>
</tr>
<tr>
<td>type 4 outliers</td>
<td>490</td>
<td>463</td>
<td>375</td>
<td>425</td>
<td>431</td>
<td>435</td>
<td>471</td>
<td>318</td>
<td>358</td>
<td>427</td>
</tr>
</tbody>
</table>

SG model

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>no outliers</td>
<td>445</td>
<td>420</td>
<td>200</td>
<td>342</td>
<td>390</td>
<td>266</td>
<td>439</td>
<td>116</td>
<td>295</td>
<td>370</td>
</tr>
<tr>
<td>type 1 outliers</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>type 2 outliers</td>
<td>380</td>
<td>378</td>
<td>338</td>
<td>301</td>
<td>254</td>
<td>364</td>
<td>449</td>
<td>267</td>
<td>439</td>
<td>368</td>
</tr>
<tr>
<td>type 3 outliers</td>
<td>378</td>
<td>324</td>
<td>316</td>
<td>318</td>
<td>373</td>
<td>307</td>
<td>399</td>
<td>3</td>
<td>370</td>
<td>380</td>
</tr>
<tr>
<td>type 4 outliers</td>
<td>517</td>
<td>495</td>
<td>369</td>
<td>469</td>
<td>452</td>
<td>406</td>
<td>509</td>
<td>357</td>
<td>411</td>
<td>442</td>
</tr>
</tbody>
</table>

Table 4.1. Observation noise degrees of freedom $v$ calculated for each of the outlier test datasets for (top) the model with Student-t observation and latent variables and (bottom) the model with Student-t observation noise and Gaussian latent variables.
A robust model should make good estimates of the mean ($\tilde{t}$) and precision ($\lambda$) of the observations whether or not there are outliers present. Figure 4.14 summarises the results of this experiment. For the datasets with no outliers and the datasets containing each of the four types of outliers, are plotted the mean and one standard deviation of the absolute difference between the estimated values and the values from which the data were generated. The most robust models are thus those with a mean close to zero and a small standard deviation. It is interesting to note that the most robust model appears to be that with Gaussian observation noise and Student-t latent variables.

It is instructive to look at the model orders estimated in each case (see figure 4.15). In the models where the observation noise is assumed to be Student-t (SS and SG), where the outliers lie in single observation noise elements, it is often the case that the estimated model order is zero, i.e. all the components have been switched off. These models have estimated $\tilde{t}$ and then all variation has been absorbed into the noise. Table 4.1 compares the observation noise degrees of freedom ($v$) estimates obtained for those models which incorporate Student-t assumptions. Note how the degrees of freedom are depressed when this type of outlier is present.

### 4.7. Robust variational FA

The PPCA-based models are easily extended to similar models based on Factor Analysis (FA). This thesis is primarily concerned with models based on PPCA and Probabilistic Canonical Correlation Analysis, so no experimental results will be shown for models based on FA, but the associated FA-based models are described for completeness.

If, instead of assuming that the observation noise across all observed variables is from the same distribution, it is assumed that it is different for each row of $\mathbf{T}$, i.e. that $p(e_{r,n}) = S(e_{r,n} | 0, \lambda_r^{-1}, v_r)$, then we arrive at a robust variational Factor Analysis model, with

\[
q(u_{r,n} | \mathbf{T}) = \mathcal{G}(u_{r,n} | \alpha_u, \beta_{u_{r,n}}) \quad (4.124)
\]

\[
\alpha_u = (\langle v_r \rangle + 1)/2 \quad (4.125)
\]

\[
\beta_{u_{r,n}} = \frac{1}{2} \left( \langle v_r \rangle + \langle (t_{r,n} - w_r : x_n - \tilde{t}_r)^2 \rangle \right) \quad (4.126)
\]

\[
q(v_r | \mathbf{T}) = \mathcal{G}(v_r | \alpha_v, \beta_{v_r}) \quad (4.127)
\]

\[
\alpha_v = a_v + N/2 \quad (4.128)
\]

\[
\beta_{v_r} = b_v - \frac{1}{2} \left( N + \sum_{n=1}^N \langle \log(u_{r,n}) \rangle - \langle u_{r,n} \rangle \right) \quad (4.129)
\]

\[
q(\lambda_r | \mathbf{T}) = \mathcal{G}(\lambda_r | \alpha_\lambda, \beta_{\lambda_r}) \quad (4.130)
\]

\[
\alpha_\lambda = a_\lambda + N/2 \quad (4.131)
\]

\[
\beta_{\lambda_r} = b_\lambda + \frac{1}{2} \sum_{n=1}^N \langle (t_{r,n} - w_r : x_n - \tilde{t}_r)^Tu_{r,n}(t_{r,n} - w_r : x_n - \tilde{t}_r) \rangle \quad (4.132)
\]
and all other posteriors as shown on page 88.

4.8. Conclusions

Chapter 3 introduces the new PPCA-AR model and describes an expectation maximisation algorithm for calculating the maximum likelihood estimates for the model parameters. A variational Bayesian model has the advantage that the model parameter are considered to be random variables and posterior probability distributions are calculated for them. As a first step towards a robust, variational Bayesian version of PPCA-AR, this chapter has described a novel variational Bayesian version of PPCA which is made robust to outliers by the assumption that both the observation noise and the latent variables are Student-t distributed, rather than Gaussian as they are in standard PPCA, and which uses ARD to estimate the model order.

It has been shown that the new model is generally good at estimating the actual parameters used to generate synthetic data, especially where the observation noise is low compared with the underlying data. Where the data are heavy-tailed, the rotational ambiguity associated with the components in the estimated $\mathbf{W}$ seems to have been ameliorated; this is attributed to the shape of the joint distribution of Student-t variables (the result of employing ARD over the components), where the probability mass is concentrated close to zero values of the variables rather than in regions where all the variables are non-zero, thus enabling the model to “lock into” the correct orientation.

The model is generally good as estimating the model order, except where there are significant outliers in individual elements of the observation noise, where it tends to estimate a mean value in $\bar{\mathbf{t}}$ and then absorb all variation into the Student-t observation noise distribution by making it very heavy tailed. Using the same set of ARD parameter variables $\delta_q$ to control the precisions of the components in $\mathbf{W}$ and the corresponding latent variables in $\mathbf{x}_n$ has been shown to be beneficial for the convergence of the model.

In the maximum likelihood PPCA-AR model explicit steps are taken to ensure that the components in $\mathbf{W}$ are orthogonal, in keeping with PPCA. In this robust, variational PPCA there is nothing explicit ensuring that the components are orthogonal. However, the prior distribution for each component is independent (each row of $\mathbf{W}$ is assigned the prior distribution $\mathcal{N}(\mathbf{w}_{r,:} | \mathbf{0}, \text{diag}(\delta))$), leading to Gaussian posterior distributions for each $\mathbf{w}_{r,:}$ that have diagonal covariance matrices, thus preserving the orthogonality.

From a computational perspective this robust, variational version of PPCA requires greater resources than the traditional maximum likelihood one. More memory is required for both the extra variables (particularly the new latent variables) and for the extra covariance matrices, and the model order should be started at its maximum possible value of $D - 1$ to afford ARD the opportunity to switch off those for which there is no evidence in the data. The extra memory could be somewhat reduced by realising that the conditional independence of many variables leads to diagonal posterior covariance matrices and both
memory and computation could be scaled back over the iterations by removing components (and the corresponding latent variables) that ARD has completely switched off.

This chapter has introduced a robust, variational PPCA model that captures spatial relationships between variables, but ignores temporal dependencies. The next chapter looks at temporal modelling using autoregression.
5. Robust variational autoregression

In chapter 3 it was shown that equipping Probabilistic Principal Component Analysis (PPCA) with temporal coupling over the latent variables using autoregressive (AR) processes is beneficial in the modelling of multivariate time-series data. To make the PPCA-AR model robust to outliers and able to model leptokurtic data, the excitation noise of the autoregressive (AR) processes will be assumed to be Student-t distributed. In this chapter a robust variational Bayesian version of AR is described, which incorporates automatic relevance determination (ARD) [Mackay, 1994; Neal, 1995] to estimate the model order. This results in posterior probability distributions for each of the model’s parameter variables from which a measure of the certainty of the predictions (given the observed data) are obtained.

AR models each observation in an ordered series as a linear combination of the \( P \) previous observations, to which is added an element of excitation noise from a random innovations process. An AR model of order \( P \) is defined as

\[
x_n = \sum_{p=1}^{P} \theta_p x_{n-p} + \epsilon_n
\]  

(5.1)

where \( x_n \) is the \( n \)th observation (of \( N \)) in \( x \), the \( \theta_p \) are the autoregressive coefficients and \( \epsilon_n \) is the excitation noise. Using (5.1) recursively to write \( x_n \) in terms of the innovations process shows that an AR model may also be viewed as a finite impulse response filter of the innovations.

5.1. Robust variational AR

Traditionally the excitation noise is presumed to be Gaussian distributed, which, due to the linearity of the AR model, means that the observations are also Gaussian distributed. In order to model heavy-tailed time series, this robust AR model assumes that the excitation noise is drawn from a Student-t distribution with precision \( \kappa \) and \( d \) degrees of freedom:

\[
p(\epsilon_n \mid 0, \kappa, d) = S(\epsilon_n \mid 0, \kappa, d)
\]  

(5.2)

Some of the material in this chapter has been published as [Christmas and Everson, 2011].
Using (2.47), this may be rewritten as a scale mixture of Gaussians, all sharing a common mean:

\[ p(\epsilon_n \mid 0, \kappa, d) = \int_0^\infty \mathcal{N}(\epsilon_n \mid 0, (\kappa z)^{-1}) \mathcal{G}(z \mid d/2, d/2) \, dz \]  
(5.3)

5.1.1. Priors

An alternative way of expressing the AR model shown in (5.1) is (following Ó Ruanaidh and Fitzgerald [1996])

\[ x = L \theta + \epsilon \]  
(5.4)

where \( L \) is the \( N \) by \( P \) matrix whose \( n \)th row contains the lags for element \( x_n \), i.e. \( (x_{n-1}, \ldots, x_{n-p}) \) and \( \epsilon \) is the vector of excitations (note that in this case the \( x \) (and \( L \)) represents observations and not, as in other chapters, latent variables; this notation is used for reasons of standardisation through subsequent chapters). Combining this with the excitation noise distribution (5.2) allows the data likelihood to be written as

\[ p(x \mid \theta, \kappa, d) = S(x \mid L \theta, \kappa, d) \]  
(5.5)

Expressing the Student-t as a scale mixture of Gaussians, this may be rewritten as

\[ p(x \mid \theta, \kappa, z) = \mathcal{N}(x \mid L \theta, (\kappa \text{diag}(z))^{-1}) \]  
(5.6)

\[ p(z_n \mid d) = \mathcal{G}(z_n \mid d/2, d/2) \]  
(5.7)

where the \( z_n \) are latent variables modifying the precision of the Gaussian mixture for each observation and \( \text{diag}(z) \) is the diagonal matrix with the \( z_n \) arranged along the diagonal.

For the AR coefficients we seek a sparse solution in which only those coefficients \( \theta_i \) for which there is support in the data are non-zero, by placing an ARD prior over each of the \( \theta_p \):

\[ p(\theta) = \prod_{p=1}^{P} \mathcal{N}(\theta_p \mid 0, \gamma_p) = \mathcal{N}(\theta \mid 0, \text{diag}(\gamma)). \]  
(5.8)

Just as the ARD precisions control the magnitude of the components in PPCA, each ARD precision, \( \gamma_p \), controls the magnitude of the associated AR coefficients, so that if \( \gamma_p \) is large \( \theta_p \) is effectively switched off. We again place common Gamma priors over the ARD precisions:

\[ p(\gamma_p) = \mathcal{G}(\gamma_p \mid a_\gamma, b_\gamma). \]  
(5.9)

If \( a_\gamma \) and \( b_\gamma \) are chosen to have the same value then this is effectively a Student-t prior over the elements of \( \theta \) (c.f. 2.47).

A Gamma prior is specified for the precision \( \kappa \):

\[ p(\kappa) = \mathcal{G}(\kappa \mid a_\kappa, b_\kappa) \]  
(5.10)

Finally, specification of the model is completed by assigning a Gamma prior to the degrees
of freedom $d$:

$$p(d) = \mathcal{G}(d | a_d, b_d) \quad (5.11)$$

In the absence of detailed prior information, uninformative priors were selected, setting $a_\lambda = b_\lambda = a_d = b_d = a_\gamma = b_\gamma = 10^{-3}$.

Figure 5.1 summarises the Bayesian AR model and the interdependencies between the model’s parameter variables.

### 5.1.2. Factorised variational Bayesian method

The joint probability of the data and parameters $\Omega = \{z, \theta, \kappa, d, \gamma\}$ may be factorised as

$$p(x, \Omega) = p(x, z, \theta, \kappa, d, \gamma) \quad (5.12)$$

$$= p(x | \theta, \kappa, z) p(\theta | \gamma) p(\kappa) p(z | d) p(d) p(\gamma). \quad (5.13)$$

Following the prior factorisation, the approximate posterior is factorised as:

$$q(\Omega | x) = q(\theta, \kappa, z, d, \gamma | x) \quad (5.14)$$

$$= q(\theta | x) q(\kappa | x) q(z | x) q(d | x) q(\gamma | x) \quad (5.15)$$

$$= q(\theta | x) q(\kappa | x) \left[ \prod_{n=1}^{N} q(z_n | x) \right] q(d | x) q(\gamma | x). \quad (5.16)$$

Using the factorised variational Bayes method to obtain approximate posterior distributions for the factorisation (5.14) and using the joint probability (5.12), we consider each group in turn.
5. Robust variational autoregression

AR coefficients, $\theta$

The approximate posterior for the AR coefficients that maximises the negative free energy, $\mathcal{F}(q)$, is given by

$$
\log(q(\theta | x)) = E_{q(\theta)}[\log(p(x | \theta, \kappa, z) p(\theta | \gamma) p(\kappa) p(z | d) p(d) p(\gamma))] 
$$

(5.17)

where $E_{q(a)}[b]$ denotes the expectation of $b$ taken with respect to the approximate posteriors of all variables except $a$. Expanding this and moving all terms not dependent on $\theta$ into a single constant term we get

$$
\log(q(\theta | x)) = E_{q(\theta)}[\log(p(x | \theta, \kappa, z)] + \log(p(\theta | \gamma)) + \text{const}
$$

(5.18)

Again, expanding this and moving all terms not dependent on $\theta$ into the constant term:

$$
\log(q(\theta | x)) = -\frac{1}{2}(\kappa)(x - L\theta)^T \text{diag}(z)(x - L\theta) - \frac{1}{2}\theta^T \text{diag}(\gamma)\theta + \text{const}
$$

(5.20)

where $\langle f(a) \rangle$ denotes the expectation of $f(a)$ with respect to $a$. Since (5.20) is quadratic in $\theta$, it can be seen that $q(\theta | x)$ is a Gaussian and we have

$$
q(\theta | x) = \mathcal{N}(\theta | \mu_{\theta}, \Sigma_{\theta})
$$

(5.21)

where

$$
\Sigma_{\theta}^{-1} = \langle \kappa \rangle L^T \text{diag}(z)L + \text{diag}(\gamma)
$$

(5.22)

$$
\mu_{\theta} = \langle \kappa \rangle \Sigma_{\theta} L^T \text{diag}(z)x
$$

(5.23)

Excitation noise precision, $\kappa$

Applying the same procedure for $\kappa$ we obtain a Gamma posterior distribution for the excitation noise precision:

$$
q(\kappa | x) = \mathcal{G}(\kappa | \alpha_\kappa, \beta_\kappa)
$$

(5.24)

where

$$
\alpha_\kappa = a_\kappa + \frac{N}{2}
$$

(5.25)

$$
\beta_\kappa = b_\kappa + \frac{1}{2}x^T \text{diag}(z)(x - 2L\theta) + \frac{1}{2} \sum_{n=1}^{N} \langle z_n \rangle \langle (L_n\theta)^2 \rangle
$$

(5.26)

and $L_n$ is a the row vector containing the lags for $x_n$. The final term may be simplified as follows:

$$
\langle (L_n\theta)^2 \rangle = \langle L_n\theta \theta^T L_n^T \rangle = L_n\langle \theta \theta^T \rangle L_n^T
$$

(5.27)
Latent variables, $z$

It is not analytically tractable to locate a joint distribution for $z$, so examining each $z_n$ individually (and dropping the constant term to aid readability):

$$
\log( q(z_n | x) )
= \frac{1}{2} \log(z_n) - \frac{1}{2} \langle \kappa \rangle \mathbb{E}_{z_n}[\left( x_n - L_n \theta \right)^T z_n (x_n - L_n \theta)] + \mathbb{E}_{z_n}[\log(G(z_n | d/2, d/2))]
$$

$$
= \left( \frac{\langle d \rangle + 1}{2} - 1 \right) \log(z_n) - \frac{1}{2} \langle \kappa \rangle \langle (x_n - L_n \theta)^T (x_n - L_n \theta) \rangle + \frac{\langle d \rangle}{2} z_n
$$

(5.28)

(5.29)

where $L_n$ is the $n$th row of $L$. On inspection we see that $q(z_n | x)$ is a Gamma distribution:

$$
q(z_n | x) = G(z_n | \alpha_z, \beta_z)
$$

(5.30)

where

$$
\alpha_z = \frac{\langle d \rangle + 1}{2}
$$

(5.31)

$$
\beta_z = \frac{\langle d \rangle}{2} + \frac{1}{2} \langle \lambda \rangle \langle (x_n - L_n \theta)^2 \rangle
$$

(5.32)

with

$$
\langle (x_n - L_n \theta)^2 \rangle = x_n^2 - 2x_n L_n \langle \theta \rangle + L_n \langle \theta \theta^T \rangle L_n^T
$$

(5.33)

As the expected value of $d$ becomes large, so that the Student-t distribution of the excitation noise approaches a Gaussian, the posterior expected value of $z_n$ (i.e. $\alpha_z/\beta_z$) tends to 1 and likelihood of $x$ tends towards the Gaussian $\mathcal{N}(x | L \theta, \kappa^{-1})$.

Degrees of freedom, $d$

For $q(d | x)$ the result, using Stirling’s approximation, is exactly the same as for robust, variational PPCA (see section 4.3.2), which is the Gamma distribution

$$
q(d | x) = G(d | \alpha_d, \beta_d)
$$

(5.34)

where

$$
\alpha_d = a_d + \frac{N}{2}
$$

(5.35)

$$
\beta_d = b_d - \frac{1}{2} \left( N + \sum_{n=1}^{N} (\langle \log(z_n) \rangle - \langle z_n \rangle) \right)
$$

(5.36)

ARD precisions, $\gamma$

Finally, the posterior distributions for the ARD precisions are found as:

$$
q(\gamma_p | x) = G(\gamma_p | \alpha_{\gamma_p}, \beta_{\gamma_p})
$$

(5.37)
where
\[
\alpha_\gamma = a_\gamma + 1 \quad (5.38)
\]
\[
\beta_{\gamma p} = b_\gamma + \frac{1}{2}\langle \theta^2_p \rangle \quad (5.39)
\]

5.2. Summary

The box on page 108 shows a summary of the approximate posterior distributions for each of the model parameters. Each approximate posterior distribution is dependent on the expected values of one or more of the others, so a closed-form algebraic solution cannot be obtained. We may arrive at a set of solutions by initialising the required expectations (setting each \( \theta_p, z_n \) and \( \gamma_p \) to 1, \( d \) to 100 and \( \kappa \) to the prior \( a_\kappa/b_\kappa \)) and then iteratively updating the estimate for each hyperparameter based on the current estimates of the values on which it depends, until convergence. The required expectations are obtained using the standard expressions (see appendix A):

\[
\langle \theta \rangle = \mu_\theta \quad (5.40)
\]
\[
\langle \kappa \rangle = \alpha_\kappa/\beta_\kappa \quad (5.41)
\]
\[
\langle z_n \rangle = \alpha_z/\beta_{zn} \quad (5.42)
\]
\[
\langle \log(z_n) \rangle = \psi(\alpha_z) - \log(\beta_{zn}) \quad (5.43)
\]
\[
\langle d \rangle = \alpha_d/\beta_d \quad (5.44)
\]
\[
\langle \gamma_p \rangle = \alpha_\gamma/\beta_{\gamma p} \quad (5.45)
\]

where \( \psi(\cdot) \) is the digamma function.

The equivalent expressions calculated for a Gaussian AR model are as follows:

\[
q(\theta | x) = \mathcal{N}(\theta | \mu_\theta, \Sigma_\theta) \quad (5.46)
\]
\[
\Sigma_\theta = ((\kappa)\mathbf{L}^T\mathbf{L} + \text{diag}(\langle \gamma \rangle))^{-1} \quad (5.47)
\]
\[
\mu_\theta = \langle \kappa \rangle \Sigma_\theta \mathbf{L}^T x \quad (5.48)
\]
\[
q(\kappa | x) = \mathcal{G}(\kappa | \alpha_\kappa, \beta_\kappa) \quad (5.49)
\]
\[
\alpha_\kappa = a_\kappa + \frac{N}{2} \quad (5.50)
\]
\[
\beta_\kappa = b_\kappa + \frac{1}{2}(x^T x - 2x^T \mathbf{L} \langle \theta \rangle + \mathbf{L} \langle \theta^T \rangle \mathbf{L}^T) \quad (5.51)
\]
\[
q(\gamma_p | x) = \mathcal{G}(\gamma_p | \alpha_\gamma, \beta_{\gamma p}) \quad (5.52)
\]
\[
\alpha_\gamma = a_\gamma + 1 \quad (5.53)
\]
\[
\beta_{\gamma p} = b_\gamma + \frac{1}{2}\langle \theta^2_p \rangle \quad (5.54)
\]

In the Student-t AR case as \( d \) tends to infinity and the excitation sequence becomes effectively Gaussian it can be seen that expressions (5.46) to (5.54) are recovered from (5.55) to (5.69).
5. Robust variational autoregression

Summary of robust, variational AR

\[ q(z_n | x) = \mathcal{G}(z_n | \alpha_z, \beta_{z_n}) \]  \hspace{1cm} (5.55)
\[
\alpha_z = \frac{\langle d \rangle + 1}{2} \]  \hspace{1cm} (5.56)
\[
\beta_{z_n} = \frac{\langle d \rangle}{2} + \frac{1}{2} \langle \kappa \rangle \left( x_n^2 - 2x_n \langle \theta \rangle + \langle \theta \theta^T \rangle \langle \theta \rangle^T \right) \]  \hspace{1cm} (5.57)
\[ q(d | x) = \mathcal{G}(d | \alpha_d, \beta_d) \]  \hspace{1cm} (5.58)
\[
\alpha_d = a_d + \frac{N}{2} \]  \hspace{1cm} (5.59)
\[
\beta_d = b_d - \frac{1}{2} \left( N + \sum_{n=1}^{N} (\langle \log(z_n) \rangle - \langle z_n \rangle) \right) \]  \hspace{1cm} (5.60)
\[
q(\theta | x) = \mathcal{N}(\theta | \mu_\theta, \Sigma_\theta) \]  \hspace{1cm} (5.61)
\[
\Sigma_\theta = (\langle \kappa \rangle \langle \theta \rangle \langle \theta \rangle^T \Sigma_\theta + \langle \gamma \rangle) \]  \hspace{1cm} (5.62)
\[
\mu_\theta = \Sigma_\theta \langle \theta \rangle + \langle \gamma \rangle \]  \hspace{1cm} (5.63)
\[ q(\gamma_p | x) = \mathcal{G}(\gamma_p | \alpha_{\gamma_p}, \beta_{\gamma_p}) \]  \hspace{1cm} (5.64)
\[
\alpha_{\gamma} = a_{\gamma} + 1 \]  \hspace{1cm} (5.65)
\[
\beta_{\gamma_p} = b_{\gamma} + \frac{1}{2} \langle \theta_p \rangle \]  \hspace{1cm} (5.66)
\[ q(\kappa | x) = \mathcal{G}(\kappa | \alpha_{\kappa}, \beta_{\kappa}) \]  \hspace{1cm} (5.67)
\[
\alpha_{\kappa} = a_{\kappa} + \frac{N}{2} \]  \hspace{1cm} (5.68)
\[
\beta_{\kappa} = b_{\kappa} + \frac{1}{2} \langle \theta \rangle \langle \theta \rangle \rangle \langle \theta \rangle^T \]  \hspace{1cm} (5.69)

The variables’ distributions were calculated in the order shown above.

5.3. Illustration: synthetic data

In order to demonstrate the performance of robust variational AR, synthetic datasets were generated which were known to conform to the model being tested. Once values for the dimensionality variables were selected, i.e. the number of samples, \(N\), and the model order \(P\), and values for other model parameters, \(\kappa\) and \(d\), were selected, the synthetic data was constructed according to (5.1), exactly as for the latent variables in PPCA-AR described in section 3.3. Selecting large values for the degrees of freedom caused the associated Student-t distribution to become more Gaussian-like.

Using synthetic datasets generated in this way, tests were performed to demonstrate that the new model

- estimates the model parameters effectively,
- estimates the model order effectively, and
- is more robust to outliers than the fully Gaussian model.
Figure 5.2. Gaussian and Student-t examples. Left: Comparison of the estimated posterior distributions for the excitation noise with the distribution used to generate the data. Centre: Estimated vs. actual AR coefficients. Right: Expected values of one-step-ahead predictions compared with the observations. N = 1500 and actual p = 10. Estimated values and variances are shown to 2 decimal places.

5.3.1. Illustration 1: parameter estimation

Figure 5.2a compares, for a dataset with N = 1500 and Gaussian excitation noise, the expected values of the variational posterior distributions with the actual values used to generate the data. Figure 5.2b shows results for observations generated with a very heavy-tailed distribution d = 0.5 excitation sequence. In both cases it is clear that the model accurately learns the coefficients and makes accurate one step ahead predictions despite the vastly different natures of the excitation sequences. The model orders, P, were 10 in both cases, but the model was trained with P = 20 in order to demonstrate the effect of ARD. The effects are clearly seen in the centre graphs where the θ_i values where i > P have been constrained to be close to zero and hence have been switched off.

The results in both examples are similar in that the variational posterior excitation noise distribution is more compact than the actual, a tendency reported by a number of authors (for example MacKay [2003], Wang and Titterington [2005], Consonni and Marin [2007]), the estimated θ values are similar to the actuals, with a tendency to be underestimated, and the reconstructions of the data are good.

The over-compactness of the estimated distribution of the excitation noise and the under-
estimation of the \( \theta \) coefficients warrants further investigation. To this end the model was trained against datasets which were created for every combination of \( \kappa \) and \( d \) between 0.01 and 10 in increments of 0.01, with \( N = 1500 \), \( P = 10 \) and a different, randomly-generated \( \theta \) in each case. The results, shown in figure 5.3, show that there is an approximately linear relationship between estimated and actual values of \( \kappa \), with the predicted value consistently over-estimated (and hence the excitation noise precision is over-estimated), while for \( d \) the relationship between estimated and actual values are non-linear and \( d \) is consistently under-estimated. It is this combination of over-estimated precision and under-estimated degrees of freedom that lead to the more compact distribution compared with the actual. The predicted and actual \( \theta \) values are highly correlated, but the plot appears twisted clockwise with respect to the black line, indicating that the magnitudes of the coefficients are slightly under-estimated.

With a relatively small number of samples it is highly unlikely that the data will be truly representative of the distribution from which it was generated. This is particularly true of a Student-t distribution whose degrees of freedom are such that the variance is infinite (i.e. \( d \leq 2 \)). This does not reduce the ability of the model to represent the data, but it does mean that the estimated parameter values are less likely to reflect the actuals.

### 5.3.2. Illustration 2: model order estimation

The effects of ARD in the model order estimation have been hinted at in the examples shown in section 5.3.1. This is now demonstrated in more detail by training the model against synthetic datasets of 1500 samples each for every combination of \( P \) from 1 to 15, and every \( \kappa \) and \( d \) in the integer range 1 to 10. The model was trained with \( P = 20 \). The prior for \( \theta \) is \( \mathcal{N}(\theta \mid 0, \text{diag}(\gamma)^{-1}) \); if the estimated value of one of the \( \theta_i \) is more than one standard deviation away from zero, i.e. \( \theta_i^2 > 1/\gamma_i \), then it is deemed it to be switched on. Figure 5.4 shows, in grey, which \( \theta_i \) are switched on in each of the 1500 test runs. The solid black lines indicate the actual model order in each case.

The lower model orders are well estimated, but the higher ones appear to be consistently under-estimated. This, however, is a consequence of the method used to generate the
5. Robust variational autoregression

Figure 5.4. The model was trained against datasets generated for all $\kappa$ and $d$ values in the integer range 1 to 10 and every $P$ from 1 to 15, giving 1500 datasets in all. Grey lines show which elements of $\theta$ were switched on in each test. Heavy black lines show the actual $P$ value for each test.

Figure 5.5. Plot of $\theta$ for 100 synthetic datasets with $P = 10$. ARD is constraining the estimated values where $i > P$ to be close to zero. Actual values where $i$ is close to $P$ tend to be small, so may be considered to be switched off.

AR coefficients: the scheme described in section 3.3 tends to produce $\theta_i$ that reduce in magnitude with increasing $i$. This is illustrated in figure 5.5 which compares the actual and estimated $\theta$ values for all the synthetic datasets generated with $P = 10$. The ARD mechanism is clearly suppressing $\theta_i$ when $i > P$, but it is also the case that there is often insufficient support in the data for the small $\theta_i$ (with $8 \lesssim i \lesssim 10$) so that they appear to be erroneously switched off and thus the model order is under-estimated.

5.3.3. Illustration 3: robustness to outliers

Using a similar method as described previously, a synthetic dataset of 500 samples was generated with Gaussian excitation noise ($\kappa = 10$) and $P = 10$. Student-t and Gaussian AR models were each trained against it, and, as figure 5.6a demonstrates, both models accurately make one-step-ahead reconstructions of the data.

Three outliers were then added to the dataset, each as positive values (i.e. in the same direction), with values 10 times the maximum size of the remaining samples. The Student-
Figure 5.6. One-step-ahead predictions plotted against actual values for (a) Gaussian data and (b) the same with the addition of 3 outliers. The black dotted lines indicate prediction = actual. The actual mean is marked with a black cross; the estimated mean with a black circle. The Student-t AR model is unaffected by the outliers; the predictions are still very good and the estimated mean coincides with the actual. The Gaussian AR model is noticeably less accurate than before: the estimated mean has moved away from the actual and the excitation noise variance is over-estimated (the plot appears twisted clockwise with respect to the diagonal).
Figure 5.7. Comparison of the 80% credibility intervals for the reconstructions of a 50-point extract of the EEG signal using an AR model with Student-t excitations (light shading) and one with Gaussian excitations (dark). The continuous black line indicates the observed signal. Credibility intervals are symmetric about the mean predictions, which for clarity are not shown. The Student-t AR model shows higher confidence than the Gaussian.

t and Gaussian AR models were each trained against this amended set. Figure 5.6b compares one-step-ahead predictions with actual values. It is not surprising that the $P$ samples immediately following each outlier are poorly predicted by the models, so these have been omitted from the plots. By comparing these with the equivalent graphs above them, it is clear that the Gaussian model has been significantly affected by the presence of the outliers, where the Student-t model is robust to them. The Gaussian AR predictions are noticeably worse than before (the points are spread away from the diagonal), the estimated mean has moved away from the actual and the noise variance has been underestimated (the plot appears twisted clockwise with respect to the diagonal). For the Student-t model the predictions do not appear to have deteriorated and the estimated mean has not moved noticeably away from the true mean.

The Gaussian AR model is forced to accommodate outliers within the single Gaussian distribution it fits to the excitation noise. This causes the mean of the estimated distribution to move away from the actual and/or the variance to be over-estimated; both of these effects are demonstrated here. While all of the distributions in the Student-t mixture of Gaussians have the same mean, their range of variances allows the overall distribution to accommodate the outliers.

### 5.4. Results: real data

EEG signals are often thought of as an example of data whose noise is heavier-tailed than Gaussian. If the data are considered to have been generated by an underlying autoregressive process, then we expect the sample precision to be the result of the innovations sequence (see (5.5)). As was shown in chapter 1 the EEG data are significantly non-Gaussian in nature.

Both the Student-t and Gaussian AR models were evaluated against this EEG signal. Where the degrees of freedom for a Student-t distribution is less than or equal to 2 the
5. Robust variational autoregression

variance is, effectively, infinite, which makes it impossible to make a direct comparison of the confidence it has in its predictions with the Gaussian AR model. Instead, for each observation, Monte Carlo sampling of 1000 predictions was used to generate an 80% credibility interval; this method was repeated for the Gaussian AR results to enable a direct comparison to be made. Figure 5.7 shows a subset of results for 50 observations. For each observation the Student-t interval is noticeably smaller and the actual observed value falls within it. In fact all of the 1150 observations lie within the 80% credibility interval of the Student-t AR model. This is not the case for the Gaussian AR model.

The Student-t model estimates the AR coefficients with high precision and a model order of approximately 12. Repeating the training of the model across all 58 channels that comprise a single observation set it is found that it identifies rather similar values in each case, whereas the Gaussian model does not. This result is demonstrated in figure 5.8. An important consequence of this is that power spectral densities calculated from the AR coefficients estimated with Student-t excitations are considerably more consistent across a subject test than estimates using Gaussian excitations.

5.5. Conclusions

In this chapter it has been shown that a Bayesian AR model based on the assumption that the excitation noise is Student-t distributed is more robust to outliers than the model where the noise is assumed to be Gaussian or a mixture of Gaussians. It has also been shown to be a generalisation of the Gaussian model, able to model data whose excitation noise is Gaussian distributed or heavier-tailed than the Gaussian distribution.

For an exact Bayesian model the Student-t assumption leads to intractable integrals in the calculation of the posterior densities for the model’s parameter variables. It has been shown that the factorised variational Bayes technique provides good approximations to those posterior distributions and is computationally efficient, but it tends to underestimate
the excitation noise variance, the degrees of freedom and the magnitudes of the model coefficients.

The incorporation of ARD priors over the AR coefficients results in sparse solutions that accurately predict the model order.

For real EEG data that is heavier-tailed than Gaussian, the Student-t model makes more accurate one-step-ahead predictions, with smaller variances, than the Gaussian model. It estimates the model order to be approximately 12 for these EEG data and provides a remarkably uniform set of AR coefficients for all 58 EEG signals in a single observation set, in contrast to the Gaussian model.

The connection between the AR coefficients and the power spectrum of the observations has long been recognised [Akaike, 1969] and exploited for the estimation of power spectra, but the variability of these estimates has been pointed out [Christini et al., 1995]. The consistency of the coefficients estimated by this Student-t AR model across EEG channels lead to much more consistent estimates of the power spectra.

The next chapter combines the robust, variational PPCA model defined in chapter 4 with the robust, variational AR model defined in this chapter to give a robust, variational version of PPCA-AR.
6. Robust variational PPCA with autoregression

Chapter 3 has introduced the new PPCA-AR model which captures temporal structure in Probabilistic Principal Component Analysis (PPCA) by applying independent autoregressive (AR) processes to each of the latent variables, and described an Expectation Maximisation (EM) algorithm for determining the maximum likelihood values for the model parameters. The E-step utilises the forward-backward algorithm to determine the posterior probability densities of the latent variables, which is tractable because Gaussian assumptions are made about the observation and excitation noise terms. While maximum likelihood PPCA-AR estimates the observation and excitation noise variances, and hence provides posterior probability distributions for the observations and latent variables, it provides only point estimates for each of the other parameters, which leads to the model underestimating the variance of the observations. The model has no automated method for estimating the model orders.

A fully Bayesian model provides probability distributions for each of the parameters (which are now regarded as random variables) and enables the incorporation of automatic relevance determination (ARD) [Mackay, 1994; Neal, 1995] to estimate the model orders, but the integrals required for an exact solution are intractable. As for robust PPCA and AR, described in chapters 4 and 5 respectively, the factorised variational Bayesian method may be used to calculate approximate posterior distributions for each of the parameter variables. This also enables the model to be modified to include robustness and the ability to model leptokurtic data, by modelling the noise terms using Student-t distributions rather than Gaussians.

In this new, robust version of PPCA-AR a mean observation parameter has been added to enable the modelling of sets of observations with a non-zero mean.

Let us start by restating the expressions that define PPCA-AR (from (3.10) and (3.11)): 

\[ \tilde{x}_n = \Theta \tilde{x}_{n-1} + \tilde{\epsilon}_n \] \hspace{1cm} (6.1)

\[ t_n = W \tilde{x}_n + \bar{t} + e_n \] \hspace{1cm} (6.2)

The \( N \) observations in the \( t_n \) are \( D \)-dimensional, the mean observation in \( \bar{t} \) enables the modelling of non-zero-mean data, the latent variables in the \( x_n \) are \( Q \)-dimensional and the model order of each of the AR processes is \( P \). The AR coefficients for the \( q \)th latent variable (i.e. the \( q \)th row of \( X \), where \( X \) is the matrix \( (x_1, \ldots, x_N) \)) are in \( \theta_q \). In the maximum likelihood model the observation noise, \( e_n \), is assumed to be sampled from the
Gaussian $\mathcal{N}(\epsilon_n | 0, \lambda^{-1}I)$ and the excitation noise, $\epsilon_n$, is assumed to be sampled from the Gaussian $\mathcal{N}(\epsilon_n | 0, \text{diag}(\kappa)^{-1})$. The AR model of order $P$ is converted to a first order model by stacking the latent variable lags into $\tilde{x}_n$ and constructing block-diagonal matrices for the $\theta_q$ and $\kappa_q^{-1}$ in $\Theta$ and $\Phi$ respectively (see section 3.1). Finally, $\tilde{W} = WF$ where $W$ is the mixing matrix and $F$ is the matrix of ones and zeros that extracts $x_n$ from the stacked $\tilde{x}_n$, such that $x_n = F\tilde{x}_n$.

6.1. Priors

Previously the Gaussian noise assumptions regarding the excitation and observation noise terms, in $\epsilon_n$ and $e_n$ respectively, led to Gaussian likelihoods for $\tilde{x}_n$ and $t_n$. In the interests of robustness these terms are now assumed to be Student-t distributed about a zero mean:

$$p(\epsilon_{q,n}) = S(\epsilon_{q,n} | 0, \kappa_q, d_q)$$

$$p(e_{r,n}) = S(e_{r,n} | 0, \lambda, v)$$

where $q$ indicates the $q$th latent variable, $r$ the $r$th observed variable and $n$ the $n$th sample. For the excitation noise each latent variable is assumed to have a different distribution, each centred on zero, but with potentially different precisions and degrees of freedom. The reasons for this will become clear shortly. The observation noise is assumed to have the same precision for each dimension, as it is for standard PPCA. If the noise for each dimension of the observations were to be modelled with different, independent Student-t distributions the results would be a model that is related to Factor Analysis (FA) (see section 6.5).

These noise assumptions lead to Student-t distributions for the likelihoods of the latent variables and observations:

$$p(\tilde{x}_n | \Theta, \tilde{x}_{n-1}, \kappa_q, d_q) = S(\tilde{x}_n | \Theta\tilde{x}_{n-1}, \kappa_q, d_q)$$

$$p(t_n | \tilde{W}, \tilde{x}_n, \tilde{t}, \lambda, v) = S(t_n | \tilde{W}\tilde{x}_n + \tilde{t}, \lambda, v)$$

As before there needs to be some method for fixing the relative scaling within the $\tilde{W}\tilde{x}_n$ term. In this case this is achieved by fixing the precisions of the latent variables, in $\kappa_q$, to 1.

Once again the Student-t distributions may be expressed as scale mixtures of Gaussians. With Gamma priors over the degrees of freedom variables ($d_q$ and $v$) and a Gamma over the observation noise precision ($\lambda$), this leads to

**Student-t distribution for observation noise:**

$$p(t_n | W, x_n, \tilde{t}, u_n, \lambda) = \mathcal{N}(t_n | \tilde{W}\tilde{x}_n + \tilde{t}, (\lambda \text{diag}(u_n))^{-1})$$

$$p(u_{r,n} | v) = \mathcal{G}(u_{r,n} | \frac{v}{2}, \frac{v}{2})$$

$$p(v) = \mathcal{G}(v | a_v, b_v)$$
\[ p(\lambda) = \mathcal{G}(\lambda \mid a_\lambda, b_\lambda) \quad (6.10) \]

**Student-t distribution for excitation noise:**

\[ p(x_1 \mid z_1) = \mathcal{N}(x_1 \mid m_0, \text{diag}(z_1)^{-1}) \quad (6.11) \]
\[ p(\tilde{x}_n \mid \Theta, \tilde{x}_{n-1}, \tilde{z}_n) = \mathcal{N}(\tilde{x}_n \mid \Theta \tilde{x}_{n-1}, \text{diag}(\tilde{z}_n)^{-1}) \quad \text{for } n > 1 \quad (6.12) \]
\[ p(z_{q,n} \mid d_q) = \mathcal{G}(z_{q,n} \mid d_q^2, d_q^2) \quad (6.13) \]
\[ p(d_q) = \mathcal{G}(d_q \mid a_d, b_d) \quad (6.14) \]

Note that the first vector of latent variables, in \( x_1 \), is assigned a different prior to the others as there is no temporally previous vector for it to depend on. As in chapters 4 and 5 two new sets of latent variables, \( u_n \) and \( z_n \), have been introduced and \( \tilde{z}_n \) is the stacked version of \( z_n \), given by \( F^T z_n \).

The mean observation, \( \bar{t} \), is assigned a Gaussian prior:

\[ p(\bar{t}) = \mathcal{N}(\bar{t} \mid m_{\bar{t}}, V_{\bar{t}}) \quad (6.15) \]

As with robust variational PPCA, the columns of \( W \) are assigned ARD priors controlled by the hyper-parameter precisions in \( \delta \):

\[ p(w_{r:} \mid \delta) = \mathcal{N}(w_{r:} \mid m_w, \text{diag}(\delta)^{-1}) \quad (6.16) \]
\[ p(\delta_q) = \mathcal{G}(\delta_q \mid a_{\delta}, b_{\delta}) \quad (6.17) \]

and, as for robust variational AR, each coefficient in \( \theta_q \) (for the \( q \)th latent variable) is assigned an ARD prior controlled by the hyper-parameter precisions in \( \gamma_q \):

\[ p(\theta_q \mid \gamma_q) = \mathcal{N}(\theta_q \mid m_{\theta_q}, \text{diag}(\gamma_q)^{-1}) \quad (6.18) \]
\[ p(\gamma_{p,q}) = \mathcal{G}(\gamma_{p,q} \mid a_{\gamma}, b_{\gamma}) \quad (6.19) \]

In robust variational PPCA (chapter 4) the ARD priors over the columns of \( W \) are also applied to the associated latent variables. This is not the case here; the rows of \( X \) are controlled only by the AR processes. Where there is no evidence in the data for a latent variable, all the AR coefficients for that variable are constrained towards zero, thus forcing it to be switched off.

If the parameters of the excitation noise distributions were the same for each row of \( X \), then this switching off would have an unwarranted effect on the estimates of the precision and degrees of freedom for the shared distribution. Hence the different degrees of freedom, \( d_q \), for each latent variable.

The graphical model for this system is shown in figure 6.1. Comparing it with the equivalent diagrams for robust variational PPCA, in figure 4.4, and robust variational AR, in figure 5.1, it may be seen that the dependencies for those parameters shown in blue (\( W, \bar{t}, \delta, \lambda, u_n \) and \( v \)) have been directly inherited from robust variational PPCA while those shown in red (\( \theta_q, \gamma_q, z_n \) and \( d_q \)) have been directly inherited from robust variational
AR. The variational derivations for the approximate posterior distributions are therefore likewise inherited. Some amendments to the inferences for the AR coefficients (the $\theta_q$) and latent variables ($z_n$) are required as they are now dependent on variables rather than observations as they were in chapter 5. The only place where entirely new derivations are required is for the latent variables.

In the absence of detailed prior information, uninformative priors are selected: $m_W = m_J = m_\theta = 0$, $V_t = 10^3$ and $a_v = b_v = a_\lambda = b_\lambda = a_\delta = b_\delta = a_\gamma = b_\gamma = a_d = b_d = 10^{-3}$.

Beal and Ghahramani [2001] describe a variational EM algorithm for the linear dynamical system whereby the E-step uses exact inference through the forward-backward algorithm to calculate posterior distributions for the latent variables, based on the posterior expectations of the other parameter variables. The posterior expectations of the latent variables are then used in the M-step, which uses variational Bayes to calculate the approximate posterior distributions for the other parameter variables. This approach is tractable because the latent variables are assumed to be Gaussian distributed. Beal [2003] describes a slightly different variational EM algorithm: the E-step is determined as before and then the parameter variables (and combinations of parameter variables) are replaced by their expectations, as calculated in the M-step.

In the robust PPCA-AR model the latent variables are assumed to have been generated from autoregressive processes with Student-t excitation noise, leading to distributions for the latent variables that are also Student-t (6.5). This assumption makes exact inference for the forward-backward algorithm intractable, but we may use the factorised variational Bayesian technique to approximate the posterior distributions. The EM algorithm is no longer required as this approximation becomes another step in the variational Bayesian iterative process.
This robust variational version of the forward-backward algorithm may be considered to be a generalisation of the standard algorithm in that it relaxes the constraint for the latent variables to be Gaussian distributed, though exact inference has now been replaced by approximations. The forward sweep may be thought of as a robust (approximate) Kalman filter and the backward sweep as a robust (approximate) Kalman smoother.

6.1.1. Robust variational forward-backward algorithm

In the maximum likelihood algorithm, the posterior distributions of the latent variables are calculated in two sweeps. The forward sweep calculates, for each \( n \) in increasing order from 1 up to \( N \),

\[
\alpha(\tilde{x}_n) = p(\tilde{x}_n \mid t_1, \ldots, t_n).
\]

The backward sweep combines \( \alpha(\tilde{x}_n) \) with \( \beta(\tilde{x}_n) = p(\tilde{x}_n \mid t_{n+1}, \ldots, t_N) \), starting at \( n = N \) and working back down to 1. Thus the posterior probability of \( \tilde{x}_n \) is given by

\[
\alpha(\tilde{x}_n) \beta(\tilde{x}_n) = p(\tilde{x}_n \mid t_1, \ldots, t_N) \tag{6.20}
\]

where \( T = (t_1, \ldots, t_N) \). The \( \alpha(\tilde{x}_n) \) and the resulting posterior probability are defined recursively as

\[
\alpha(\tilde{x}_n) = p(t_n \mid \tilde{x}_n) \int p(\tilde{x}_n \mid \tilde{x}_{n-1}) \alpha(\tilde{x}_{n-1}) \, d\tilde{x}_{n-1} \tag{6.21}
\]

\[
p(\tilde{x}_n \mid T) = \alpha(\tilde{x}_n) \int p(\tilde{x}_{n+1} \mid \tilde{x}_n) \, d\tilde{x}_{n+1} \tag{6.22}
\]

Forward sweep

In the fully Bayesian model we must introduce into (6.21) further conditioning of the latent variables on all the other parameters in the model. With \( \Omega \) representing the set containing all the other parameters, (6.21) must be rewritten as

\[
\alpha(\tilde{x}_n) = \int \left[ p(t_n \mid \tilde{x}_n, \Omega) \int p(\tilde{x}_n \mid \tilde{x}_{n-1}, \Omega) \alpha(\tilde{x}_{n-1}) \, d\tilde{x}_{n-1} \right] p(\Omega) \, d\Omega \tag{6.23}
\]

Applying the variational method, \( \alpha(\tilde{x}_n) \) is replaced by its variational approximation, \( q^{(\alpha)}(\tilde{x}_n) \), and we may write

\[
\log(q^{(\alpha)}(\tilde{x}_n)) = \mathbb{E}_{\tilde{x}_n} \left[ \log \left( p(t_n \mid \tilde{x}_n, \Omega) p(\tilde{x}_n \mid \tilde{x}_{n-1}, \Omega) q^{(\alpha)}(\tilde{x}_{n-1}) p(\Omega) \right) \right] \tag{6.24}
\]

where \( \mathbb{E}_{\tilde{x}_n}[b] \) denotes the expectation of \( b \) taken with respect to the approximate posteriors of all variables except \( a \).

One of the assumptions that makes the variational approach tractable is that the approximate posterior distributions are conditionally independent. This allows \( q^{(\alpha)}(X) \) to be expressed as \( \prod_{n=1}^N q^{(\alpha)}(\tilde{x}_n) \) and thus \( \tilde{x}_{n-1}, \tilde{x}_n \) and \( \tilde{x}_{n+1} \) to be treated as independent variables when conditioned on \( T \). Moving all terms in (6.24) not dependent on \( \tilde{x}_n \) into a
single constant term gives

$$\log(q^{(\alpha)}(\tilde{x}_n)) = E_{\tilde{x}_n} \left[ \log \left( p(t_n | \tilde{x}_n, \Omega) p(\tilde{x}_n | \tilde{x}_{n-1}, \Omega) \right) \right] + \text{const} \quad (6.25)$$

It would appear at first glance that the recursive nature of the forward sweep has been broken as the $q^{(\alpha)}(\tilde{x}_{n-1})$ term in (6.24) does not include any dependency on $\tilde{x}_n$ and so has been absorbed into the constant term. However, the remaining expression does contain terms dependent on expectations of $\tilde{x}_{n-1}$ with respect to its approximate posterior distribution, so the link to previous instances of $\tilde{x}_n$ is, in fact, maintained.

Following this definition through the usual variational procedure, with $\bar{W}$ defined as $WF$ and $\bar{z}_n$ as $F^Tz_n$ ($F$ being the fixed matrix of zeros and ones that extracts $x_n$ from $\tilde{x}_n$, as before) and dropping the constant term to aid readability we obtain

$$\log(q^{(\alpha)}(\tilde{x}_n)) = E_{\tilde{x}_n} \left[ \log \left( p(t_n | W, x_n, \tilde{f}, \lambda, u_n) p(\tilde{x}_n | \Theta, \tilde{x}_{n-1}, z_n) \right) \right] \quad (6.26)$$

$$= E_{\tilde{x}_n} \left[ \log \left( \mathcal{N}(t_n | \bar{W}\tilde{x}_n + \bar{f}, \text{diag}(\lambda u_n)^{-1})\mathcal{N}(\tilde{x}_n | \Theta\tilde{x}_{n-1}, \text{diag}(\bar{z}_n)^{-1}) \right) \right] \quad (6.27)$$

$$= -\frac{1}{2} E_{\tilde{x}_n} \left[ \tilde{x}_n^T \left( \lambda W^T \text{diag}(u_n) W + \text{diag}(\bar{z}_n) \right) \tilde{x}_n \right. \right.

$$\left. \left. - 2\tilde{x}_n^T \left( \lambda W^T \text{diag}(u_n)(t_n - \bar{f}) + \text{diag}(\bar{z}_n)\Theta\tilde{x}_{n-1} \right) \right] \quad (6.28)$$

This may be recognized as the Gaussian $\mathcal{N}(\tilde{x}_n | \tilde{\mu}_n^{(\alpha)}, \tilde{\Sigma}_n^{(\alpha)})$, where

$$\tilde{\Sigma}_n^{(\alpha)} = \left( \langle \lambda \rangle \langle \tilde{W}^T \text{diag}(u_n) \tilde{W} \rangle + \text{diag}(\bar{z}_n) \right)^{-1} \quad (6.29)$$

$$\tilde{\mu}_n^{(\alpha)} = \tilde{\Sigma}_n^{(\alpha)} \left( \langle \lambda \rangle \langle \tilde{W}^T \text{diag}(u_n) \rangle (t_n - \bar{f}) + \text{diag}(\bar{z}_n) \Theta \tilde{x}_{n-1} \right) \quad (6.30)$$

and $\langle f(a,b) \rangle$ denotes the expectation of $f(a,b)$ with respect to the approximate posterior distributions of $a$ and $b$.

A slightly different formulation is required for $\tilde{x}_1$ as it has a different prior due to its position at the beginning of the time sequence. Expression (6.23) becomes

$$\alpha(\tilde{x}_1) = \int p(t_1 | \tilde{x}_1, \Omega) p(\tilde{x}_1 | \Omega) p(\Omega) \, d\Omega \quad (6.31)$$

and following through the same variational process results in the Gaussian $\mathcal{N}(\tilde{x}_1 | \tilde{\mu}_1^{(\alpha)}, \tilde{\Sigma}_1^{(\alpha)})$, where

$$\tilde{\Sigma}_1^{(\alpha)} = \left( \langle \lambda \rangle \langle \tilde{W}^T \text{diag}(u_1) \tilde{W} \rangle + \text{diag}(\bar{z}_1) \right)^{-1} \quad (6.32)$$

$$\tilde{\mu}_1^{(\alpha)} = \tilde{\Sigma}_1^{(\alpha)} \left( \langle \lambda \rangle \langle \tilde{W}^T \text{diag}(u_1) \rangle (t_1 - \bar{f}) + \text{diag}(\bar{z}_1) \mu_0 \right) \quad (6.33)$$

These calculations are dependent on the expectation $\langle \tilde{W}^T \text{diag}(u_n) \tilde{W} \rangle$, which may be rewritten as $F^T(W^T \text{diag}(u_n) W)F$. Taking each element of the matrix individually, we
may write

\[ \langle W^T \text{diag}(u_n) W \rangle_{i,j} = \sum_{r=1}^{D} \langle w_{r,n} w_{r,i} w_{r,j} \rangle \]

(6.34)

\[ \sum_{r=1}^{D} \langle w_{r,n} \rangle \langle w_{r,i} w_{r,j} \rangle \]

(6.35)

Since the rows of \( W \) are posteriorly independent, this simplifies to

\[ \langle W^T \text{diag}(u_n) W \rangle_{i,j} = \left\{ \begin{array} {c} \sum_{r=1}^{D} \langle u_{r,n} \rangle \langle w_{r,i} w_{r,j} \rangle & \text{if } i = j \\ \sum_{r=1}^{D} \langle u_{r,n} \rangle \langle w_{r,i} w_{r,j} \rangle & \text{if } i \neq j \end{array} \right\} \]

(6.36)

**Backward sweep**

Following a similar procedure as for the forward sweep, in the fully Bayesian formulation (6.22) must be rewritten as

\[
p(\tilde{x}_n | T) = \alpha(\tilde{x}_n) \int \left[ \int p(\tilde{x}_{n+1} | \tilde{x}_n, \Omega) d\tilde{x}_{n+1} \right] p(\Omega) d\Omega \tag{6.37}
\]

Applying the variational method we may write this as

\[ \log(q(\tilde{x}_n | T)) = E_{\tilde{x}_n} \left[ \log \left( q^{(\alpha)}(\tilde{x}_n) p(\tilde{x}_{n+1} | \tilde{x}_n, \Omega) p(\Omega) \right) \right] \tag{6.38} \]

Moving all terms not dependent on \( \tilde{x}_n \) into a single constant term gives

\[ \log(q(\tilde{x}_n | T)) = E_{\tilde{x}_n} \left[ \log \left( q^{(\alpha)}(\tilde{x}_n) p(\tilde{x}_{n+1} | \tilde{x}_n, \Omega) \right) \right] + \text{const} \tag{6.39} \]

The recursion appears in terms dependent on the expectation of \( \tilde{x}_{n+1} \) with respect to its approximate posterior distribution, so the link to subsequent instances of \( \tilde{x}_n \) is maintained.

Following this definition through the usual variational procedure, again dropping the constant term to aid readability:

\[ \log(q(\tilde{x}_n | T)) = E_{\tilde{x}_n} \left[ \log \left( q^{(\alpha)}(\tilde{x}_n) p(\tilde{x}_{n+1} | \tilde{x}_n, \Omega) \right) \right] \tag{6.40} \]

\[ = E \left[ \log \left( p(\tilde{x}_n | \tilde{x}_{n+1}, \Omega) p(\tilde{x}_{n+1} | \Theta, \tilde{x}_n, z_{n+1}) \right) \right] \]

\[ = E \left[ \log \left( \mathcal{N}(\tilde{x}_n | \tilde{\mu}_n, \tilde{\Sigma}_n) \mathcal{N}(\tilde{x}_{n+1} | \Theta \tilde{x}_n, \text{diag}(z_{n+1})^{-1}) \right) \right] \tag{6.41} \]

\[ = -\frac{1}{2} E \left[ \tilde{x}_n^T \left( \tilde{\Sigma}_n^{(\alpha)} \right)^{-1} + \Theta^T \text{diag}(z_{n+1}) \Theta \right] \tilde{x}_n \]

\[ -2\tilde{x}_n^T \left( \tilde{\Sigma}_n^{(\alpha)} \right)^{-1} \tilde{\mu}_n + \Theta^T \text{diag}(z_{n+1}) \tilde{x}_n \]

(6.42)

This may be recognized as the Gaussian \( \mathcal{N}(\tilde{x}_n | \tilde{\mu}_n, \tilde{\Sigma}_n) \), where

\[ \tilde{\Sigma}_n = \left( \left( \tilde{\Sigma}_n^{(\alpha)} \right)^{-1} + \Theta^T \text{diag}(z_{n+1}) \Theta \right)^{-1} \]

(6.43)

\[ \tilde{\mu}_n = \tilde{\Sigma}_n \left( \left( \tilde{\Sigma}_n^{(\alpha)} \right)^{-1} \tilde{\mu}_n + \Theta^T \text{diag}(z_{n+1}) \tilde{x}_n \right) \]

(6.44)
For the last instance of the latent variables in the time series, i.e. where \( n = N \), the forward sweep results in \( \alpha(\hat{x}_N) \) which is defined as \( p(\hat{x}_N | t_1, \ldots, t_N) \). This is just \( p(\hat{x}_n | T) \), the posterior probability for \( \hat{x}_n \). So where \( n = N \) the posterior distribution is the Gaussian \( p(\hat{x}_N | T) = \mathcal{N}(\hat{x}_N | \mu_N^{(\alpha)}, \Sigma_N^{(\alpha)}) \).

These calculations are dependent on the expectation \( \langle \Theta^T \text{diag}(\tilde{z}_n+1) \Theta \rangle \). If we consider each latent variable separately, then this expectation for the \( q \)th latent variable and \( n \)th sample is \( \langle \theta_q \theta_q^T \rangle \langle z_{q,n}+1 \rangle \). The full \( \langle \Theta^T \text{diag}(\tilde{z}_n+1) \Theta \rangle \) may be constructed as a block diagonal matrix of the individual matrices for each \( q \).

The posterior distribution for \( x_n \) may easily be extracted from that for \( \hat{x}_n \) using the \( F \) matrix, giving \( q(x_n | T) = \mathcal{N}(x_n | F \hat{\mu}_n, F \hat{\Sigma}_n F^T) \).

Inspecting (6.30) and (6.44), it may be seen that the expectation of \( \hat{x}_n \) is made up of three elements: the prediction from the previous time-step \( (\Theta \mu_{n-1} - \tilde{t}) \), an adjustment proportional to the current observation brought back into the latent space \( (\tilde{W}^T (t_n - \bar{t})) \) and an adjustment proportional to the prediction from the next time-step \( (\Theta^T \mu_{n+1}) \). The adjustments are weighted according to the relative magnitudes of the predicted precisions, so that more weight is given to the element that has greater certainty.

### 6.1.2. AR coefficients, \( \theta_q \)

In chapter 5 the lag vector associated with a single latent variable value was represented by a row of the lag matrix \( L \), which was constructed from the observations. In PPCA-AR it is more convenient to represent the lags by the \( \tilde{x}_n \) vectors. These are no longer observations, but variables in the inference procedure. Hence the lags will be represented as expectations with respect to approximate posterior distributions when included in the calculations of other variables in the factorised variational Bayesian process.

The expression for a single \( \theta_q \) (from (5.19)) may be rewritten with the change of notation for the PPCA-AR model, and dropping the constant term for readability, as

\[
\log(q(\theta_q | T)) = \mathbb{E}_{\theta_q} \left[ \log \left( \mathcal{N}(\theta_q | 0, \text{diag}(\gamma_q)^{-1}) \right) + \sum_{n=2}^{N} \log \left( \mathcal{N}(x_{q,n} | \theta_q^T \tilde{x}_{q,n-1}, z_{q,n}-1 \right) \right]
\]

where \( \tilde{x}_{q,n} \) is the \( P \)-dimensional column vector that represents the lags for \( x_{q,n} \) (i.e. the portion of \( \tilde{x}_n \) that belongs to the \( q \)th element of \( x_n \)). This leads to

\[
\log(q(\theta_q | T)) = -\frac{1}{2} \mathbb{E}_{\theta_q} \left[ \theta_q^T \text{diag}(\gamma_q) \theta_q + \sum_{n=2}^{N} z_{q,n} (x_{q,n} - \theta_q^T \tilde{x}_{q,n-1})^2 \right]
\]

where \( \tilde{x}_{q,n} \) is the \( P \)-dimensional column vector that represents the lags for \( x_{q,n} \) (i.e. the portion of \( \tilde{x}_n \) that belongs to the \( q \)th element of \( x_n \)). This leads to

\[
\log(q(\theta_q | T)) = -\frac{1}{2} \mathbb{E}_{\theta_q} \left[ \theta_q^T \left( \sum_{n=2}^{N} z_{q,n} \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T + \text{diag}(\gamma_q) \right) \theta_q - 2\theta_q \sum_{n=2}^{N} z_{q,n} x_{q,n} \tilde{x}_{q,n-1} \right]
\]
Thus \( q(\theta_q | T) \) may be recognised as the Gaussian \( \mathcal{N}(\theta_q | \mu_{\theta_q}, \Sigma_{\theta_q}) \), where

\[
\Sigma_{\theta_q} = \left( \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle + \text{diag}(\langle \gamma_q \rangle) \right)^{-1}
\]

(6.48)

\[
\mu_{\theta_q} = \Sigma_{\theta_q} \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle \tilde{x}_{q,n} \rangle
\]

(6.49)

This calculation is dependent on the expectation \( \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle \). Taking each element of this \( P \times P \) matrix separately:

\[
\langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle_{i,j} = \begin{cases} 
\langle \tilde{x}_{n-1} \tilde{x}_{n-1}^T \rangle_{(q-1)p+i,(q-1)p+i} & \text{if } i = j \\
\langle \tilde{x}_{(q-1)p+i,n-1} \rangle \langle \tilde{x}_{(q-1)p+j,n-1} \rangle & \text{if } i \neq j
\end{cases}
\]

(6.50)

6.1.3. Latent variables, \( z_n \)

It is not analytically tractable to locate a joint distribution for each \( z_n \), so the expression for each \( z_{q,n} \) individually (from (5.28)) may be rewritten, with the change of notation for PPCA-AR, as

\[
\log( q(z_{q,n} | T) ) = \mathbb{E}_{z_{q,n}} \left[ \log \left( \mathcal{N}(x_{q,n} | \theta_q^T \tilde{x}_{q,n-1}, \gamma_{q,n}^{-1}) \right) + \log \left( \mathcal{G}(z_{q,n} | \frac{d_q}{2}, \frac{d_q}{2}) \right) \right]
\]

(6.51)

(6.52)

This leads to

\[
\log( q(z_{q,n} | T) ) = \mathbb{E}_{z_{q,n}} \left[ \frac{1}{2} \log(z_{q,n}) - \frac{1}{2} (x_{q,n} - \theta_q^T \tilde{x}_{q,n-1})^2 z_{q,n} + \left( \frac{d_q}{2} - 1 \right) \log(z_{q,n}) - \frac{d_q}{2} z_{q,n} \right]
\]

(6.53)

\[
= \mathbb{E}_{z_{q,n}} \left[ \left( \frac{d_q}{2} + \frac{1}{2} - 1 \right) \log(z_{q,n}) \right.
\]

\[
- \frac{1}{2} \left( d_q + x_{q,n}^2 - 2x_{q,n} \theta_q^T \tilde{x}_{q,n-1} + \text{trace}(\theta_q \theta_q^T \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T) \right) z_{q,n}
\]

(6.54)

Thus \( q(z_{q,n} | T) \) may be recognised as the Gamma \( \mathcal{G}(z_{q,n} | \alpha_{z_q}, \beta_{z_q,n}) \), where

\[
\alpha_{z_q} = \frac{\langle d_q \rangle + 1}{2}
\]

(6.55)

\[
\beta_{z_q,n} = \frac{1}{2} \left( \langle d_q \rangle + \langle x_{q,n}^2 \rangle - 2 \langle x_{q,n} \rangle \langle \theta_q^T \tilde{x}_{q,n-1} \rangle + \text{trace}(\langle \theta_q \theta_q^T \rangle \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle) \right)
\]

(6.56)

The prior for the first latent variable vector is different from the rest, due to its position at the beginning of the time series. This leads to different posterior definitions for the
elements of $z_1$:

$$
\log( q(z_{q,1} | T)) = \mathbb{E}_{z_{q,1}} \left[ \log( N(x_{q,1} | 0, z_{q,1})^{-1}) \right] + \log \left( G(z_{q,1} | \frac{d_q}{2}, \frac{d_q}{2}) \right) (6.57)
$$

$$
= \mathbb{E}_{z_{q,1}} \left[ \frac{1}{2} \log(z_{q,1}) - \frac{1}{2} x_{q,1}^2 z_{q,1} + \left( \frac{d_q}{2} - 1 \right) \log(z_{q,1}) - \frac{d_q}{2} z_{q,1} \right] (6.58)
$$

$$
= \mathbb{E}_{z_{q,1}} \left[ \left( \frac{d_q}{2} + \frac{1}{2} - 1 \right) \log(z_{q,1}) - \frac{1}{2} (d_q + x_{q,1}^2) z_{q,1} \right] (6.59)
$$

This leads to the Gamma distribution $G(z_{q,1} | \alpha_{z_q}, \beta_{z_q,1})$ for $q(z_{q,1} | T)$, where

$$
\alpha_{z_q} = \langle d_q \rangle + \frac{1}{2} (d_q + x_{q,1}^2) (6.60)
$$

$$
\beta_{z_q,1} = \frac{1}{2} \left( \langle d_q \rangle + \langle x_{q,1}^2 \rangle \right) (6.61)
$$

### 6.2. Summary

The full results for robust variational PPCA-AR, incorporating the inherited derivations from chapters 4 and 5, are shown on pages 126 and 127.

Each approximate posterior distribution is dependent on the expected values of one or more of the others, so a closed-form algebraic solution cannot be obtained. We may arrive at a set of solutions by initialising the required expectations (from maximum likelihood Gaussian PPCA for $W, X$ and $\lambda$, setting $\bar{t}$ to the mean observation, each $u_{r,n}, z_{q,n}, \delta_q$ and $\gamma_{p,q}$ to 1, and $v$ and each $d_q$ to 100) and then iteratively updating the estimate for each hyperparameter based on the current estimates of the values on which it depends, until convergence. The required current expectations are obtained using the standard expressions (see appendix A)

$$
\langle w_{r,c} \rangle = \mu_{w_c} (6.62)
$$

$$
\langle w_{r}^T w_{r,\cdot} \rangle = \Sigma_{w_{r}} + \mu_{w_{r}}^T \mu_{w_{r}} (6.63)
$$

$$
\langle w_{r,q}^2 \rangle = \langle w_{r}^T w_{r,\cdot} \rangle_{q,q} (6.64)
$$

$$
\langle \delta_q \rangle = \alpha_{\delta} / \beta_{\delta_q} (6.65)
$$

$$
\langle \lambda \rangle = \alpha_{\lambda} / \beta_{\lambda} (6.66)
$$

$$
\langle u_{r,n} \rangle = \alpha_{u} / \beta_{u_{r,n}} (6.67)
$$

$$
\langle \log(u_{r,n}) \rangle = \psi(\alpha_{u}) - \log(\beta_{u_{r,n}}) (6.68)
$$

$$
\langle x_{n} \rangle = \mu_{x_n} (6.69)
$$

$$
\langle x_{n} x_{n}^T \rangle = \Sigma_{x_{n}} + \mu_{x_n}^T \mu_{x_n} (6.70)
$$

$$
\langle x_{q,n}^2 \rangle = \langle x_{n} x_{n}^T \rangle_{q,q} (6.71)
$$

$$
\langle \bar{t} \rangle = \mu_{\bar{t}} (6.72)
$$

$$
\langle \bar{t}^T \bar{t} \rangle = \Sigma_{\bar{t}} + \mu_{\bar{t}}^T \mu_{\bar{t}} (6.73)
$$

$$
\langle \bar{t}_{r}^2 \rangle = \langle \bar{t}^T \bar{t} \rangle_{r,r} (6.74)
$$
6. Robust variational PPCA with autoregression

The variables’ distributions were calculated in the order shown below.

\[ q^{(a)}(\tilde{x}_n) = \mathcal{N}(\tilde{x}_n | \tilde{\mu}_n^{(a)}, \tilde{\Sigma}_n^{(a)}) \] (6.75)

\[ \tilde{\Sigma}_n^{(a)} = \left( \langle \lambda \rangle \langle \tilde{W}^T \rangle \text{diag}(\tilde{u}_n) \tilde{W} \right) + \text{diag}(\langle \tilde{z}_n \rangle) \] (6.76)

\[ \tilde{\mu}_1^{(a)} = \tilde{\Sigma}_1^{(a)} \left( \langle \lambda \rangle \langle \tilde{W}^T \rangle \text{diag}(\langle u_1 \rangle)(t_1 - \tilde{t}) + \text{diag}(\langle \tilde{z}_1 \rangle)\mu_0 \right) \] (6.77)

\[ \tilde{\mu}_n^{(a)} = \tilde{\Sigma}_n^{(a)} \left( \langle \lambda \rangle \langle \tilde{W}^T \rangle \text{diag}(\langle u_n \rangle)(t_n - \tilde{t}) + \text{diag}(\langle \tilde{z}_n \rangle)\langle \Theta \rangle \langle \tilde{x}_{n-1} \rangle \right) \] (6.78)

\[ q(x_n | T) = \mathcal{N}(x_n | F\tilde{\mu}_n, F\Sigma_n F^T) \] (6.79)

\[ \tilde{\Sigma}_n^{-1} = (\tilde{\Sigma}_n^{(a)})^{-1} + \langle \Theta^T \rangle \text{diag}(\tilde{z}_{n+1}) \Theta \] (6.80)

\[ \tilde{\Sigma}_N = \tilde{\Sigma}_N \] (6.81)

\[ \tilde{\mu}_n = \tilde{\Sigma}_n \left( (\tilde{\Sigma}_n^{(a)})^{-1} \tilde{\mu}_n^{(a)} + \langle \Theta^T \rangle \text{diag}(\langle \tilde{z}_n \rangle) \langle \tilde{x}_{n+1} \rangle \right) \] (6.82)

\[ \tilde{\mu}_N = \tilde{\mu}_N^{(a)} \] (6.83)

\[ q(w_r: | T) = \mathcal{N}(w_r: | \mu_{w_r}, \Sigma_{w_r}) \] (6.84)

\[ \Sigma_{w_r}^{-1} = \text{diag}(\langle \delta \rangle) + \langle \lambda \rangle \sum_{n=1}^{N} (u_{r,n}) (x_n x_n^T) \] (6.85)

\[ \mu_{w_r} = \Sigma_{w_r} \langle \lambda \rangle \sum_{n=1}^{N} (u_{r,n}) (t_{r,n} - \langle \tilde{t}_r \rangle) \] (6.86)

\[ q(\delta_q | T) = \mathcal{G}(\delta_q | \alpha_\delta, \beta_\delta) \] (6.87)

\[ \alpha_\delta = a_\delta + \frac{D}{2} \] (6.88)

\[ \beta_\delta = b_\delta + \frac{1}{2} \sum_{r=1}^{D} \langle w_{r,q}^2 \rangle \] (6.89)

\[ q(\lambda | T) = \mathcal{G}(\lambda | \alpha_\lambda, \beta_\lambda) \] (6.90)

\[ \alpha_\lambda = a_\lambda + ND/2 \] (6.91)

\[ \beta_\lambda = b_\lambda + \frac{1}{2} \sum_{n=1}^{N} \langle (t_n - Wx_n - \tilde{t})^T \text{diag}(u_n)(t_n - Wx_n - \tilde{t}) \rangle \] (6.92)

\[ q(\theta | T) = \mathcal{N}(\theta | \mu_\theta, \Sigma_\theta) \] (6.93)

\[ \Sigma_{\theta_q}^{-1} = \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle + \text{diag}(\langle \gamma_q \rangle) \] (6.94)

\[ \mu_{\theta_q} = \Sigma_{\theta_q} \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle x_{q,n} \rangle \langle \tilde{x}_{q,n-1} \rangle \] (6.95)

\[ q(\gamma_{p,q} | T) = \mathcal{G}(\gamma_{p,q} | \alpha_\gamma, \beta_{\gamma_{p,q}}) \] (6.96)

\[ \alpha_\gamma = a_\gamma + 1 \] (6.97)

\[ \beta_{\gamma_{p,q}} = b_\gamma + \frac{1}{2} \langle \theta_{p,q}^2 \rangle \] (6.98)

\[ q(\tilde{t} | T) = \mathcal{N}(\tilde{t} | \mu_\tilde{t}, \Sigma_\tilde{t}) \] (6.99)
6. Robust variational PPCA with autoregression

\[
\Sigma^{-1}_t = V^{-1}_t + \langle \lambda \rangle \sum_{n=1}^{N} \text{diag}(\langle u_n \rangle) \quad (6.100)
\]

\[
\mu_t = \Sigma_t \langle \lambda \rangle \sum_{n=1}^{N} \text{diag}(\langle u_n \rangle) (t_n - \langle W \rangle \langle x_n \rangle) \quad (6.101)
\]

\[
q(u_{r,n} | T) = \mathcal{G}(u_{r,n} | \alpha_u, \beta_{u_{r,n}}) \quad (6.102)
\]

\[
\alpha_u = \langle \langle v \rangle + 1 \rangle / 2 \quad (6.103)
\]

\[
\beta_{u_{r,n}} = \frac{1}{2} \left( \langle v \rangle + \langle (t_{r,n} - w_r \cdot x_n - \bar{t}_r)^2 \rangle \right) \quad (6.104)
\]

\[
q(v | T) = \mathcal{G}(v | \alpha_v, \beta_v) \quad (6.105)
\]

\[
\alpha_v = a_v + ND / 2 \quad (6.106)
\]

\[
\beta_v = b_v - \frac{1}{2} \left( ND + \sum_{n=1}^{N} \sum_{r=1}^{D} (\log(u_{r,n}) - \langle u_{r,n} \rangle) \right) \quad (6.107)
\]

\[
q(z_{q,n} | T) = \mathcal{G}(z_{q,n} | \alpha_z, \beta_{z_{q,n}}) \quad (6.108)
\]

\[
\alpha_z = \frac{\langle d_q \rangle + 1}{2} \quad (6.109)
\]

\[
\beta_{z_{q,1}} = \frac{1}{2} \left( \langle d_q \rangle + \langle x_{q,1}^2 \rangle \right) \quad (6.110)
\]

\[
\beta_{z_{q,n}} = \frac{1}{2} \left( \langle d_q \rangle + \langle x_{q,n}^2 \rangle - 2 \langle x_{q,n} \rangle \langle \theta_q^T \rangle \langle \tilde{x}_{q,n-1} \rangle + \text{trace}(\theta_q \theta_q^T) \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle \right) \quad (6.111)
\]

\[
q(d_q | T) = \mathcal{G}(d_q | \alpha_d, \beta_{d_q}) \quad (6.112)
\]

\[
\alpha_d = a_d + N / 2 \quad (6.113)
\]

\[
\beta_{d_q} = b_d - \frac{1}{2} \left( N + \sum_{n=1}^{N} (\log(z_{q,n}) - \langle z_{q,n} \rangle) \right) \quad (6.114)
\]

The following definitions are required (with blkdiag(\cdot) being a block diagonal matrix):

\[
(W^T \text{diag}(u_n)W)_{i,j} = \begin{cases} 
\sum_{r=1}^{D} \langle u_{r,n} \rangle \langle w_r^T \cdot w_r \rangle_{i,i} & \text{if } i = j \\
\sum_{r=1}^{D} \langle u_{r,n} \rangle \langle w_r \rangle_{i,j} & \text{if } i \neq j 
\end{cases} \quad (6.115)
\]

\[
\langle \Theta^T \text{diag}(z_{n+1}) \Theta \rangle = \text{blkdiag}(\langle \theta_q \theta_q^T \rangle \langle z_{q,n+1} \rangle) \quad (6.116)
\]

\[
\langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle = \begin{cases} 
\langle \tilde{x}_{n-1} \tilde{x}_{n-1}^T \rangle_{(q-1)p+i,(q-1)p+i} & \text{if } i = j \\
\langle \tilde{x}_{(q-1)p+i,n-1} \rangle \langle \tilde{x}_{(q-1)p+i,n-1} \rangle_{(q-1)p+j,n-1} & \text{if } i \neq j 
\end{cases} \quad (6.117)
\]

and

\[
\langle (t_{r,n} - w_r \cdot x_n - \bar{t}_r)^2 \rangle = \langle t_{r,n}^2 \rangle + \text{trace}(\langle w_r^T \cdot w_r \rangle \langle x_n x_n^T \rangle) + \langle \bar{t}_r^2 \rangle - 2 \langle t_{r,n} \rangle \langle \langle w_r \rangle \cdot \langle x_n \rangle + \langle \bar{t}_r \rangle \rangle + 2 \langle w_r \rangle \langle x_n \rangle \langle \bar{t}_r \rangle \quad (6.118)
\]

\[
\langle (t_n - W x_n - \bar{t})^T \text{diag}(u_n) (t_n - W x_n - \bar{t}) \rangle = t_n^2 \text{diag}(\langle u_n \rangle) + \langle x_n W^T \text{diag}(u_n) W x_n \rangle + \text{trace}(\text{diag}(\langle u_n \rangle) \langle \bar{t} \bar{t}^T \rangle)
\]

\[
- t_n^2 \text{diag}(\langle u_n \rangle) \langle W \rangle \langle x_n \rangle + \langle \bar{t} \rangle - \langle W \rangle \langle x_n \rangle + \langle \bar{t} \rangle \text{diag}(\langle u_n \rangle) \rangle t_n + \langle \bar{t} \rangle \text{diag}(\langle u_n \rangle) \langle W \rangle \langle x_n \rangle + \langle x_n \rangle^{T} \langle W \rangle^{T} \text{diag}(\langle u_n \rangle) \langle \bar{t} \rangle \quad (6.119)
\]
6. Robust variational PPCA with autoregression

\[ \langle z_{q,n} \rangle = \alpha_z / \beta z_{q,n} \] (6.119)

\[ \langle \log(z_{q,n}) \rangle = \psi(\alpha_z) - \log(\beta z_{q,n}) \] (6.120)

\[ \langle v \rangle = \alpha_v / \beta_v \] (6.121)

\[ \langle d_q \rangle = \alpha_d / \beta d_q \] (6.122)

\[ \langle \theta_q \rangle = \mu_\theta \] (6.123)

\[ \langle \theta_q^\top \theta_q^\top \rangle = \Sigma_\theta + \mu_\theta \mu_\theta^\top \] (6.124)

\[ \langle \theta^2_{p,q} \rangle = \langle \theta_q^\top \theta_q^\top \rangle_{p,p} \] (6.125)

\[ \langle \gamma_{p,q} \rangle = \alpha_\gamma / \beta_{\gamma p,q} \] (6.126)

where \( \psi(\cdot) \) is the digamma function.

Having formulated this model, straightforward adjustments can be made to change the assumptions about the distributions of the noise. If all the values in the latent variables for the observation noise (in the \( u_n \)) take the value 1, then the Student-t distribution from (6.3), which may be written as

\[
\int N(e_{r,n} | 0, (\lambda u_{r,n})^{-1}) G(u_{r,n} | v^2, v^2) \, du_{r,n},
\]

becomes just \( N(e_{r,n} | 0, \lambda^{-1}) \) and the observation noise assumption is now Gaussian. Similarly, setting all the values in the latent variables for the excitation noise (in the \( z_n \)) to 1 converts the excitation noise assumption to Gaussian. Thus an implementation of this model is easily able to support the four possible Student-t/Gaussian combinations of noise assumptions.

In PPCA-AR constraints applied to the shapes of the \( \Theta, \Phi \) and \( W \) matrices differentiate the model from a general linear dynamical system. In the maximum likelihood formulation the individual \( \theta_q \) and \( \kappa_q \) are calculated separately and then the associated \( \Theta \) and \( \Phi \) constructed from them, and the orthogonality of the components in \( W \) is enforced via Lagrange multipliers to ensure that \( W^\top W \) is the identity matrix. In this variational approach the \( \theta_q \) are again calculated individually and the \( \Theta \) constructed from them, and the \( \kappa_q \) are all set to 1, so the issue does not arise. The components in \( W \) are constrained only by independence assumptions in the prior and the independence of the latent variables. These ensure that the covariance matrix for each row of \( W \) is diagonal. It is not isotropic because the \( \kappa_q = 1 \) assumption ensures that any variance in the \( Wx_n \) term must be expressed as variance in the elements of \( W \). The variational treatment has not altered the constraints on the model and has generalised the noise assumptions from Gaussian to Student-t; thus robust variational PPCA-AR may be seen as a generalisation of the maximum likelihood version.

The Bayesian approach results in a probability distribution for the observations from which a reconstruction may be made. The spread of the distribution will indicate the model’s uncertainty regarding the reconstruction. Using \( \Omega \) to represent the set of all the model’s parameter variables, we would need to perform the integration \( \int p(T | \Omega) \, d\Omega \). This is insoluble, but the factorised variational technique leads to an intuitive approximation:

\[
\log(q(t_n | T)) = E_{\lambda t_n} \left[ N \left( t_n | Wx_n + \bar{t}, (\lambda \text{diag}(u_n))^{-1} \right) \right] + \text{const} \quad (6.127)
\]
So \( q(t_n | T) \) is just \( \mathcal{N} (t_n | \langle W \rangle \langle x_n \rangle + \langle \bar{t} \rangle, (\langle \lambda \rangle \text{diag} (\langle u_n \rangle))^{-1}) \). The approximate posterior distribution of every observation is Gaussian, but each one has a different mean and precision. It may be a bit surprising that this is not a Student-t. This is because the approximation is based on the posterior expectations of \( \lambda \) and \( u_n \) rather than an integration over their entire (Gamma) distributions.

6.3. Illustration: synthetic data

It is necessary to demonstrate that robust variational PPCA-AR is able to estimate model parameters effectively, which means constructing synthetic data which conforms to the model’s assumptions and for which the parameters are known. The first illustration demonstrates the estimation of the noise distributions through the \( \lambda, d \) and \( v \) parameters; the second demonstrates the estimation of the more structural parameters, \( \bar{t}, W \) and the \( \theta_q \); the third looks at the variances of the approximate posterior distributions; the fourth demonstrates model order estimation; the final illustration demonstrates the ability of the model to impute missing values.

Stationary AR signals of order \( P \), for the \( q \)th latent variable, were generated exactly as for the maximum likelihood model (see section 3.3 on page 64). The synthetic observations were then generated according to the expression \( t_n = Wx_n + \bar{t} + e_n \), where \( W \) is some \( D \times Q \) matrix, \( \bar{t} \) is some random Gaussian distributed vector and \( e_n \) is a \( D \)-dimensional vector whose \( r \)th element is randomly drawn from Student-t distribution with zero mean, precision \( \lambda \) and \( v \) degrees of freedom. The latent variables are each generated independently from AR processes of order \( P \), using a different, randomly-generated \( \theta_q \) (as above), precision \( \kappa_q = 1 \) and potentially different degrees of freedom, \( d_q \), for each variable.

6.3.1. Illustration 1: estimating noise parameters

Datasets were generated with \( D = 5, N = 200, Q = 3 \) and \( P = 10 \) and with every combination of \( \lambda \) in the set \( \{0.01, 0.1, 1, 10, 100, 1000\} \) and \( v \) and \( d_q \) the integer values from 1 to 10. The elements of \( W \) were randomly sampled from a Gaussian distribution with zero mean and unit variance and different \( \theta_q \) values were generated for each latent variable. The model was trained against each of them using the known \( Q \) and \( P \) values (to minimise the effect of ARD). Figure 6.2 shows that there is a close correlation between the model’s estimates of \( \lambda \) and the actual values used to generate the data. There is a tendency to underestimate larger precisions, but the relationship between estimated and actual is approximately linear.

Figure 6.3 compares the estimated and actual observation noise degrees of freedom, \( v \). The points have been coloured according to the the actual \( \lambda \) value used to generate the data. For small values of actual \( v \), where the variance of the distribution is effectively infinite \( (v \leq 2) \) there is a reasonably good correspondence, with a tendency to over-estimate the value. For larger values of \( v \) the results depend very much on the actual noise precision:
6. Robust variational PPCA with autoregression

Figure 6.2. Comparison of estimated observation noise precision $\lambda$ with actual values for each of 600 trials. Note the logarithmic ordinate.

Figure 6.3. Comparison of estimated and actual observation noise degrees of freedom, $v$. Each point is coloured according to the actual observation noise precision and a small amount of horizontal scatter added to each point for visualisation.

where there is high precision, i.e. small amounts of noise, the estimate tends to indicate a Gaussian distribution, while for low precision the estimate tends to indicate a more heavy-tailed distribution than was actually the case.

The comparison of estimated and actual latent variables’ degrees of freedom, $d_q$, in figure 6.4 is similarly coloured according to the the actual $\lambda$ values. There is a marked tendency to estimate the excitation noise distribution as Gaussian, though where the actual $d_q$ is 1 (and the variance is effectively infinite) there is a better correlation between the estimates and actuals.

6.3.2. Illustration 2: estimating $\bar{t}$, $W$ and $\theta$

Figure 6.5 shows a plot of estimated against actual values for the mean observation, $\bar{t}$. There is generally an approximately linear correlation between the two, with those estimates generated from datasets with high observation noise precision (unsurprisingly) generally better estimated than those from datasets with low observation noise precision.
6. Robust variational PPCA with autoregression

Figure 6.4. Comparison of estimated and actual latent variables degrees of freedom, $d_q$. Each point is coloured according to the actual observation noise precision and a small amount of horizontal scatter added to each point for visualisation.

Figure 6.5. Comparison of estimated and actual mean observation values, $\bar{t}$. Each point is coloured according to the actual observation noise precision.
Figure 6.6. Hinton diagrams comparing the estimated and actual mixing matrix, \( \mathbf{W} \), where the estimate is good. (a) the actual \( \mathbf{W} \) used to generate the data; (b) the estimate produced by the model; (c) the estimate produced by the model after a Procrustes transformation has been applied; (d) the Procrustes transformation matrix.

Figure 6.7. Hinton diagrams comparing the estimated and actual mixing matrix, \( \mathbf{W} \), where the estimate is bad. (a) the actual \( \mathbf{W} \) used to generate the data; (b) the estimate produced by the model; (c) the estimate produced by the model after a Procrustes transformation has been applied; (d) the Procrustes transformation matrix.
The results for the estimates of $W$ are similarly better where there is low noise and worse where there is high noise. As described in section 2.5.1, the model is only able to estimate the components in $W$ up to a sign, permutation and rotation, so the Procrustes transformation is used to transform the estimate in the direction of the actual value. Figure 6.6 shows an example of a $W$ where the model’s estimate is close to the actual. The figure contains four Hinton diagrams: (a) the actual $W$ used to generate the data; (b) the estimate produced by the model; (c) the estimate produced by the model after a Procrustes transformation has been applied; (d) the Procrustes transformation matrix. The correlations between each column of the actual $W$ and the corresponding column of the transformed estimated matrix are good, but the relative magnitudes of the components in the estimate are different. The shape of the Procrustes transformation matrix indicates that virtually no rotation has been required; each row and column contains only one sizeable value. The large values are not on the main diagonal, so some permutation has been involved, and the large squares are not all white, so some sign change has also been involved, as can be seen by comparing figures 6.6(a) and (b).

Figure 6.7 shows an example of a $W$ where the model’s estimate is not close to the actual. The figure contains four Hinton diagrams, as before. The estimated matrix obviously bears no resemblance to the actual, either before the transformation or after it. This does not mean that the model is unable to represent the data, just that it has found a different representation of it from the actual structure used to generate it.

The correspondence between the magnitude of the observation noise and the quality of the estimates of $W$ are shown in two heat maps in figure 6.8. For the $W$ for each of the tests, the correlation between each component in the actual $W$ and the corresponding component in the transformed estimated $W$ was calculated, resulting in three correlation values for each test. The absolute value of the correlations were taken, as we are interested only in the magnitude, not the sign. The left-hand heat map in 6.8 shows these three correlation values as three columns, where each row represents one test. The tests have been sorted by the actual observation noise precision value used to generate the data. The dominance of red in the bottom of the map shows that $W$ tends to be well estimated for low levels of noise. The right-hand heat map is generated from the Procrustes transformation matrix for each test and is similarly ordered by the actual observation noise precision. Each row of the map is generated by taking the Procrustes transformation matrix, sorting its rows into ascending order (as a way of removing the permutations), taking the absolute value (we are interested only in the magnitude and not the sign), and then reshaped into a vector. A perfect result will have values of 1 in columns 3, 5 and 7 of the heat map and zeros in other columns. Occasionally ARD switches off all the components in $W$, resulting in a Procrustes transformation matrix which is the identity. This misleading result has been removed by setting all the values for that row to 0.5 for the heat map. Again the tests have been sorted by the actual observation noise precision value used to generate the data. It can be seen that there is a strong red band in column 3 across the whole range of $\lambda$ values; the bands in columns 5 and 7 are not so obvious, indicating that one column of the estimated $W$ is closely correlated with the actual, but the other two columns are less so. The concentration of green near the top of the heat map (where the observation
Figure 6.8. Heat maps for determining the quality of the model’s estimates of the mixing matrix, \( W \). See text for details.
noise levels are high) are caused by the model switching off all the components in $W$; the signals are being represented by a mean value, with all variation absorbed into the noise term.

As described previously, the model is only able to estimate the components in $W$ up to a sign, permutation and rotation transformation. This affects the estimation of the latent variables and hence the AR coefficients. The AR coefficients may only be directly compared where no rotation transformation of the estimate of $W$ has occurred, and any change of sign and permutation of the latent variables have been taken into account. Figure 6.9 shows an example of the comparison between actual and estimated $\theta_q$ for the three rows of one dataset, showing that the correspondence is generally good, but with a tendency for the coefficients to be underestimated.

Figure 6.9. Comparison of estimated and actual AR coefficients ($\theta_q$) for the three rows of one synthetic dataset. Actual values are shown as black circles with a black line to guide the eye; estimated values are shown as red crosses.

6.3.3. Illustration 3: variances of approximate posterior distributions

Two of the datasets used in the previous illustrations were selected, one where the precision was large ($\lambda = 100$, with $d_q = 2$ and $v = 2$), and hence approximate posteriors that are good estimates of the actual distributions are expected, and one where the precision is small ($\lambda = 1$, with $d_q = 10$ and $v = 10$), and hence approximate posteriors that are less good estimates of the actual distributions are expected. The model was trained against each of them until convergence and then the magnitudes of the variances inspected for each posterior distribution. In both cases the standard deviations for the latent variables were too small to see on a box plot.

Figure 6.10 show the box plots for $W$, $\bar{t}$, the $\theta_q$ and the $d_q$ respectively, with the box covering one standard deviation above and below the expected value in each case. In each figure the plot on the left shows the results for the high-precision dataset (i.e. the
Figure 6.10. Expectations ± one standard deviations from the approximate posterior distributions for (a) \( W \), (b) \( \bar{t} \), (c) \( d \), and (d) \( \theta_q \) estimated for, on the left, a synthetic dataset with high precision (\( \lambda = 100 \)) and, on the right, a synthetic dataset with lower precision (\( \lambda = 1 \)). In each case the expectation is shown as a red line, with a blue box indicating one standard deviation above and below.
6. Robust variational PPCA with autoregression

Figure 6.11. Estimation of model order: components in $W$. Each point has been coloured according to the actual observation noise precision and offset laterally by a small amount for clarity. Where $\lambda$ is high, the model order estimation is generally correct. For lower $\lambda$ values the model may be either under- or over-estimated, with no discernable pattern.

<table>
<thead>
<tr>
<th>variable</th>
<th>High-precision dataset</th>
<th>Lower-precision dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>30.45 ± 1.36</td>
<td>0.81 ± 0.04</td>
</tr>
<tr>
<td>$\nu$</td>
<td>5.51 ± 0.25</td>
<td>365.23 ± 16.33</td>
</tr>
</tbody>
</table>

The variances are generally larger (with respect to the magnitude of the expected values) for the lower-precision dataset for $W$ and $\bar{t}$ and about the same for the other variables. It is, perhaps, not surprising that the dataset with higher observation noise precision should show lower variances for the estimates of some parameter variables.

6.3.4. Illustration 4: estimating model orders

One hundred synthetic datasets were generated with $\nu$ and $d_q$ set to 2 and $\kappa_q = 1$. In each case $\lambda$ was randomly selected to be one of $\{0.01, 1, 100\}$, the PPCA model order, $Q$, was set to a random integer between 1 and 10 and the AR model order, $P$, to a random integer between 0 and 8. Different, randomly-generated $W$ and $\theta_q$ were used for each of them. The dimensionality of the observations, $D$, was set to $Q + 2$.

The model was trained against each dataset until convergence, with $Q$ set to $D - 1$ and $P$ set to 10 so that ARD should switch off at least one component and two AR coefficients in each test. Figure 6.11 compares the actual number of components with the number estimated by the model. Each point has been offset laterally by a small amount for clarity. The points are coloured according to the actual observation noise precision. Where $\lambda$ is high, the model order estimation is generally correct. For lower $\lambda$ values the model may be either under- or over-estimated, with no discernable pattern.

Where ARD switches off components the elements in the $W$ matrix are set to zero, making the estimated model order easy to identify. For the coefficients in each $\theta_q$ the results are less clear as the values are never constrained to be exactly zero. In chapter 5, if the
estimated value of a coefficient was more than one standard deviation away from zero, i.e. \( \theta_{q,i}^2 > 1/\gamma_{q,i} \), then it was deemed to be switched on. However, in this model that test concludes that all coefficients are switched off. Instead figure 6.12 looks at the relative magnitude of the coefficients over all the tests, separated into plots for each actual model order. Each diagram shows box plots for \( \theta_{q,i} \) for the actual value (in black) and the three different values of \( \lambda \) for each model order, from \( P = 0 \) to \( P = 8 \). In each case it may be seen that the magnitude of the estimates is significantly greater for coefficients where \( i \leq P \), indicating that it is indeed selecting the correct model orders.

For each plot in figure 6.12, figure 6.13 contains the corresponding plot of standard deviations for the posterior distributions of the AR coefficients. Where the noise is low (\( \lambda = 100 \)), the standard deviations of those coefficients that are switched on is generally higher than those that are switched off. Where the noise is high (\( \lambda = 0.01 \)), there is not much variation in standard deviations across all the coefficients.

### 6.3.5. Illustration 5: imputing missing values

The ability of PPCA-AR to impute missing values for datasets where the noise is significantly heavier-tailed than Gaussian is demonstrated by training the model against 50 different synthetic datasets for proportions of missing data from 0 to 90% in increments of 5%. The patterns of gaps are different for each dataset.

As for the maximum likelihood version of PPCA-AR, elements of the observations are registered as missing through a new matrix, \( M \), of the same dimensionality as the observations in \( T \). Each value, \( m_{i,j} \), is an indicator variable which takes the value 1 if the value \( t_{i,j} \) is observed or 0 if it is missing. Initially all the values are set to 1. A gap is introduced into the data by uniformly randomly selecting a row and column of \( M \) as the centre, sampling from a Poisson distribution of mean 3 and then setting the appropriate values in that row of \( M \) to zero. Gaps are generated independently for each row of \( M \) to ensure that approximately the same proportion of values are missing for each variable. Different \( M \) matrices are generated for 0% to 90% missing at 5% intervals, with the gaps being cumulative, meaning that the 20% missing \( M \) is the 15% missing \( M \) with 5% more values missing. These percentages are approximate as the algorithm adds gaps until the desired percentage is exceeded.

Each dataset is generated with parameters \( D = 12, N = 200, Q = 6, P = 10, \lambda = 10, \) each \( \kappa_q = 10, \) each \( d_q = 2 \) and \( v = 2 \). Both robust variational PPCA-AR and maximum likelihood PPCA-AR were trained against each dataset until convergence for each of the missing patterns. The model orders for both were set to the actual values to enable direct comparison. The models were compared by measuring the mean squared error of the reconstruction of the observed and missing values separately. Figure 6.14 shows the mean errors (full line) plotted against the proportions of missing data, along with one standard deviation above the mean (dotted line). The results for robust variational PPCA-AR are shown in red and for maximum likelihood PPCA-AR in blue.
Figure 6.12. Box plots of the relative magnitudes of the AR coefficients over all the tests, separated into separate plots for each actual model order. Each diagram shows box plots for $\theta_{q,i}$ for the actual value (in black) and the three different values of $\lambda$ for each model order, from $P = 0$ to $P = 8$. In each case it may be seen that the magnitude of the estimates is significantly greater for coefficients where $i \leq P$, indicating that it is indeed detecting the correct model orders.
Figure 6.13. Box plots of the standard deviations of the $\theta_{q,i}$ associated with the plots shown in figure 6.12.
Figure 6.14. Imputation of missing data by PPCA-AR, comparing the maximum likelihood model (in blue) with the robust variational model (in red). For each model the line represents the mean value over 50 tests and the dotted line represents one standard deviation above the mean. Each of the tests uses a different synthetic dataset and a different pattern of missing values.

Figure 6.15. Comparison of the AR coefficients for one latent variable for increasing proportions of missing data. The surface plot on the left is for the maximum likelihood model; that on the right for the robust variational model. Note that as more data becomes missing ARD in the robust variational model is switching the coefficients off as there becomes less evidence for them. This does not happen in the maximum likelihood model.
It may be seen that for lower proportions of missing data (up to about 45%) the maximum likelihood model is better at reconstructing the observed data than the robust variational model, even though the noise is Student-t and not Gaussian. However, in all cases for the missing data and for higher proportions of missing values for the observed data the robust variational model is better.

It is also instructive to look at the estimated AR coefficients as more data is missing. Figure 6.15 shows a surface plot for each model (maximum likelihood on the left, robust variational model on the right) which shows the same AR coefficient vector estimated for differing amounts of missing data. Note that as the proportion of missing data increases, ARD in the robust variational model switches the coefficients off as the evidence for them diminishes.

6.4. Results: real data

The model was trained against the ONS data, with $Q$ set to the maximum possible value of $D - 1$ and $P$ set to the arbitrary value of 20. Due to time constraints the training was limited to 2000 iterations. Figure 6.16 shows the Hinton diagram for the estimated $W$ for these data. None of the components have been switched off by ARD, though the first 2 or 3 components seem to be more significant than the rest. Figure 6.17 shows the $\theta_q$ values associated with each of the 22 latent variables. This indicates that the model orders of the autoregressive processes varies between 1 (e.g. component 1) to approximately 12 (for component 14). However, the plots of the degrees of freedom by iteration shown in figure 6.18 indicate that the model has not yet converged. It seems, from this evidence, that both the observation and excitation noise are likely to be Gaussian since the values for $v$ and $d_q$ are tending to become large.

The NASA data present more of a challenge to the algorithm due to the dimensionality of the data ($D$ is 113,593) and data is genuinely missing, so we can only inspect the results.
Figure 6.17. Plot of the $\theta_q$ estimated for the ONS data by robust variational PPCA-AR.

Figure 6.18. Plots of the degrees of freedom estimated for the ONS data by robust variational PPCA-AR by iteration for (a) the observation noise and (b) the excitation noise. These show that convergence has not yet been reached, but the large values indicate that both noise elements are likely to be approximately Gaussian.
of the imputations of the missing data to see if they seem plausible. In order to attain results in a reasonable time-frame and using the computing power available, values for $Q$ of 30 and $P$ of 10 were used, and only a $100 \times 200$ section of the pictures was used. The results were very variable: figure 6.19 shows before and after images for two frames where the estimation of the missing values appears to be good, while figure 6.20 shows the same for two frames where human intuition would say that the estimation is implausible.

None of the components in $W$ have been switched off by ARD, but all the AR coefficients have been constrained to be close to zero and their ARD parameter variables all have large values. This indicates that the PPCA model order ($Q$) is probably higher than 30 and that there is not enough evidence for a temporal model in the data. Figure 6.21 shows how much data is missing from these picture sections. With approximately 84% of the values missing for pixels on the left-hand side of the image and the fact that ARD is switching off the AR coefficients due to lack of evidence for them in that small amount of data, it is not surprising that the estimations are of low quality.

This robust variational PPCA-AR model was also trained against a single observation set of the EEG data, consisting of the traces from 62 electrodes. Figure 6.22 shows the Hinton diagram for $W$. Of the components, 17 of the 61 have been switched off, indicating a model order of 44. The first three components appear to be more significant than the others; their magnitude compared to the others seems to indicate that there are three significant latent signals supporting the observed data. The latent variables for these three signals are shown in figure 6.23. The trigger for the evoked response was applied at about observation 200; the first latent variable may represent a response to the trigger, but it is impossible to say without further evidence.

The results are particularly interesting with regards to the $\theta_q$, which are shown as a surface plot in figure 6.24 (with the $\theta_q$ for the switched off latent variables not shown). Notice how uniform the estimates are across the different electrodes, as was noted in chapter 5. PPCA-AR also estimates the model order for the AR processes to be about 12.
Figure 6.19. Two examples of plausible estimates for the missing data in NASA SeaWiFS images. On the lower image the estimation on the right hand side of the picture seems to be better than that on the left.
Figure 6.20. Two examples of implausible estimates for the missing data in NASA SeaWiFS images.
Figure 6.21. Number of observed instances for each pixel in the 92 NASA SeaWiFS images. The blue pixels represent approximately 84% missing values, while the red pixels represent approximately 56% missing values.

Figure 6.22. Hinton diagram of the $W$ estimated for the EEG data. Note that components 31, 34, 36, 41-48, 50, 53, 55-57 and 60 have been switched off.
Figure 6.23. Plots of the three most significant latent variables for the EEG data.

Figure 6.24. Surface plot of the $\theta_q$ for the 44 latent variables that have not been switched off. A value of $P = 20$ was used to train the model.
6. Robust variational PPCA with autoregression

6.5. Robust variational FA-AR

If, instead of assuming that the observation noise across all observed variables is from the same distribution, it is assumed that each observed variable has a different precision, \( \lambda_r \) and degrees of freedom, \( v_r \), i.e. that \( p(e_{r,n}) = S(e_{r,n} \mid 0, \lambda_r^{-1}, v_r) \) rather than \( S(e_{r,n} \mid 0, \lambda^{-1}, v) \), then we arrive at model based on Factor Analysis: robust variational FA-AR. The approximate posteriors for the parameter variables are the same as for PPCA-AR (as shown on pages 126 and 127), except for the following:

\[
q(u_{r,n} \mid \mathbf{T}) = \mathcal{G}(u_{r,n} \mid \alpha_u, \beta_{u_{r,n}}) \tag{6.128}
\]

\[
\alpha_u = \frac{(v_r) + 1}{2} \tag{6.129}
\]

\[
\beta_{u_{r,n}} = \frac{1}{2} \left( v_r + \langle (t_{r,n} - \mathbf{w}_r; \mathbf{x}_n - \bar{t}_r)^2 \rangle \right) \tag{6.130}
\]

\[
q(v_r \mid \mathbf{T}) = \mathcal{G}(v_r \mid \alpha_v, \beta_{v_r}) \tag{6.131}
\]

\[
\alpha_v = a_v + N/2 \tag{6.132}
\]

\[
\beta_{v_r} = b_v - \frac{1}{2} \left( N + \sum_{n=1}^{N} \langle \log(u_{r,n}) \rangle - \langle u_{r,n} \rangle \right) \tag{6.133}
\]

\[
q(\lambda_r \mid \mathbf{T}) = \mathcal{G}(\lambda_r \mid \alpha_{\lambda}, \beta_{\lambda_r}) \tag{6.134}
\]

\[
\alpha_{\lambda} = a_{\lambda} + N/2 \tag{6.135}
\]

\[
\beta_{\lambda_r} = b_{\lambda} + \frac{1}{2} \sum_{n=1}^{N} \langle (t_{r,n} - \mathbf{w}_r; \mathbf{x}_n - \bar{t}_r)^T u_{r,n} (t_{r,n} - \mathbf{w}_r; \mathbf{x}_n - \bar{t}_r) \rangle \tag{6.136}
\]

6.6. Conclusions

Chapter 3 introduced the PPCA-AR model which combines PPCA and AR to capture both spatial and temporal structure in multivariate time-series data and in chapters 4 and 5, robust variational versions of PPCA and AR were described. In this chapter these have been combined into robust variational PPCA-AR, which extends the maximum likelihood model by incorporating Student-t noise assumptions for robustness and the ability to model leptokurtic data, and ARD to estimate the model orders. The model uses a robust variational version of the forward-backward algorithm to infer the approximate posterior distributions of the latent variables and approximate posterior distributions are obtained for all the model’s parameter variables; these distributions provide a measure of uncertainty in the estimated values.

Where the observation noise is small compared with the underlying signal, the algorithm makes generally good estimates of the parameter variables, with a tendency to underestimate the many of them. In these cases the model order is also generally accurately estimated. When used to impute the values of missing data in synthetic leptokurtic datasets, the model provides consistently better estimates than the maximum likelihood model.

When ARD switches off a PPCA component the elements of the vector are exactly zero, but for the AR coefficients, although ARD constrains the switched off values to be smaller...
than those that are switched on, there is no such clear-cut method of determining the model order. However, as the proportion of missing data increases, diminishing the evidence for a temporal model, ARD causes the AR coefficients to be switched off, in contrast to the maximum likelihood model.

The variational solution for PPCA-AR may be compared with the maximum likelihood solution described in chapter 3. In order to do this we must make some assumptions about the expected values of some of the parameter variables to ensure that the comparison is fair. Since the maximum likelihood model did not include ARD, the expected values of the ARD precision variables are set to zero (that is all PPCA components and all AR coefficients are switched on). The expected value of the mean observation is set to zero. The maximum likelihood model made Gaussian assumptions about the noise, so the expected values of the latent variables in the variational Bayesian model \(u_n\) and \(z_n\) are all set to 1 to replicate this. Lastly the parameters of the prior distributions are set to zero, rendering them extremely uninformative. With these assumptions in mind we may calculate the expected values of \(\lambda, W\) and \(\theta_q\). Starting with \(\lambda\):

\[
\langle \lambda^{-1} \rangle = \beta_\lambda / \alpha_\lambda \tag{6.137}
\]

\[
= 2 \frac{1}{ND} \sum_{n=1}^{N} \langle (t_n - Wx_n - \bar{t})^T \text{diag}(u_n)(t_n - Wx_n - \bar{t}) \rangle \tag{6.138}
\]

\[
= \frac{1}{ND} \sum_{n=1}^{N} \text{trace} (t_n^Tw_n^T - W(t_n^T(t_n - \bar{t})) + W(t_nx_n^T)x_n^T) \tag{6.139}
\]

which is the same as the maximum likelihood solution in (3.37). For each row of \(W\):

\[
\langle w_{r,:} \rangle = \left( \text{diag}(\langle \delta \rangle) + \langle \lambda \rangle \sum_{n=1}^{N} \langle u_{r,n}x_nx_n^T \rangle \right)^{-1} \langle \lambda \rangle \sum_{n=1}^{N} \langle u_{r,n} \rangle \langle x_n \rangle (t_{r,n} - \langle \bar{t}_r \rangle) \tag{6.140}
\]

\[
= \left( \sum_{n=1}^{N} \langle x_nx_n^T \rangle \right)^{-1} \sum_{n=1}^{N} \langle x_n \rangle t_{r,n} \tag{6.141}
\]

Combining all the rows of \(W\) gives

\[
\langle W \rangle = \left( \sum_{n=1}^{N} \langle x_nx_n^T \rangle \right)^{-1} \sum_{n=1}^{N} \langle x_n \rangle t_n^T \tag{6.142}
\]

which is equivalent to the maximum likelihood solution in (3.36). Finally, for the \(\Theta\):

\[
\langle \theta_q \rangle = \left( \text{diag}(\langle \gamma_q \rangle) + \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle \right)^{-1} \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle x_{q,n} \rangle \langle \tilde{x}_{q,n-1} \rangle \tag{6.143}
\]

\[
= \left( \sum_{n=2}^{N} \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle \right)^{-1} \sum_{n=2}^{N} \langle x_{q,n} \rangle \langle \tilde{x}_{q,n-1} \rangle \tag{6.144}
\]

which again is equivalent to the maximum likelihood solution in (3.34). So the maximum likelihood solution represents the expectation of the equivalent variational result.
From a computational perspective this robust version of PPCA-AR has greater resource usage than the maximum likelihood Gaussian version. More memory is required for both the extra variables (particularly the new latent variables) and for the extra precision matrices. However, the variational version executes more quickly, which appears to be due to the lower computational burden of the variational forwards-backwards algorithm.
7. Robust variational CCA with autoregression

The PPCA-AR model described previously tries to capture spatio-temporal relationships within a single set of multivariate time-series data by using independent autoregressive (AR) processes for each of the latent variables in Probabilistic Principal Component Analysis (PPCA). For EEG analysis researchers are interested in identifying a specific evoked response, i.e. one or more signals that are triggered by some sort of event. These signals may be picked up by more than one of the electrodes positioned around the test subject’s head, so a model such as PPCA-AR could be used to isolate them. The problem is that the electrodes may pick up many shared signals in addition to the evoked response and training PPCA-AR against a single observation set would identify all of them.

A solution is to perform the evoked response test several times and then search for signals that are shared across the multiple observation sets. Consider the case, without loss of generality, of two observation sets, \( T^{(1)} \) and \( T^{(2)} \), where each of them (generically referred to as \( T^{(*)} \)) is made up of \( N \) observation vectors, \( t^{(*)}_1 \) to \( t^{(*)}_N \). The dimensionality of each observation set may be different, say \( D^{(1)} \) and \( D^{(2)} \) respectively. The stacked \( D \)-dimensional observations may be modelled as a latent variable model of a similar form to that of PPCA-AR (with the dependence of each \( x_n \) on \( x_{n-1} \) suppressed for now):

\[
\begin{pmatrix}
  t^{(1)}_n \\
  t^{(2)}_n
\end{pmatrix} =
\begin{pmatrix}
  W^{(1)} \\
  W^{(2)}
\end{pmatrix} x_n +
\begin{pmatrix}
  \bar{t}^{(1)} \\
  \bar{t}^{(2)}
\end{pmatrix} +
\begin{pmatrix}
  e^{(1)}_n \\
  e^{(2)}_n
\end{pmatrix}
\]  

(7.1)

which may be written as

\[
t_n = Wx_n + \bar{t} + e_n
\]  

(7.2)

As before, \( W \) is the \( D \times Q \) mixing matrix, \( x_n \) is the \( n \)th vector of latent variables which are shared between the two observation sets, \( \bar{t} \) allows a non-zero mean observation and \( e_n \) is the observation noise. In Gaussian PPCA-AR the observation noise is assumed to be Gaussian distributed with \( e_n \sim \mathcal{N}(0, \lambda^{-1}I) \). This leads to a likelihood for the observations of \( \mathcal{N}(t_n | Wx_n + \bar{t}, \lambda^{-1}I) \). Note that if we use PPCA-AR in this way, then the maximum value that \( Q \) can be is one smaller that the smaller of the \( D^{(*)} \), i.e. \( Q < \min(D^{(*)}) \).

Each of the observation sets is assumed to be independent. The noise assumption of PPCA-AR preserves this independence, but a more realistic observation noise model for the EEG data would have a full covariance matrix, such that the likelihood for the observations is given by \( \mathcal{N}(t_n | Wx_n + \bar{t}, B^{-1}) \), where \( B \) is a block diagonal precision matrix made up of
the $D^{(1)} \times D^{(1)}$ matrix $B^{(1)}$ and the $D^{(2)} \times D^{(2)}$ matrix $B^{(2)}$ associated with observation sets 1 and 2 respectively. With this assumption in place, the model described thus far is identical to that of Bach and Jordan’s [2005] Probabilistic Canonical Correlation Analysis (ProbCCA). Just as for PPCA-AR, we may incorporate AR processes for each of the latent variables to provide temporal modelling within ProbCCA, resulting in a new model, ProbCCA-AR.

The $Q$-dimensional latent variables in $x_n$ are each modelled as an AR process of order $P$, with $\theta_q$ as the vector of AR coefficients for the $q$th latent variable. Robustness is achieved by modelling the excitation noise, $\epsilon_{q,n}$, as a Student-t distribution with precision $\kappa_q$ and degree of freedom $d_q$:

$$\text{p}(\epsilon_{q,n} | \kappa_q, d_q) = S(\epsilon_{q,n} | 0, \kappa_q, d_q) \quad (7.3)$$

This order $P$ model is converted to a first order model by stacking the latent variable lags into $\tilde{x}_n$ and constructing a block diagonal matrix for the $\theta_q$ in $\Theta$ (see section 3.1), giving the now-familiar combination of expressions

$$\tilde{x}_n = \Theta \tilde{x}_{n-1} + \epsilon_n \quad (7.4)$$
$$t_n = \tilde{W} \tilde{x}_n + \tilde{\bar{t}} + e_n \quad (7.5)$$

where $\tilde{W} = WF$ and $F$ is the fixed matrix of ones and zeros that extracts $x_n$ from $\tilde{x}_n$. As per PPCA-AR the $\kappa_q$ are set to 1 to control the scale ambiguity in the $Wx_n$ term.

The ProbCCA-AR model introduced in this chapter models the observation noise as Gaussian and the AR excitation noise as independent Student-t distributions over each latent variable. From the results in chapter 6 it appears that PPCA-AR tends to capture the heavy-tailed nature of the data in the excitation noise in preference to the observation noise, so this model is expected to be robust to outliers and able to model leptokurtic data. Replacing the Gaussian assumption with a Student-t is left for future work.

As for PPCA-AR, the integrals required for Bayesian ProbCCA-AR are intractable, so a variational approximation method is used, and automatic relevance determination (ARD) [Mackay, 1994; Neal, 1995] is incorporated to estimate the model orders.

The calculations and graphical models in this chapter represent a system with two observation sets, but this is for clarity only; as will be seen in the results section the model may be applied to any number of observation sets.

### 7.1. Priors

Expressing the Student-t distribution as a scale mixture of Gaussians (with the introduction of a new set of latent variables in $z_n$), the noise assumptions lead to the following probability distributions:

$$\text{p}(t_n | W, x_n, \tilde{\bar{t}}, B) = \mathcal{N}(t_n | Wx_n + \tilde{\bar{t}}, B^{-1}) \quad (7.6)$$
\[
p(x_1 | z_1) = \mathcal{N}(x_n | m_0, \text{diag}(z_1)^{-1}) \quad (7.7)
\]
\[
p(\tilde{x}_n | \Theta, x_{n-1}, z_n) = \mathcal{N}(\tilde{x}_n | \Theta x_{n-1}, \text{diag}(\tilde{z}_n)^{-1}) \quad \text{for } n > 1 \quad (7.8)
\]
\[
p(z_{q,n} | d_q) = \mathcal{G}(z_{q,n} | d_{\alpha q}, d_{\beta q}) \quad (7.9)
\]
\[
p(d_q) = \mathcal{G}(d_q | a_d, b_d) \quad (7.10)
\]

where \(\tilde{z}_n\) is the stacked version of \(z_n\), given by \(F^Tz_n\). Note that the first vector of latent variables, in \(x_1\), is assigned a different prior to the others as there is no temporally previous vector for it to depend on.

The observation noise precision, \(B\), is a block diagonal matrix and each block is assigned a Wishart prior, which is the conjugate prior for a Gaussian precision matrix:

\[
p(B^{(\ast)}) = \mathcal{W}(B^{(\ast)} | M_B, v_B) \quad (7.11)
\]

and a Gaussian prior is assigned to the mean observation, \(\bar{t}\):

\[
p(\bar{t}) = \mathcal{N}(\bar{t} | m_{\bar{t}}, V_{\bar{t}}) \quad (7.12)
\]

As with PPCA-AR, the columns of \(W\) are assigned ARD priors controlled by the hyperparameter precisions in \(\delta\):

\[
p(w_{r,:} | \delta) = \mathcal{N}(w_{r,:} | m_w, \text{diag}(\delta)^{-1}) \quad (7.13)
\]

\[
p(\delta_q) = \mathcal{G}(\delta_q | a_\delta, b_\delta) \quad (7.14)
\]

and each AR coefficient in \(\theta_q\) (for the \(q\)th latent variable) is assigned an ARD prior controlled by the hyperparameter precisions in \(\gamma_q\):

\[
p(\theta_q | \gamma_q) = \mathcal{N}(\theta_q | m_{\theta}, \text{diag}(\gamma_q)) \quad (7.15)
\]

\[
p(\gamma_{p,q}) = \mathcal{G}(\gamma_{p,q} | a_{\gamma}, b_{\gamma}) \quad (7.16)
\]

Wang [2007] introduces a variational Bayesian CCA model based on Bach and Jordan’s [2005] ProbCCA. It uses ARD to control the magnitude of the components in each of the \(W^{(\ast)}\), with separate sets of ARD precision parameters for each observation set, such that \(p(w_{r,:}^{(1)} | \delta^{(1)}) = \mathcal{N}(w_{r,:}^{(1)} | m_w, \text{diag}(\delta^{(1)})^{-1})\) and \(p(w_{r,:}^{(2)} | \delta^{(2)}) = \mathcal{N}(w_{r,:}^{(2)} | m_w, \text{diag}(\delta^{(2)})^{-1})\). This definition may be useful (as will be seen in section 7.5), but classical CCA implies that the corresponding components in each \(W^{(\ast)}\) should be switched on or off together.

The graphical model for this system is shown in figure 7.1. Comparing it with the equivalent diagram for robust variational PPCA-AR (in figure 6.1), it may be seen that the dependencies for the parameter variables circled in red (those that are not directly dependent on the likelihood) are identical to those in the robust variational PPCA-AR model and so their variational derivations will be identical.

In the absence of any prior information, uninformative priors are selected: \(m_w = m_0 = m_{\bar{t}} = m_\theta = 0\), \(V_{\bar{t}} = 10^3I\) and \(a_\delta = b_\delta = a_\gamma = b_\gamma = a_d = b_d = 10^{-3}\). The parameters for
7. Robust variational CCA with autoregression

Figure 7.1. Graphical model for the robust variational ProbCCA-AR model with shared ARD parameters over each $W^{(\ast)}$. Those parameters circled in red have identical dependencies to the same parameters in the robust variational PPCA-AR model. Note that the first vector of latent variables, $x_1$, has different dependencies from those shown.

the Wishart prior for the observation noise precision are discussed in section 7.2.3.

7.2. Factorised variational Bayesian method

As remarked above, the parameters highlighted in red in figure 7.1 (that is $\delta$, $\theta_q$, $\gamma_q$, $z_n$ and $d_q$) have the same dependencies as for the robust variational PPCA-AR model and so the derivations of the approximate posterior distributions are exactly the same and are not included in this section, though they are included in the summary on pages 161 and 162. The derivations for the remaining parameters are as follows.

7.2.1. Mean observation, $\bar{t}$

For the mean observation, $\bar{t}$, the only difference in the variational inference is in the form of the observation noise precision matrix, and the result is very similar to that for the
robust variational PPCA-AR model:

\[
\log(q(\bar{\mathbf{t}} | \mathbf{T})) = \mathbb{E}_{\bar{\mathbf{t}}} \left[ \sum_{n=1}^{N} \log(p(t_n | \mathbf{W}, x_n, \bar{\mathbf{t}}, \mathbf{B})) + \log(p(\bar{\mathbf{t}})) \right] + \text{const} \tag{7.17}
\]

\[
= \mathbb{E}_{\bar{\mathbf{t}}} \left[ \sum_{n=1}^{N} \log(\mathcal{N}(t_n | \mathbf{W}x_n + \bar{\mathbf{t}}, \mathbf{B}^{-1})) + \log(\mathcal{N}(\bar{\mathbf{t}} | \mathbf{0}, \mathbf{V}_\bar{\mathbf{t}})) \right] + \text{const} \tag{7.18}
\]

\[
= -\frac{1}{2} \left( \bar{\mathbf{t}}^T (\mathbf{V}_\bar{\mathbf{t}}^{-1} + N(\mathbf{B})) \bar{\mathbf{t}} - 2\bar{\mathbf{t}}^T \mathbf{B} \sum_{n=1}^{N} (t_n - \langle \mathbf{W} \rangle \langle x_n \rangle) \right) + \text{const} \tag{7.19}
\]

It can be seen that \(q(\bar{\mathbf{t}} | \mathbf{T})\) is the Gaussian, \(\mathcal{N}(\bar{\mathbf{t}} | \mu_\bar{\mathbf{t}}, \Sigma_\bar{\mathbf{t}})\), where

\[
\Sigma_\bar{\mathbf{t}} = \left( \mathbf{V}_\bar{\mathbf{t}}^{-1} + N(\mathbf{B}) \right)^{-1} \tag{7.20}
\]

\[
\mu_\bar{\mathbf{t}} = \Sigma_\bar{\mathbf{t}} \langle \mathbf{B} \rangle \sum_{n=1}^{N} (t_n - \langle \mathbf{W} \rangle \langle x_n \rangle) \tag{7.21}
\]

7.2.2. Mixing matrix, \(\mathbf{W}\)

The derivation for the approximate posterior for \(\mathbf{W}\) is similar in form to PPCA-AR (dropping the \(\text{const}\) term from now on to aid readability):

\[
\log(q(\mathbf{W} | \mathbf{T})) = \mathbb{E}_{\mathbf{W}} \left[ \sum_{n=1}^{N} \log(p(t_n | \mathbf{W}, x_n, \bar{\mathbf{t}}, \mathbf{B})) + \sum_{r=1}^{D} \log(p(w_{r,:} | \delta)) \right] \tag{7.22}
\]

\[
= \mathbb{E}_{\mathbf{W}} \left[ \sum_{n=1}^{N} \log(\mathcal{N}(t_n | \mathbf{W}x_n + \bar{\mathbf{t}}, \mathbf{B}^{-1})) + \sum_{r=1}^{D} \log(\mathcal{N}(w_{r,:} | \mathbf{0}, \text{diag}(\delta)^{-1})) \right] \tag{7.23}
\]

\[
= -\frac{1}{2} \mathbb{E}_{\mathbf{W}} \left[ \sum_{n=1}^{N} \left( x_n^T \mathbf{W}^T \mathbf{B} x_n - 2(t_n - \bar{\mathbf{t}})^T \mathbf{B} x_n \right) + \sum_{r=1}^{D} w_{r,:} \text{diag}(\delta) w_{r,:}^T \right] \tag{7.24}
\]

but the precision matrix leads to a more complicated calculation for the terms based on the likelihood (the scalar precision, \(\lambda\), for PPCA-AR enabled a straightforward rearrangement). The first term may be rearranged as follows:

\[
x_n^T \mathbf{W}^T \mathbf{B} x_n = \text{trace}(\mathbf{W} x_n x_n^T \mathbf{W}^T) \tag{7.25}
\]

\[
= \sum_{r=1}^{D} w_{r,:} x_n x_n^T \mathbf{W}^T \mathbf{B}_{:,r} \tag{7.26}
\]

\[
= \sum_{r=1}^{D} \text{trace}(\mathbf{B}_{:,r} w_{r,:} x_n x_n^T \mathbf{W}^T) \tag{7.27}
\]

\[
= \sum_{r=1}^{D} \sum_{s=1}^{D} \mathbf{B}_{s:r} w_{r,:} x_n x_n^T w_{s,:}^T \tag{7.28}
\]
7. Robust variational CCA with autoregression

\[
D \sum_{r=1}^{D} \left( w_r; B_{r,r} x_n x_n^T w_{r,:}^T + w_{r,:} \sum_{s=1}^{D} B_{s,r} x_n x_n^T w_{s,:}^T \right) = \sum_{r=1}^{D} \left( w_{r,:} x_n (t_n - \bar{t})^T B_{r,:} \right) \quad (7.29)
\]

The approximate posterior distributions for each row of \( W \) are independent. This is a quadratic expression in \( w_r,: \). The second term of (7.24) may also be rearranged to give an expression in \( w_r,: \):

\[
(t_n - \bar{t})^T B W x_n = \text{trace}(W x_n (t_n - \bar{t})^T B) = \sum_{r=1}^{D} w_{r,:} x_n (t_n - \bar{t})^T B_{r,:} \quad (7.30)
\]

Putting (7.29) and (7.31) back into (7.24) and now considering each row of \( W \) separately gives

\[
\log(q(w_r,: \mid T)) = -\frac{1}{2} \mathbb{E}_{B \mid W} \left[ \sum_{n=1}^{N} \left( w_{r,:} B_{r,r} x_n x_n^T w_{r,:}^T + w_{r,:} \sum_{s=1}^{D} B_{s,r} x_n x_n^T w_{s,:}^T - 2 w_{r,:} x_n (t_n - \bar{t})^T B_{r,:} \right) \right] = -\frac{1}{2} \left[ w_{r,:} \left( \text{diag}(\delta) + \langle B_{r,:} \rangle \sum_{n=1}^{N} \langle x_n x_n^T \rangle \right) w_{r,:}^T - 2 w_{r,:} \sum_{n=1}^{N} \langle x_n \rangle (t_n - \bar{t})^T \langle B_{r,:} \rangle - \frac{1}{2} \sum_{s=1}^{D} \langle B_{s,r} \rangle \langle x_n x_n^T \rangle \langle w_{s,:}^T \rangle \right] \quad (7.32)
\]

It can be seen that \( q(w_r,: \mid T) \) is the Gaussian, \( \mathcal{N}(w_r,: \mid \mu_{w_r}, \Sigma_{w_r}) \), where

\[
\Sigma_{w_r} = \left( \text{diag}(\delta) + \langle B_{r,:} \rangle \sum_{n=1}^{N} \langle x_n x_n^T \rangle \right)^{-1} \quad (7.34)
\]

\[
\mu_{w_r} = \Sigma_{w_r} \sum_{n=1}^{N} \langle x_n \rangle (t_n - \bar{t})^T \langle B_{r,:} \rangle - \frac{1}{2} \sum_{s=1}^{D} \langle B_{s,r} \rangle \langle x_n x_n^T \rangle \langle w_{s,:}^T \rangle \quad (7.35)
\]

7.2.3. Precision matrix, \( B \)

The approximate posterior for each \( B^{(\ast)} \) must be calculated separately. In the following expressions the \( \ast \) has been dropped from each variable for clarity.

\[
\log(q(B \mid T)) = \mathbb{E}_{B \mid W} \left[ \log(p(B)) + \sum_{n=1}^{N} \log(p(t_n \mid W, x_n, \bar{t}, B)) \right] \quad (7.36)
\]
A straightforward expansion of the summation term gives (dropping the

\[ E_B \left[ \log(W(B | M_B, v_B)) + \sum_{n=1}^{N} \log(N(t_n | Wx_n + \bar{t}, B^{-1})) \right] \quad (7.37) \]

\[ = \frac{1}{2} E_B \left[ (v_B - D + 1) \log |B| - \text{trace}(M_B^{-1}B) + N \log |B| \right. \]
\[ - \sum_{n=1}^{N} \text{trace} \left( (t_n - Wx_n - \bar{t})(t_n - Wx_n - \bar{t})^T B \right) \]
\[ = \frac{v_B + N - D - 1}{2} \log |B| \]
\[ - \frac{1}{2} \text{trace} \left( \left( M_B^{-1} + \sum_{n=1}^{N} (t_n - Wx_n - \bar{t})(t_n - Wx_n - \bar{t})^T \right) B \right) \quad (7.39) \]

It can be seen that each \( q(B^{(*)} | T) \) is a Wishart, \( W(B^{(*)} | \Delta^{(*)}, \chi) \), where

\[ \chi = v_B + N \]
\[ \Delta^{(*)} = \left( M_B^{-1} + \sum_{n=1}^{N} (t_n^{(*)} - W^{(*)}x_n - \bar{t}^{(*)})(t_n^{(*)} - W^{(*)}x_n - \bar{t}^{(*)})^T \right)^{-1} \quad (7.41) \]

We choose a very vague prior for each \( B^{(*)} \), with \( v_B \to 0 \) and \( M_B^{-1} \to 0 \).

A straightforward expansion of the summation term gives (dropping the \(^{*}\) for clarity)

\[ \langle (t_n - Wx_n - \bar{t})(t_n - Wx_n - \bar{t})^T \rangle \]
\[ = t_n^T t_n + \langle Wx_n x_n^T W^T + (\bar{t} \bar{t}^T) \rangle \]
\[ - t_n \left( \langle W \rangle \langle x_n \rangle + \langle \bar{t} \rangle \right)^T - \left( \langle W \rangle \langle x_n \rangle + \langle \bar{t} \rangle \right) t_n^T \]
\[ + \langle W \rangle \langle x_n \rangle \langle \bar{t} \rangle + \langle \bar{t} \rangle \langle x_n \rangle^T \langle W \rangle \]
\[ = t_n^T t_n + \langle Wx_n x_n^T W^T + (\bar{t} \bar{t}^T) \rangle \quad (7.42) \]

The \( \langle Wx_n x_n^T W^T \rangle \) matrix may be constructed on an element-by-element basis, taking advantage of the fact that the rows of \( W \) are independent:

\[ \langle Wx_n x_n^T W^T \rangle_{r,s} = \begin{cases} \text{trace}(\langle x_n x_n^T \rangle \langle w_{r,:} \rangle \langle w_{s,:}^T \rangle) & \text{if } r = s \\ \langle w_{r,:} \rangle \langle x_n x_n^T \rangle \langle w_{s,:}^T \rangle & \text{if } r \neq s \end{cases} \quad (7.43) \]

### 7.2.4. Latent variables, \( x_n \)

In section 6.1.1 a robust variational version of the forward-backward algorithm is described for the case where the observation noise matrix is diagonal. In the section this is further generalised to take into account the assumption that the noise has a full precision matrix.
Forward sweep

Following the same procedure as described in section 6.1.1, the approximate posterior $q^{(a)}(\tilde{x}_n) \approx \alpha(\tilde{x}_n) = p(\tilde{x}_n | t_1, \ldots, t_n)$ is found from:

$$
\log(q^{(a)}(\tilde{x}_n)) = \mathbb{E}_{\tilde{x}_n} \left[ \log(p(t_n | \tilde{W}, \tilde{x}_n, \tilde{t}, B)) + \log(p(\tilde{x}_n | \Theta, \tilde{x}_{n-1}, \tilde{z}_n)) \right]
$$

$$
= \mathbb{E}_{\tilde{x}_n} \left[ \log(p(t_n | \tilde{W}\tilde{x}_n + \tilde{t}, B^{-1})) + \log(N(\tilde{x}_n | \Theta\tilde{x}_{n-1}, \text{diag}(\tilde{z}_n)^{-1})) \right]
$$

$$
= -\frac{1}{2} \mathbb{E}_{\tilde{x}_n} \left[ \tilde{x}_n^T \tilde{W}^T B \tilde{x}_n - 2\tilde{x}_n^T \tilde{W}^T B(t_n - \tilde{t}) \right. \\
- \tilde{x}_n^T \text{diag}(\tilde{z}_n)\tilde{x}_n - 2\tilde{x}_n^T \text{diag}(\tilde{z}_n)\Theta\tilde{x}_{n-1} \\
$$

$$
= -\frac{1}{2} \left[ \tilde{x}_n^T \left( \langle \tilde{W}^T B \tilde{W} \rangle + \text{diag}(\tilde{z}_n) \right) \tilde{x}_n \\
- 2\tilde{x}_n^T \left( \langle \tilde{W}^T \rangle (t_n - \langle \tilde{t} \rangle) + \text{diag}(\langle \tilde{z}_n \rangle) \langle \tilde{x}_{n-1} \rangle \right) \right]
$$

The $\langle \tilde{W}^T B \tilde{W} \rangle$ term may be expressed as $F^T (\tilde{W}^T B \tilde{W}) F$ and constructed on an element-by-element basis, taking advantage of the fact that the rows of $W$ are independent:

$$
\langle \tilde{W}^T B \tilde{W} \rangle_{i,j} = \langle w_{\cdot,i} \tilde{W} \tilde{W} \rangle_{i,j} = \langle \sum_{r=1}^D \sum_{s=1}^D w_{r,i} B_{r,s} w_{s,j} \rangle
$$

$$
= \sum_{r=1}^D \left( \langle B_{r,r} \rangle \langle w_{r,i} w_{r,j} \rangle + \sum_{s=1}^D \langle B_{r,s} \rangle \langle w_{r,i} \rangle \langle w_{s,j} \rangle \right)
$$

where $\langle w_{r,i} w_{r,j} \rangle$ is just $\langle w_{r,i} \cdot w_{r,j} \rangle_{i,j}$. This piecemeal construction leads to the following definition:

$$
\langle \tilde{W}^T B \tilde{W} \rangle = \sum_{r=1}^D \left( \langle B_{r,r} \rangle \langle \tilde{w}_{r,i} \cdot \tilde{w}_{r,j} \rangle + \sum_{s=1}^D \langle B_{r,s} \rangle \langle \tilde{w}_{r,i} \rangle \langle \tilde{w}_{s,j} \rangle \right)
$$

It can be seen that $q^{(a)}(\tilde{x}_n)$ is a Gaussian, $N(\tilde{x}_n | \tilde{\mu}^{(a)}_{\tilde{x}_n}, \tilde{\Sigma}^{(a)}_{\tilde{x}_n})$, where

$$
\tilde{\Sigma}^{(a)}_{\tilde{x}_n} = \left( \langle \tilde{W}^T B \tilde{W} \rangle + \text{diag}(\langle \tilde{z}_n \rangle) \right)^{-1}
$$

$$
\tilde{\mu}^{(a)}_{\tilde{x}_n} = \tilde{\Sigma}^{(a)}_{\tilde{x}_n} \left( \langle \tilde{W}^T \rangle (t_n - \langle \tilde{t} \rangle) + \text{diag}(\langle \tilde{z}_n \rangle) \langle \tilde{x}_{n-1} \rangle \right)
$$

A slightly different formulation is required for $\tilde{x}_1$ as it has a different prior due to its position at the beginning of the time sequence:

$$
\log(q^{(a)}(\tilde{x}_1)) = \mathbb{E}_{\tilde{x}_1} \left[ \log(p(t_1 | \tilde{W}, \tilde{x}_1, \tilde{t}, B)) + \log(p(\tilde{x}_n | 0, \tilde{z}_n)) \right]
$$
7. Robust variational CCA with autoregression

\[ \log(p(t_1 | \tilde{W}\tilde{x}_1 + \tilde{t}, B^{-1})) + \log(N(\tilde{x}_1 | 0, \text{diag}(\tilde{z}_n)^{-1})) \]  

(7.55)

which gives rise to an approximate posterior distribution for \( \tilde{x}_1 \) of \( N(\tilde{x}_1 | \tilde{\mu}_{x_1}^{(a)}, \tilde{\Sigma}_{x_1}^{(a)}) \), where

\[
\tilde{\Sigma}_{x_1}^{(a)} = \left( \langle \tilde{W}^T B \tilde{W} \rangle + \text{diag}(\langle \tilde{z}_1 \rangle) \right)^{-1} 
\]

(7.56)

\[
\tilde{\mu}_{x_1}^{(a)} = \tilde{\Sigma}_{x_1}^{(a)} \langle \tilde{W} \rangle^T (B)(t_1 - \langle \tilde{t} \rangle) 
\]

(7.57)

**Backward sweep**

For the last instance of the latent variables in the time series, i.e. where \( n = N \), the forward sweep results in \( \alpha(\tilde{x}_N) = p(\tilde{x}_n | t_1, \ldots, t_N) \). This is just \( p(\tilde{x}_n | T) \), the posterior probability for \( \tilde{x}_n \). So where \( n = N \) the posterior distribution is the Gaussian \( N(\tilde{x}_N | \tilde{\mu}_N^{(a)}, \tilde{\Sigma}_N^{(a)}) \).

For the remaining \( \tilde{x}_n \) for values of \( n \) from \( N - 1 \) down to 1, the \( \alpha(\tilde{x}_n) \approx q^{(a)}(\tilde{x}_n) \) from the forward sweep is combined with \( \beta(\tilde{x}_n) \) in the backward sweep, giving an expression that is identical to that for PPCA-AR in (6.39) in section 6.1.1:

\[ \log(q(\tilde{x}_n | T)) = \mathbb{E}_{x_n} \left[ \log(q^{(a)}(\tilde{x}_n)) + \log(p(\tilde{x}_{n+1} | \tilde{x}_n, \Theta, \tilde{z}_{n+1})) \right] \]

(7.58)

Hence the result is the same: \( q(\tilde{x}_n | T) \) is the Gaussian \( N(\tilde{x}_n | \tilde{\mu}_n, \tilde{\Sigma}_n) \), where

\[
\tilde{\Sigma}_n = \left( \langle \tilde{\Sigma}_n^{(a)} \rangle^{-1} + \langle \Theta^T \text{diag}(\tilde{z}_{n+1}) \Theta \rangle \right)^{-1} 
\]

(7.59)

\[
\tilde{\mu}_n = \tilde{\Sigma}_n \left( \langle \tilde{\Sigma}_n^{(a)} \rangle^{-1} \tilde{\mu}_n^{(a)} + \langle \Theta^\top \rangle \text{diag}(\langle \tilde{z}_n \rangle) \langle \tilde{x}_{n+1} \rangle \right) 
\]

(7.60)

Here \( \tilde{\Sigma}_n \) is dependent on the expectation \( \langle \Theta^T \text{diag}(\tilde{z}_{n+1}) \Theta \rangle \). If we consider each latent variable separately, then this expectation for the \( q \)th latent variable for the \( n \)th sample is \( \langle \theta_q \theta_q^\top \rangle_{\tilde{z}_{n+1}} \). The full \( \langle \Theta^\top \rangle \text{diag}(\tilde{z}_{n+1}) \Theta \) may be constructed as a block diagonal matrix of these individual matrices.

The posterior distribution for \( x_n \) may easily be extracted from that for \( \tilde{x}_n \) using the \( F \) matrix, giving \( q(x_n | T) \) as \( N(x_n | F\tilde{\mu}_n, F\tilde{\Sigma}_nF^\top) \).

**7.2.5. Summary**

The full results for robust variational ProbCCA-AR, incorporating the inherited derivations from chapter 6, are shown on pages 161 and 162.

Each approximate posterior distribution is dependent on the expected values of one or more of the others, so a closed-form algebraic solution can not be obtained. We may arrive at a set of solutions by initialising the required expectations (perhaps based on the results of maximum likelihood Gaussian PPCA-AR) and then iteratively updating the estimate for each hyperparameter based on the current estimates of the values on which it depends, until convergence. The required current expectations are obtained using the
7. Robust variational CCA with autoregression

The variables’ distributions were calculated in the order shown below.

\[
q^{(\alpha)}(\tilde{x}_n) = \mathcal{N}(\tilde{x}_n | \tilde{\mu}_n^{(\alpha)}, \tilde{\Sigma}_n^{(\alpha)}) \tag{7.61}
\]

\[
\tilde{\Sigma}_n^{(\alpha)} = \left(\tilde{W}^T B \tilde{W} + \text{diag}((\tilde{z}_n))\right)^{-1} \tag{7.62}
\]

\[
\tilde{\mu}_1^{(\alpha)} = \Sigma_1^{(\alpha)} \langle \tilde{W} \rangle^T (B)(t_1 - \langle \tilde{t} \rangle) \tag{7.63}
\]

\[
\tilde{\mu}_n^{(\alpha)} = \Sigma_n^{(\alpha)} \left(\langle \tilde{W} \rangle^T (B)(t_n - \langle \tilde{t} \rangle) + \text{diag}((\tilde{z}_n))\langle \tilde{x}_{n-1} \rangle\right) \tag{7.64}
\]

\[
q(x_n | T) = \mathcal{N}(x_n | F \tilde{\mu}_n, F \Sigma_n F^T) \tag{7.65}
\]

\[
\Sigma_n^{-1} = \left(\Sigma_n^{(\alpha)}\right)^{-1} + (\Theta^T \text{diag}(\tilde{z}_n) \Theta) \tag{7.66}
\]

\[
\Sigma_N = \tilde{\Sigma}_N^{(\alpha)} \tag{7.67}
\]

\[
\tilde{\mu}_n = \Sigma_n \left(\Sigma_n^{(\alpha)}\right)^{-1} \tilde{\mu}_n^{(\alpha)} + (\Theta^T \text{diag}(\tilde{z}_n))\langle \tilde{x}_{n+1} \rangle \tag{7.68}
\]

\[
\tilde{\mu}_N = \tilde{\Sigma}_N^{(\alpha)} \tag{7.69}
\]

\[
q(w_{r,:} | T) = \mathcal{N}(w_{r,:} | \mu_{w_r}, \Sigma_{w_r}) \tag{7.70}
\]

\[
\Sigma_{w_r}^{-1} = \text{diag}(\tilde{\delta}) + \langle B_{r,r} \rangle \sum_{n=1}^{N} \langle x_n x_n^T \rangle \tag{7.71}
\]

\[
\mu_{w_r} = \Sigma_{w_r} \sum_{n=1}^{N} \left(\langle x_n \rangle\langle t_n - \langle \tilde{t} \rangle \rangle^T \langle B_{r,r} \rangle - \frac{1}{2} \sum_{s=1}^{D} \langle B_{s,r} \rangle \langle x_n x_n^T \rangle \langle w_{s,:} \rangle^T \right) \tag{7.72}
\]

\[
q(\delta_q | T) = \mathcal{G}(\delta_q | \alpha_\delta, \beta_\delta) \tag{7.73}
\]

\[
\alpha_\delta = a_\delta + D/2 \tag{7.74}
\]

\[
\beta_\delta = b_\delta + \frac{1}{2} \sum_{r=1}^{D} (w_{r,q})^2 \tag{7.75}
\]

\[
q(t | T) = \mathcal{N}(t | \mu_t, \Sigma_t) \tag{7.76}
\]

\[
\Sigma_t^{-1} = V_t^{-1} + N(B) \tag{7.77}
\]

\[
\mu_t = \Sigma_t (B) \sum_{n=1}^{N} \langle t_n - \langle W \rangle \langle x_n \rangle \rangle \tag{7.78}
\]

\[
q(B^{(\gamma)}) | T) = \mathcal{W}(B^{(\gamma)} | \Delta^{(\gamma)}, \chi) \tag{7.79}
\]

\[
\chi = N \tag{7.80}
\]

\[
(\Delta^{(\gamma)})^{-1} = \sum_{n=1}^{N} ((t_n^{(\gamma)} - W^{(\gamma)} x_n - \tilde{t}^{(\gamma)})(t_n^{(\gamma)} - W^{(\gamma)} x_n - \tilde{t}^{(\gamma)})^T) \tag{7.81}
\]

\[
q(\theta | T) = \mathcal{N}(\theta | \mu_\theta, \Sigma_\theta) \tag{7.82}
\]

\[
\Sigma_{\theta}^{-1} = \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle \tilde{x}_{q,n-1} \rangle^T + \text{diag}(\gamma_q) \tag{7.83}
\]

\[
\mu_{\theta} = \Sigma_{\theta} \sum_{n=2}^{N} \langle z_{q,n} \rangle \langle x_{q,n} \rangle \langle \tilde{x}_{q,n-1} \rangle \tag{7.84}
\]

\[
q(\gamma_{p,q} | T) = \mathcal{G}(\gamma_{p,q} | \alpha_\gamma, \beta_{p,q}) \tag{7.85}
\]

\[
\alpha_\gamma = a_\gamma + 1 \tag{7.86}
\]

\[
\beta_{p,q} = b_\gamma + \frac{1}{2} (\theta_{p,q}^2) \tag{7.87}
\]
7. Robust variational CCA with autoregression

The following definitions are required (with \( \text{blkdiag}(\cdot) \) being a block diagonal matrix):

\[
\begin{align*}
q(z_{q,n} \mid \mathbf{T}) &= \mathcal{G}(z_{q,n} \mid \alpha_z, \beta_{z_{q,n}}) \quad (7.88) \\
\alpha_z &= \frac{\langle d_q \rangle + 1}{2} \quad (7.89) \\
\beta_{z_{q,n}} &= \frac{1}{2} \left( \langle d_q \rangle + \langle x_{q,n}^2 \rangle - 2\langle x_{q,n} \rangle \langle \theta_q^T \rangle \langle \tilde{x}_{q,n-1} \rangle \\
&\quad + \text{trace}(\langle \theta_q \theta_q^T \rangle \langle \tilde{x}_{q,n-1} \tilde{x}_{q,n-1}^T \rangle) \right) \quad (7.90) \\
q(d_q \mid \mathbf{T}) &= \mathcal{G}(d_q \mid \alpha_d, \beta_{d_q}) \quad (7.91) \\
\alpha_d &= a_d + N/2 \quad (7.92) \\
\beta_{d_q} &= b_d - \frac{1}{2} \left( N + \sum_{n=1}^{N} (\langle \log(z_{q,n}) \rangle - \langle z_{q,n} \rangle) \right) \quad (7.93)
\end{align*}
\]

standard expressions (see appendix A)

\[
\begin{align*}
\langle w_{r,:} \rangle &= \mu_w \quad (7.99) \\
\langle w_{r,:}^T w_{r,:) \rangle &= \Sigma_w + \mu_w^T \mu_w \quad (7.100) \\
\langle w_{r,q}^2 \rangle &= \langle w_{r,:}^T w_{r,:) \rangle_{q,q} \quad (7.101) \\
\langle \delta_q \rangle &= \alpha_d / \beta_{d_q} \quad (7.102) \\
\langle \mathbf{B}^{(*)} \rangle &= \chi \Delta^{(*)} \quad (7.103) \\
\langle u_{r,n} \rangle &= \alpha_u / \beta_{u_{r,n}} \quad (7.104) \\
\langle \log(u_{r,n}) \rangle &= \psi(\alpha_u) - \log(\beta_{u_{r,n}}) \quad (7.105) \\
\langle x_{n} \rangle &= \mu_x \quad (7.106) \\
\langle x_{n} x_{n}^T \rangle &= \Sigma_x + \mu_x^T \mu_x \quad (7.107) \\
\langle x_{q,n}^2 \rangle &= \langle x_{n} x_{n}^T \rangle_{q,q} \quad (7.108)
\end{align*}
\]
\[\langle \bar{t} \rangle = \mu_t \quad (7.109)\]

\[\langle \bar{t}^T \rangle = \Sigma_t + \mu_t \mu_t^T \quad (7.110)\]

\[\langle \bar{t}_r^2 \rangle = \langle \bar{t}^T \rangle_{r,r} \quad (7.111)\]

\[\langle z_{q,n} \rangle = \alpha_z / \beta_{z_{q,n}} \quad (7.112)\]

\[\langle \log(z_{q,n}) \rangle = \psi(\alpha_z) - \log(\beta_{z_{q,n}}) \quad (7.113)\]

\[\langle v \rangle = \alpha_v / \beta_v \quad (7.114)\]

\[\langle d_{q} \rangle = \alpha_d / \beta_{d_{q}} \quad (7.115)\]

\[\langle \theta_q \rangle = \mu_{\theta} \quad (7.116)\]

\[\langle \theta_q \theta_q^T \rangle = \Sigma_{\theta} + \mu_{\theta} \mu_{\theta}^T \quad (7.117)\]

\[\langle \theta_{p,q}^2 \rangle = \langle \theta_{p,q} \theta_{q}^T \rangle_{p,p} \quad (7.118)\]

\[\langle \gamma_{p,q} \rangle = \alpha_\gamma / \beta_{\gamma_{p,q}} \quad (7.119)\]

where \(\psi(\cdot)\) is the digamma function.

Having formulated this model, straightforward adjustments can be made to change the assumptions of the distributions of the excitation noise. If all the values in the excitation noise latent variables (the \(z_{n}\)) take the value 1, then the Student-t distribution from (7.3), which may be written as \(\int \mathcal{N}(\epsilon_{q,n} | 0, z_{q,n}^{-1}) \mathcal{G}(z_{q,n} | d_q, d_q) \, dz_{q,n}\), becomes just \(\mathcal{N}(\epsilon_{q,n} | 0, 1)\) and the excitation noise assumption is now Gaussian.

The expectations of the model’s variables are initialised by running a few iterations of robust variational PPCA-AR against the stacked observation sets (see equations (7.1) and (7.2)). This provides straightforward starting positions for all variables apart from the observation noise precisions in \(B^{(\cdot)}\). The latter are initialised to \(\lambda I\), where \(\lambda\) is the scalar observation noise precision from PPCA-AR.

### 7.3. Illustration: synthetic data

The main concern for CCA is to identify latent signals that are shared between observation sets. As such the illustrations in this section concentrate on the model’s ability to estimate the values of the \(W\) and \(\theta_q\) values and the latent variables over a range of different noise distributions and to estimate the correct model orders.

Synthetic data was generated as for the robust variational PPCA-AR model (see section 6.3), but a single set of latent variables was used to generate all the observation sets for a single test of ProbCCA-AR. For each dataset, two observation sets of 200 samples were generated from the same set of 2-dimensional latent variables, i.e. \(N = 200\) and \(Q = 2\). The dimensionality of the observation sets was \(D^{(1)} = 5\) and \(D^{(2)} = 4\). Each latent variable was generated from an AR process of order \(P = 5\), with different \(\theta_q\) and a Student-t excitation noise distribution having precision \(\kappa_q = 1\) and degrees of freedom, \(d_q\), set to one of \(\{2,3,4,100\}\).
The precision matrix, $B$, was constructed by generating a random $D^{(2)} \times D^{(2)}$ matrix $A$, with Gaussian distributed elements, calculating the sample covariance matrix, $AA^T$, multiplying it by one of \{0.01,1,100\} to give low, medium and high observation noise and then setting $B$ to be its inverse.

Against each of these datasets, robust variational ProbCCA-AR was trained for 2000 iterations with $Q$ set to 3 (one lower than the minimum $D^{(2)}$) and $P = 8$. It would be expected that ARD should cause one of the components to be switched off and three of the AR coefficients in each case.

### 7.3.1. Illustration 1: parameter estimation

Taking one example for each of the low, medium and high observation noise datasets, figure 7.2 shows Hinton diagrams of examples of (row 1) actual, (row 2) estimated and (row 3) transformed estimated $W$ matrices and (row 4) the Procrustes transformation matrix where the excitation noise distribution is heavy-tailed ($d_q = 2$). Figure 7.3 shows the same but for cases where the excitation noise is approximately Gaussian ($d_q = 100$). Note that the posterior standard deviations are very small compared with the magnitudes of the elements in the estimated $W$ matrices in each case.

When the excitation noise is heavy-tailed, the estimates of $W$ are generally good, with the unavoidable sign and permutation ambiguities, but no significant rotation. The scale of each component is different compared with the actuals because any variance in the latent variables is pushed into the components. In the low noise example in (d) the third component has not been switched off, but it is very similar to the second component. In the high noise example in (f) the third component has remained switched on to accommodate the extra noise. In these two cases the components best correlated with the actuals have been used in the Procrustes transformation.

When the excitation noise is approximately Gaussian (figure 7.3) it seems that ARD is not able to switch off components, even for low levels of observation noise.

Using the same examples, figures 7.4 shows box plots of the estimated $\theta_q$ values (in red, with one standard deviation above and below shown as a blue box) plotted over the actual values (in black) used to generate the data. There are two graphs for each example, one for each latent variable. For low and medium observation noise (in (a),(b), (d) and (e)) ARD seems to be generally successful at switching off the appropriate components, even for the Gaussian excitation noise examples. The estimates in the heavy-tailed excitation noise, low and medium observation noise examples (in (a) and (b)) are similar to the actuals. As has previously been noted, the $\theta_q$ values are generally underestimated.

The ProbCCA-AR model constrains the latent variables to have a variance of 1. In order to compare the estimated latent variables with those that were used to generate the data, they must be scaled appropriately. Figure 7.5 is a plot of the actual values against the scaled estimated values where the excitation noise is heavy-tailed and the observation noise is (a) low, (b) medium and (c) high. The actual latent variables are shown in blue.
Figure 7.2. Hinton diagrams of (row 1) actual, (row 2) estimated and (row 3) transformed estimated $W$ and (row 4) the Procrustes transformation matrix for (left) low, (centre) medium and (right) high observation noise. The excitation noise has a heavy-tailed distribution ($d_q = 2$).
Figure 7.3. Hinton diagrams of (row 1) actual, (row 2) estimated and (row 3) transformed estimated $W$ and (row 4) the Procrustes transformation matrix for (left) low, (centre) medium and (right) high observation noise. The excitation noise has an approximately Gaussian distribution ($d_q = 100$).
Figure 7.4. Plots showing the mean (red) and one standard deviation (blue) of the estimated \( \theta_q \) values superimposed on the actual values (black) used to generate the data for (left) low, (centre) medium and (right) high observation noise and excitation noise with (top) a heavy-tailed distribution (\( d_q = 2 \)) and (bottom) an approximately Gaussian distribution (\( d_q = 100 \)).
Figure 7.5. Plots of actual latent variable values (in blue) against scaled estimated values (in red, with one standard deviation shown in green) for (a) low, (b) medium and (c) high observation noise and excitation noise with a heavy-tailed distribution ($d_q = 2$).
Figure 7.6. Plots of actual latent variable values (in blue) against scaled estimated values (in red, with one standard deviation shown in green) for (a) low, (b) medium and (c) high observation noise and excitation noise with an approximately Gaussian distribution ($d_q = 100$).
with the estimates in red. Green indicates the region of one standard deviation above and below the estimate. The scaling factor is shown in red on each plot. Where none of the components have been switched off, the latent variables associated with those components included in the Procrustes transformation described above are drawn. Note that while the sign and permutation have been corrected, no rotation has been applied to the latent variables. For low and medium observation noise the fit is remarkably good.

Figure 7.6 shows the same, but for the Gaussian excitation noise examples. In this case the latent variables have been rotated in accordance with the Procrustes transformation calculated above. The estimates do not seem to be as accurate as they are for the heavy-tailed noise and the standard deviations are more prominent.

### 7.3.2. Illustration 2: model order estimation

As stated above, the actual model orders for each of the datasets was $Q = 2$ and $P = 5$. With the model trained with values of $Q = 3$ and $P = 8$, one component (i.e. one vector in $W$) and three AR coefficients should be switched off by ARD in each case.

Figure 7.7 shows the median and quartile ranges for the columns of $W$. The columns for each individual $W$ have sorted into descending order of magnitude to allow for the permutation ambiguity. It can clearly be seen that the model reliably switches off the third component.

Figure 7.8 compares the median and quartiles of the elements in each $\theta_q$: plot (a) shows the actual AR coefficients used to generate the data while (b) shows those estimated by the model. Where a component in $W$ has been switched off, the corresponding $\theta_q$ is also set to zero, so these vectors have not been included in the calculation of the mean and quartile values in (b). As was seen for the PPCA-AR model, those AR coefficients $\theta_{p,q}$ where $p$ is greater than the actual model order of 5 are clearly being constrained in magnitude to be close to zero, but are not being entirely switched off, and the magnitude of the coefficients is generally underestimated.
7. Robust variational CCA with autoregression

7.4. Results: real data

For the EEG data we wish to discover an underlying signal that is shared between multiple observation sets. In order to achieve this it is desirable to train the model against as many observation sets as possible. Each observation set in this example is made up of the measurements from 62 electrodes over 800 time steps. However, training the model against multiple sets of 62 signals is very slow, so a subset of 10 electrodes has been selected in which an evoked response has been identified [Lavric, 2009]. The model was trained for these 10 signals in each of 5 observation sets for 16700 iterations, with $Q$ set to 9 (one less than $D$) and $P$ to 30.

Figure 7.9 shows the Hinton diagram for the estimated $W$. While ARD has not entirely switched off any components, it is clear that the second component is much more significant than the rest. This is made obvious in figure 7.10 which shows the median and quartiles for each component.

The significant second component is plotted in figure 7.11 split into the subvectors associated with each observation set. There is a similarity in the shapes of each of these plots, with higher contributions from the latent signal detected around electrodes 43 and 44 and lower contributions at electrodes 47, 48 and 40.

Figure 7.12 shows the posterior mean and one standard deviation for each element in the first two components in $W$. The standard deviations are generally small with respect to the magnitudes of the means.

The latent signal associated with the significant component is shown in figure 7.13. The black dotted line marks the time step at which the trigger for the evoked response was made; there does appear to be some reaction to it in the signal.

Figure 7.14 shows the AR coefficients, $\theta_q$, for each of the 9 latent signals. As was noted in
Figure 7.9. The $W$ estimated by the model. While ARD has not entirely switched off any components, clearly the second component is more significant than the others.

Figure 7.10. The median (in red) and quartile values (in blue) for the components in the estimated $W$, highlighting the significance of the second component.
Figure 7.11. The significant component split into the subvectors associated with each observation set and plotted separately. There is a similarity in the shapes of these plots, with higher contributions from the latent signal detected around electrodes 43 and 44 and lower contributions at electrodes 47, 48 and 40.

Figure 7.12. The mean (red) and one standard deviation (blue) of the elements in (a) the first component and (b) the significant second component in $W$. 
7. Robust variational CCA with autoregression

Figure 7.13. The latent signal associated with the significant component. The black dotted line marks the time step at which the trigger for the evoked response was made.

Figure 7.14. Surface plot of the AR coefficients, $\theta_{pq}$, associated with each of the 9 latent variables. As for PPCA-AR in chapter 6 the set is remarkably uniform.

Figure 7.15. Heatmap of the observation noise precision matrix, $B$, estimated by the model. The black lines outline each of the five individual $B^{(\ast)}$ (for the five sets of observations) that make up the block-diagonal matrix.
chapter 6, the model finds a remarkably uniform set. An important consequence of this is that power spectral densities calculated from the AR coefficients estimated by this model are consistent across a subject test.

Figure 7.15 shows a heat map of the estimated observation noise precision matrix $B$. The matrix is block diagonal, with each block (outlined in black) calculated separately; the off-diagonal blocks are zero. While there is a strong diagonal element, it is clear that the data indicates full precision matrices for each observation set.

The observation noise in this model is assumed to be Gaussian. Interestingly, given that the data are leptokurtic in nature, the estimated degrees of freedom for the Student-t excitation noise are all large (greater than 300), implying that the signals are being modelled as Gaussian. Each of the latent variables is an average over the five observation sets; for a spike to appear in the latent variables it would have to appear in all the observation sets. Future work on this model will assume that the observation noise is Student-t distributed to capture the leptokurtic nature of the data.

7.5. Extension to the ProbCCA-AR model

The model described thus far (“classical” ProbCCA-AR), in line with traditional CCA, assumes that all the observation sets share one or more latent signals, while PPCA-AR finds latent signals that are shared by the variables in a single set of observations. A hybrid of these two models (“hybrid” ProbCCA-AR), and a generalisation of both of them, is able to identify latent signals that support any subset of the observation sets, including the special cases of one or all observation sets. Using classical ProbCCA-AR against EEG data, the model may be trained against observation sets for one evoked response and then, separately, for controls (or a different evoked response), and the differences determined by inspecting the structure of the results. Hybrid ProbCCA-AR may be trained against all the observation sets in one go, the expectation being that those observation sets in the evoked response group should be easily distinguished from those in the control group by the structure of the components.

Hybrid ProbCCA-AR is derived from classical ProbCCA-AR by a simple change to the ARD parameters that control the precision of the components in $W$. Classical ProbCCA-AR uses the same set of ARD parameters for each observation set, while hybrid ProbCCA-AR uses different sets of ARD parameters for each observation set. This makes a small change to the graphical model, as shown in figure 7.16.

The amendment to the model is straightforward, starting with small changes to the priors for each row, $r$, of $W^{(\ast)}$ and the now separate ARD parameter variables:

$$p(w^{(\ast)}_{r;\ast}) = \mathcal{N}(w^{(\ast)}_{r;\ast} | 0, \text{diag}(\delta^{(\ast)})^{-1})$$

$$p(\delta^{(\ast)}_{q}) = \mathcal{G}(\delta^{(\ast)}_{q} | a_\delta, b_\delta)$$

The effect of these changes on the expressions for the posteriors is also small; the posterior
7. Robust variational CCA with autoregression

Figure 7.16. Subset of the graphical model for the robust variational ProbCCA-AR model with different ARD parameters over each \( \mathbf{W}^{(*)} \). The elements in red indicate the change from the previous model. All elements not included are identical to those shown in figure 7.1.

for each row of \( \mathbf{W}^{(*)} \) is the slightly modified Gaussian \( q(\mathbf{w}^{(*)}_{n} | \mathbf{T}) = \mathcal{N}(\mathbf{w}^{(*)}_{n} | \boldsymbol{\mu}^{(*)}_{n}, \Sigma^{(*)}_{w_n}) \)

\[
\Sigma^{(*)}_{w_n} = \text{diag}(\langle \delta^{(*)} \rangle) + \langle B^{(*)}_{r,r} \rangle \sum_{n=1}^{N} (\mathbf{x}_n \mathbf{x}_n^T) \quad (7.122)
\]

\[
\boldsymbol{\mu}^{(*)}_{w_n} = \Sigma^{(*)}_{w_n} \sum_{n=1}^{N} \left( \mathbf{x}_n (\mathbf{t}^{(*)}_n - \langle \tilde{\mathbf{t}}^{(*)} \rangle)^\text{T} (\mathbf{B}^{(*)}_{s,r}) - \frac{1}{2} \sum_{s=1}^{D} \langle B^{(*)}_{s,r} \rangle \langle \mathbf{x}_n \mathbf{x}_n^T \rangle \langle \mathbf{w}^{(*)}_{s}\rangle^\text{T} \right) \quad (7.123)
\]

(c.f. equations (7.71) and (7.72)) and the posteriors for the different ARD parameter variables are the Gamma distributions \( q(\delta^{(*)}_{q} | \mathbf{T}) = \mathcal{G}(\delta^{(*)}_{q} | \alpha_{\delta^{(*)}}, \beta_{\delta^{(*)}}) \)

\[
\alpha_{\delta^{(*)}} = a_{\delta} + \frac{D^{(*)}}{2} \quad (7.124)
\]

\[
\beta_{\delta^{(*)}} = b_{\delta} + \frac{1}{2} \sum_{r=1}^{D} \langle (\mathbf{w}_{r,q}^{(*)})^2 \rangle \quad (7.125)
\]

(c.f. equations (7.74) and (7.75)) with all the other posterior definitions remaining unchanged.

As for classical ProbCCA-AR, the expectations of the model's variables are initialised using robust variational PPCA-AR. However, in this case PPCA-AR is run for a few iterations against each observation set separately. The resulting latent variables are then stacked to initialise \( \mathbf{X} \) and \( \mathbf{W} \) is constructed as a block-diagonal matrix of the separate \( \mathbf{W} \) matrices from PPCA-AR. As before, the observation noise precision matrices, \( \mathbf{B}^{(*)} \), are are initialised to \( \lambda \mathbf{I} \), where \( \lambda \) is the scalar observation noise precision from PPCA-AR.

To test this new model a synthetic dataset was generated as before, but with different
Table 7.1. Table showing which of the 8 latent variables, and their corresponding components, contributes to each of the 4 observation sets.

<table>
<thead>
<tr>
<th>Observation set</th>
<th>$D^{(*)}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 7.17. (a) The actual $\mathbf{W}$ used to generate the synthetic data. (b) The $\mathbf{W}$ estimated by classical ProbCCA-AR. (c) The $\mathbf{W}$ estimated by hybrid ProbCCA-AR. (d) The $\mathbf{W}$ estimated by hybrid ProbCCA-AR, transformed in the direction of the actual $\mathbf{W}$ using the Procrustes transformation.

observation sets being constructed from different subsets of latent variables. It consisted of four observation sets, with dimensionalities 5, 6, 7 and 3 respectively, and 8 latent variables. Table 7.1 shows which of latent variables and corresponding components contributed to the signals in each observation set. The full $\mathbf{W}$ for this dataset is shown in figure 7.17a. The noise was Gaussian, with low levels of observation noise.

With the smallest observation dimension $D^{(*)}$ being 3, the classical ProbCCA-AR model will find a maximum of 2 latent variables. Training it against this dataset results in the estimated $\mathbf{W}$ shown in figure 7.17b. It appears to have identified the signal shared by the first three observation sets and a second which is shared between sets 1, 3 and 4. Performing a Procrustes transformation on the estimated $\mathbf{W}$ might make the link to the actual more obvious, but it is very difficult to tell which columns of the actual $\mathbf{W}$ to compare it against, especially since the second identified component seems to be a conglomeration of more than one of the originals.

The new, hybrid version of ProbCCA-AR was also trained against this dataset, using a dimensionality for the latent variables of $Q = \sum_{i=1}^{C} D^{(*)} - 1$ (where $C$ is a count of the number of observation sets; 4 in this case). The estimated $\mathbf{W}$ from the trained model is shown in figure 7.17c. Clearly it is identifying a number of latent variables that are shared between different subsets of observation sets and switching off a number of components.
7. Robust variational CCA with autoregression

Figure 7.18. The transformation matrix calculated by using the Procrustes method to transform the $W$ estimated by hybrid ProbCCA-AR in the direction of the actual $W$. It represents differences in sign and permutation of the components, but there is negligible rotation ambiguity.

Those components that are clearly switched on were transformed in the direction of the actual $W$ using the Procrustes transformation, resulting in a transformed version of the estimated $W$, shown in figure 7.17d and a transformation matrix, shown in figure 7.18. The new estimate shows a very good correspondence with the actual and the transformation matrix highlights the fact that while the sign and permutation of the estimated components are different from the actuals, there is no rotation ambiguity.

This new model has significantly higher resource requirements than the original due to the big increase in the dimensionality of the latent variables. It takes longer to run and uses more memory. Although promising, it will require faster algorithms or approximations for large datasets.

7.6. Conclusions

The previous chapter introduces robust variational PPCA-AR, which identifies latent variables shared between multivariate time-series observations in a single observation set. In this chapter the model is extended to create two different robust variational ProbCCA-AR models that identify latent variables shared between observation sets. Here a dataset is made up of two or more observation sets that contain the same number of observations with potentially differing dimensionalities. The first model (“classical” ProbCCA-AR) is based on classical CCA and identifies latent variables that are shared by all the observation sets; the second (“hybrid” ProbCCA-AR) extends this to enable it to identify latent variables that are shared between subsets of the observation sets. The latter requires a big increase in the dimensionality of the latent variables and is, therefore, significantly slower to train and uses significantly greater computing resources.

Classical ProbCCA-AR has been shown to be effective at identifying shared latent signals in synthetic datasets and, where the observation noise is small compared with the signal, generally identifies the correct model orders. Against five observation sets of evoked response EEG data it appears to have identified a single shared signal that seems to represent an evoked response. This is encouraging, but substantially more testing is required before
any significance can be ascribed to these results. The model needs to be trained against other observation sets for the same evoked response and against control observation sets.

Hybrid ProbCCA-AR has successfully identified latent variables in observation sets with complicated dependencies, but, again, substantially more testing is required to prove the model.

Both of these models assume that the observation noise is Gaussian distributed. Further work is required to bring them into line with PPCA-AR by making the assumptions Student-t.
8. Conclusions

In this thesis I have introduced two new latent variable models based on Probabilistic Principal Component Analysis (PPCA) and Probabilistic Canonical Correlation Analysis (ProbCCA) that capture both spatial and temporal information in multivariate time-series data. The first model, PPCA-AR, finds latent variables that support a single set of observations; the second, ProbCCA-AR, finds latent variables that support two or more sets of observations. In both cases independent autoregressive (AR) processes are applied to the latent variables to capture temporal relationships. These AR processes of order $P$ are converted to first order processes allowing both PPCA-AR and ProbCCA-AR to be represented as linear dynamical systems. This enables a forward-backward (Kalman filter/smoother) algorithm to be used to calculate the posterior distributions of the latent variables. These two models are fundamentally the same, except in the assumptions they make regarding the probability distribution of the observation noise: PPCA-AR, like PPCA, assumes that the noise precision matrix is diagonal, while ProbCCA-AR, like ProbCCA, assumes full precision matrices for each set of observations.

The concept behind the new models was introduced in chapter 3 with a maximum likelihood version of PPCA-AR with Gaussian noise assumptions. An Expectation Maximisation (EM) algorithm is used; taking advantage of the Gaussian noise assumption for the observation noise, the E-step uses the standard forward-backward algorithm to calculate the posterior distributions of the latent variables. It was shown that PPCA-AR may be viewed as a generalisation of PPCA.

While the maximum likelihood solution leads to posterior distributions for the latent variables, it provides only point estimates for each of the other parameters in the model. The next step was to move to a Bayesian scheme where the model parameters are viewed as random variables and the result is posterior distributions for all of them. The precisions or variances of these distributions provide a measure of the model’s certainty in its estimates, given the available data. The integrals required for exact Bayesian inference are intractable, so a variational method is used to provide approximate posterior distributions. A factorised scheme is employed which makes assumptions about independence (factorisation) between the approximate posteriors and hence avoids the long-winded process of minimising the Kullback-Leibler divergence through differentiation and the additional assumptions that process requires regarding the form of the approximate posterior distributions.

The factorised variational Bayesian method enables beneficial changes to be made to the model to generalise it further. Firstly the Gaussian assumptions for the noise are
replaced by the Student-t, which is an infinite mixture of Gaussians with a common mean and Gamma-distributed precision. It has been demonstrated that this makes the models more robust to outliers and better able to model the heavy-tailed distributions that are often found in real data. Secondly, automatic relevance determination (ARD) priors are introduced to provide automatic estimation of the model orders. This causes the PPCA/ProbCCA components and AR coefficients to be “switched off”, i.e. constrained to be close to zero, where there is no evidence for them in the data. These two elements are introduced by steps into PPCA-AR by first creating a new robust variational PPCA model in chapter 4 and new robust variational AR model in chapter 5 and then combining them into a new robust variational PPCA-AR model in chapter 6. The latter is shown to be a further generalisation of maximum likelihood PPCA-AR, with the maximum likelihood version representing the expectation of the equivalent Gaussian variational version. Both the PPCA and AR models were shown to be more robust to outliers than the standard Gaussian equivalents.

Changing the noise model from Gaussian to Student-t makes the exact forward-backward algorithm intractable. The new robust variational version introduced in chapter 6 is shown to be notionally equivalent to the exact version; both methods calculate the values of the latent variables at the current time as a weighted sum of contributions from the observation at the current time, the latent variables at the previous time and the latent variables at the following time. The weighting is according to the estimated precisions of the three contributions, with the more precise contributions being given a greater weighting. Although the results of this algorithm are approximate posterior distributions for the latent variables, it may be viewed as a generalisation of the exact method as the assumptions about the priors have been relaxed from Gaussian to an infinite mixture of Gaussians through the Student-t.

The PPCA model with Student-t latent variables demands that the sources are independent rather than merely decorrelated, so it essentially conforms to an Independent Component Analysis (ICA) model in which the sources are modelled by Student-t distributions. Note, however, that while the components in PPCA (and in PPCA-AR and ProbCCA-AR) are orthogonal, this is almost certainly not the case for a full ICA model, and the latter incorporates a wider range of source distributions. In particular the Student-t is incapable of representing platykurtic distributions and the PPCA Student-t model may not, therefore, be expected to be capable of separating such sources. For further discussion see Tipping and Lawrence’s [2005] “generalized component analysis” which also uses a Student-t source model. PPCA-AR with Gaussian latent variables is closely related to Pearlmutter and Parra’s [1997] “context-sensitive ICA” and also to the more generally coupled ICA models using generalised power distributions introduced by Penny et al. [2001], both of which use temporal information in the sources to enhance their separation.

PPCA-AR finds a shared subspace within a single observation set. ProbCCA-AR finds a subspace which is shared between two or more observation sets. Although this seems at first glance to be a big difference, in chapter 7 it was shown that in fact the fundamental difference between the two is that PPCA-AR assumes that the observation noise
8. Conclusions

is isotropic, that is the precision matrix is diagonal with the same value in each diagonal element, while ProbCCA-AR assumes that each observation set has a full precision matrix. As a slight simplification, ProbCCA-AR has been introduced with the assumption of Gaussian observation noise, but still with Student-t excitation noise for robustness and to model leptokurtic data.

Two different methods of applying ARD priors to the components of ProbCCA-AR lead to one model similar to traditional ProbCCA, which finds one subspace to support all the observation sets, and one model which may be thought of as a hybrid of this ProbCCA-AR and PPCA-AR, which find subspaces shared between all the observation sets and any subsets of them.

So in all, six new models have been introduced: maximum likelihood PPCA-AR, robust variational PPCA, AR and PPCA-AR and two different ProbCCA-AR models. In the first instance each model was trained against synthetic data which were constructed to fit the model’s assumptions and for which the actual parameter values were known. Unsurprisingly the magnitude of the observation noise relative to the signals was shown to be a major factor in the accuracy of the models’ estimations of their parameters. While there is no way of resolving the direction and sign ambiguities in the PPCA/ProbCCA components, other types of rotational ambiguity seem to be resolved where the models assume the noise to be Student-t distributed, especially where the data is heavy-tailed, though this is dependent on the level of observation noise. This resolution is attributed to the star-shaped multivariate Student-t distribution, where regions of higher probability spread along the axes representing each variable, allowing the estimation process to “lock” into a correct direction. This is in contrast to the Gaussian multivariate distribution which is spherical, providing no information regarding the rotation.

ARD has been proven to be effective at estimating the model orders. For PPCA, constraining the latent variables with the same ARD parameters as their associated components lead to generally sparser models that converged more quickly than Bishop’s [1999a] model which constrained only the components. However, this extra sparsity was at the expense of underestimating the actual model order. For AR, constraining the coefficients was effective at estimating the model order, but, unlike for the PPCA components, this did not generally result in the coefficients being entirely switched off (i.e. their values were not set to exactly zero). In fact, even where there is no temporal structure, the coefficients are still not entirely switched off. A simple test was described for determining whether a coefficient to be switched off. For the combined PPCA-AR and ProbCCA-AR models separate ARD parameters were used to constrain the PPCA/ProbCCA components and the AR coefficients. The latent variables were constrained only by the ARD priors over the AR coefficients, but it seems that the additional control exerted by PPCA/ProbCCA means that the measure for determining whether the coefficients are switched off is no longer viable (it always considers them all to be switched off). However, ARD was shown to constrain the magnitude of the coefficients to be close to zero outside the actual model order. Higher observation noise leads to an over-estimation of the number of components, as components are switched on to account for the noise. With the AR coefficients the
reverse is true: higher noise camouflages the temporal correlations leading to a reduction in the evidence for temporal model and hence causing the magnitude of the coefficients to be more constrained towards zero.

The new models have been demonstrated on three real datasets: the ONS data which contains 494 observations of 23-dimensional data; the 92 consecutive daily observations of the 113593-dimensional SeaWiFS satellite images supplied by NASA, where significant portions of the images are missing due to cloud cover; and the EEG data where each of the multiple observation sets contain 800 observations of 62-dimensional variables that are strongly leptokurtic.

The ONS data provides a complete and relatively low-dimensional set of data on which to demonstrate the ability of PPCA-AR and contains some obvious and significant outliers. PPCA-AR estimates the temporal model order to vary between 1 and about 12 for different components and both the observation and excitation noise contributions to be Gaussian (the expectations of the degrees of freedom variables are estimated to be large).

The NASA data represent a significant challenge to PPCA-AR due to its high dimensionality and the substantial proportion of data that is missing. Running on a Linux server with effectively 4 CPUs and 32Gb RAM, it took several days just to process 200 iterations of a subsection of the pictures. Even with these limitations, PPCA-AR does indicate useful information: despite the fact that the data is intuitively a time series, ARD is causing all the AR coefficients to be switched off, showing that there is not enough evidence in the observed data for a temporal model.

The robust variational models provide an interesting result for the EEG data. This was first highlighted by the AR model which showed a remarkably uniform estimation for the AR coefficients across all the electrodes of a single observation set, leading to a much more uniform set of power spectra. This was compared with the Gaussian model where there was no uniformity. This result was repeated with the robust variational version of PPCA-AR, which additionally indicated that there were two or three significant spatial components. ProbCCA-AR confirmed the result by estimating that two components were shared across multiple observation sets for the same evoked response test.

For these datasets it has been assumed that there is no \textit{a priori} knowledge regarding the prior probability distributions for the parameter variables and uninformative priors have been used. If more information was available, informative priors would increase the rate of convergence.

### 8.1. Future work

Looking first at practical considerations, a limitation of the new latent variable models (particularly the variational Bayesian models) is the amount of computer resource required to train them, both in CPU time and memory. There are two reasons for this: firstly full covariance matrices are stored in memory, which is particularly an issue for the latent
8. Conclusions

variables due to their dimensionality and the number of observations, and secondly, since we wish to use ARD to determine the model orders, the starting point for the estimated model orders must be the maximum possible.

While it is safe to say that increasing the potential model orders (i.e. the upper limits placed on ARD) increases the computing resource required, no formal work has been done to far to quantify the relationship, particularly between the potential model orders and the computation time. There is also an unexplored relationship between the initialisation of the model parameters and computation time, though informal testing has determined that random starting positions lead to significantly higher numbers of iterations required to achieve convergence.

For many of the variables full covariance matrices are not required; independence assumptions lead to the covariances being diagonal and so only the variances need to be stored. This would lead to a significant saving in memory as all the latent variable variances could be stored in a single matrix, rather than in $N$ separate covariance matrices as at present.

For highly dimensional data such as the NASA images it is not feasible to start the PPCA model order at its maximum value, so some manual estimation is still required. However, as ARD causes PPCA/ProbCCA components or AR coefficients to be switched off, the dimensionality of the variables could be reduced so that the process speeds up for later iterations. For the PPCA components this is relatively straightforward as they tend to be entirely switched off (i.e. the values of the elements in the component vector become exactly zero). For the AR coefficients the results are not so cut and dried and some method of thresholding the values is required.

Given the large number of extra parameters required to support the models with Student-t noise assumptions, these models require greater computing resources than those with Gaussian assumptions and converge more slowly. It has been shown that it is easy to switch between these assumptions within the algorithms, so the most efficient model for a given set of data may be selected experimentally. The results in chapters 4 and 7 seem to suggest that making Gaussian assumptions regarding the observation noise and Student-t assumptions regarding the excitation noise makes the models robust to outliers and still able to model leptokurtic data.

For ProbCCA-AR the issue of how to initialise the starting point of the algorithm remains. Some testing has found that a random start point is not productive, so for now using a few iterations of PPCA-AR to initialise the components and latent variables is the best option. If the data consists of just two observation sets then initialisation with CCA is a better option. Further work is required to measure the effects of this selection of initialisation procedure.

At present ProbCCA-AR assumes that the observation noise is Gaussian distributed, leading to the use of a Wishart conjugate prior for the precision matrix. For completeness the model should be extended to support Student-t observation noise, for which a different conjugate prior is required.
In terms of the real data to which these new models have been applied, further work is required for both the NASA and EEG datasets. Although PPCA-AR was not able to find evidence of temporal correlations within the NASA data due to the proportion of data that is missing, intuitively it seems very likely that there is an underlying temporal process at work. Other datasets are available which provide more complete (or at least differently-missing) observation sets for the same geographic area, have observations that were recorded over the same temporal period and which seem likely to have an input into the underlying process. For example, measurements of sea surface temperature and currents are recorded daily by buoys. Combining these parallel observation sets with the SeaWiFS images using one of the ProbCCA-AR models could benefit the estimation of the missing regions.

The application of ProbCCA-AR to the EEG data has produced some interesting results, but in order to prove that the results are significant the model needs to be trained against more than one group of observation sets for the same evoked response test and compared with results for control groups and groups for other evoked responses. As evoked response tests look for changes that occur in the brain signals in a response to a trigger, extending ProbCCA-AR to incorporate a changepoint model is required. It also seems likely that the evoked signal travels through the brain, potentially changing the magnitude of the contribution to particular electrodes over time. To capture this the model would have to support a mixing matrix which changes smoothly over the course of the observations.

8.2. Summary

This thesis has described new variational Bayesian latent variable models, PPCA-AR and ProbCCA-AR, where independent AR processes are used to model the latent variables in PPCA and ProbCCA respectively. They are made robust to outliers by modelling the noise using the heavy-tailed Student-t distribution and the model orders are estimated using automatic relevance determination. The thesis has also described a maximum likelihood, Gaussian version of PPCA-AR and new robust, variational PPCA and AR models.
Appendices
A. Probability Distributions

This appendix summarises the main properties of the standard probability distributions referred to in this thesis and some of their key statistics: $E[x]$ is the expected value of $x$, $\text{var}[x]$ its variance and $\text{cov}[x]$ its covariance.

A.1. Gamma

With $x$, $a$ and $b$ all greater than zero:

$$G(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \quad (A.1)$$

$$E[x] = \frac{a}{b} \quad (A.2)$$

$$\text{var}[x] = \frac{a}{b^2} \quad (A.3)$$

$$E[\log(x)] = \psi(a) - \log(b) \quad (A.4)$$

The Gamma is the conjugate distribution for a univariate Gaussian precision.

A.2. Gaussian

For a univariate variable, $x$, with mean $\mu$ and precision $\lambda$:

$$N(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp \left( -\frac{\lambda}{2} (x - \mu)^2 \right) \quad (A.5)$$

$$E[x] = \mu \quad (A.6)$$

$$\text{var}[x] = \lambda^{-1} \quad (A.7)$$

$$E[x^2] = \lambda^{-1} + \mu^2 \quad (A.8)$$

In the multivariate case, with $\mathbf{x}$ as a $D$-dimensional vector:

$$N(\mathbf{x} \mid \mu, \Lambda^{-1}) = \sqrt{\frac{|\Lambda|}{2\pi}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Lambda (\mathbf{x} - \mu) \right) \quad (A.9)$$

$$E[\mathbf{x}] = \mu \quad (A.10)$$

$$\text{cov}[\mathbf{x}] = \Lambda^{-1} \quad (A.11)$$

$$E[\mathbf{xx}^T] = \Lambda^{-1} + \mu \mu^T \quad (A.12)$$
\( \Lambda \) is a \( D \times D \) symmetric, positive definite matrix.

### A.3. Student-t

For a univariate variable, \( x \), with mean \( \mu \), precision \( \lambda \) and degrees of freedom \( d \) (where \( d > 0 \)):

\[
S(x \mid \mu, \lambda, d) = \frac{\Gamma \left( \frac{d+1}{2} \right)}{\Gamma \left( \frac{d}{2} \right)} \left( \frac{\lambda}{\pi d} \right)^{\frac{1}{2}} \left( 1 + \frac{\lambda(x - \mu)^2}{d} \right)^{-\frac{d+1}{2}}
\]

\[
= \int_0^\infty \mathcal{N}(x \mid \mu, (\lambda z)^{-1}) \mathcal{G}(z \mid \frac{d}{2}, \frac{d}{2}) \, dz \tag{A.14}
\]

\[
E[x] = \mu \quad \text{for } d > 1 \tag{A.15}
\]

\[
\text{var}[x] = \frac{d}{\lambda(d - 2)} \quad \text{for } d > 2 \tag{A.16}
\]

### A.4. Wishart

With degrees of freedom \( \nu \) greater than zero and scale matrix \( W \):

\[
\mathcal{W}(W \mid \Delta, \nu) = c |W|^{-\frac{\nu}{2}} \exp \left( -\frac{1}{2} \text{trace}(\Delta^{-1}W) \right)
\]

\[
\text{where } c = |\Delta|^{-\frac{\nu}{2}} \left( 2^\frac{\nu D}{2} \pi^{\frac{D(D-1)}{4}} \prod_{i=1}^D \Gamma \left( \frac{\nu + 1 - i}{2} \right) \right)^{-1}
\]

\[
E[|W|] = \nu W \tag{A.18}
\]

\[
E[\log |W|] = \sum_{i=1}^D \psi \left( \frac{\nu + 1 - i}{2} \right) + D \log(2) + \log |\Delta| \tag{A.19}
\]

The Wishart is the conjugate distribution for a multivariate Gaussian precision matrix.
B. Maximum Likelihood estimation for Gaussian and Student-t distribution parameters

Given a set of $N$ univariate samples, $x$, from an unknown distribution, we may use maximum likelihood to estimate the parameters of a Gaussian or Student-t distribution to fit the data.

B.1. Gaussian maximum likelihood estimation

To fit a Gaussian of the form $\mathcal{N}(x_n | \mu, \sigma^2)$, the log likelihood to be maximised is defined as

$$\log(p(x | \mu, \sigma^2)) = -\frac{1}{2} \sum_{n=1}^{N} \left( \log(2\pi) + \log(\sigma^2) + \frac{1}{2\sigma^2} (x_n - \mu)^2 \right)$$  \hspace{1cm} (B.1)

Differentiating with respect to $\mu$ gives:

$$\frac{\partial}{\partial \mu} \log(p(x | \mu, \sigma^2)) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (-2x_n - 2\mu)$$  \hspace{1cm} (B.2)

Setting this to zero gives the maximum likelihood estimate for $\mu$:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$  \hspace{1cm} (B.3)

Differentiating B.1 with respect to $\sigma^2$ gives:

$$\frac{\partial}{\partial \sigma^2} \log(p(x | \mu, \sigma^2)) = -\frac{1}{2\sigma^2} \left( N - \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right)$$  \hspace{1cm} (B.4)

Setting this to zero gives the maximum likelihood estimate for $\sigma^2$:

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2$$  \hspace{1cm} (B.5)
B.2. Student-t maximum likelihood estimation

Using the form shown in (2.47), we may introduce a new latent variable, \( z \):

\[
p(x_n | \mu, \lambda, z_n) = \mathcal{N}(x_n | \mu, (\lambda z_n)^{-1})
\]

\[
p(z_n | d) = \mathcal{G}(z_n | \frac{d}{2}, \frac{d}{2})
\]

To fit a Student-t of the form \( S(x_n | \mu, \lambda, d) \), the complete data log likelihood to be maximised is defined as

\[
\log(p(x, z | \mu, \lambda, d)) = \sum_{n=1}^{N} \left( -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\lambda) + \frac{1}{2} \log(z_n) - \frac{\lambda z_n^2}{2} (x_n - \mu)^2 \
- \log(\Gamma(\frac{d}{2})) + \frac{d}{2} \log(\frac{d}{2}) + (\frac{d}{2} - 1) \log(z_n) - \frac{d}{2} z_n \right)
\]

The values of the parameters may be determined using an EM algorithm. The E-step calculates the probability distribution of the latent variables in \( z \). Using logarithms to make the calculations easier:

\[
\log(p(z_n | x_n, \mu, \lambda, d)) \propto \log(p(x_n | \mu, \lambda, z_n) p(z_n | d))
\]

\[
= \log(\mathcal{N}(x_n | \mu, (\lambda z_n)^{-1})) + \log(\mathcal{G}(z_n | \frac{d}{2}, \frac{d}{2}))
\]

\[
= \frac{1}{2} \log(z_n) - \frac{\lambda}{2} (x_n - \mu)^2 z_n + \left( \frac{d}{2} - 1 \right) \log(z_n) - \frac{d}{2} z_n + \text{const}
\]

\[
= \left( \frac{d}{2} + \frac{1}{2} - 1 \right) \log(z_n) - \left( \frac{d}{2} + \frac{\lambda}{2} (x_n - \mu)^2 \right) z_n + \text{const}
\]

which may be recognised as the Gamma distribution \( \mathcal{G}(z_n | b) \), where \( a = \frac{d+1}{2} \) and \( b = \frac{d+\lambda(x_n-\mu)^2}{2} \). For the M-step two expectations are required: \( \langle z_n \rangle = \frac{a}{b} \) and \( \langle \log(z_n) \rangle = \psi(a) - \log(b) \).

For the M-step \( z \) is replaced in (B.8) by its expectations:

\[
\frac{\partial}{\partial \mu} \log(p(x, z | \mu, \lambda)) = \frac{1}{2} \sum_{n=1}^{N} \frac{\partial}{\partial \mu} \left( \langle z_n \rangle \lambda (x_n^2 - 2x_n \mu + \mu^2) \right)
\]

\[
= \lambda \sum_{n=1}^{N} (-x_n \langle z_n \rangle + \mu \langle z_n \rangle)
\]

Setting this to zero gives the maximum likelihood estimate for \( \mu \) as

\[
\mu_{\text{ML}} = \frac{\sum_{n=1}^{N} x_n \langle z_n \rangle}{\sum_{n=1}^{N} \langle z_n \rangle}
\]
B. Maximum Likelihood estimation for Gaussian and Student-t distribution parameters

For $\lambda$ we get

$$\frac{\partial}{\partial \lambda} \log(p(x, z | \mu, \lambda)) = \frac{N}{2} \frac{1}{\lambda} - \frac{1}{2} \sum_{n=1}^{N} (z_n)(x_n - \mu)^2$$

(B.16)

Setting this to zero gives the maximum likelihood estimate for $\lambda$ as

$$\lambda_{ML} = \frac{N}{\sum_{n=1}^{N} (z_n)(x_n - \mu_{ML})^2}$$

(B.17)

Finally, for $d$ we get

$$\frac{\partial}{\partial d} \log(p(x, z | \mu, \lambda, d)) = \frac{\partial}{\partial d} \left[ N \left( - \log(\Gamma(\frac{d}{2})) + \frac{d}{2} \log(\frac{d}{2}) \right) + \sum_{n=1}^{N} \left( \frac{d}{2} \langle \log(z_n) \rangle - \frac{d}{2} \langle z_n \rangle \right) \right]$$

=B.18

$$= \frac{N}{2} \left( -2\psi(\frac{d}{2}) + \log(d) + 1 - \log(2) \right) + \frac{1}{2} \sum_{n=1}^{N} \left( \langle \log(z_n) \rangle - \langle z_n \rangle \right)$$

(B.19)

The value for $d$ that makes this zero must be determined numerically.
Bibliography


Beal, M. and Ghahramani, Z. (2002). The variational Bayesian EM algorithm for in-


Lorenz, E. (1956). Empirical orthogonal functions and statistical weather prediction. Scientific report no. 1, statistical forecasting project, MIT, Department of Meteorology.


tions of state calculations by fast computing machines. Journal of Chemical Physics,

report 514, MIT.


Naylor, J. and Smith, A. (1982). Applications of a method for the efficient computation of

Toronto, Canada.

algorithm that justifies incremental, sparse, and other variants, pages 355–368. MIT
Press, Cambridge MA.

Neuhaus, J., C., and Wrigley (1954). The quartimax method: an analytical approach to

Newcomb, S. (1886). A generalized theory of the combination of observations so as to


tsdintro.asp. Data extracted on 9th April 2009.


chapter Maximum likelihood blind source separation: a context-sensitive generalization

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space.

University Press.

Bibliography


