Article:
Baldacchino, T., Worden, K. and Rowson, J. (2017) Robust nonlinear system identification: Bayesian mixture of experts using the t-distribution. Mechanical Systems and Signal Processing, 85. pp. 977-992. ISSN: 0888-3270

https://doi.org/10.1016/j.ymssp.2016.08.045

Robust Nonlinear System Identification: Bayesian Mixture of Experts Using the *t*-Distribution

Tara Baldacchino[a,*], Keith Worden[a], Jennifer Rowson[a]

[a]*Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, UK.*

Abstract

A novel variational Bayesian mixture of experts model for robust regression of bifurcating and piece-wise continuous processes is introduced. The mixture of experts model is a powerful model which probabilistically splits the input space allowing different models to operate in the separate regions. However, current methods have no fail-safe against outliers. In this paper, a robust mixture of experts model is proposed which consists of Student-*t* mixture models at the gates and Student-*t* distributed experts, trained via Bayesian inference. The Student-*t* distribution has heavier tails than the Gaussian distribution, and so it is more robust to outliers, noise and non-normality in the data. Using both simulated data and real data obtained from the Z24 bridge this robust mixture of experts performs better than its Gaussian counterpart when outliers are present. In particular, it provides robustness to outliers in two forms: unbiased parameter regression models, and robustness to overfitting/complex models.

Keywords: Outliers, robust estimation, Student-*t* distribution, variational Bayes, mixture of experts, bifurcating mechanical structures.

1. Introduction

When violations of modelling assumptions by the underlying data-generating process occur, robust system identification methods need to be considered in order to ensure unbiased models. A simple example is the violation of normality of residuals which renders ordinary least squares system identification inaccurate. Robust modelling methodologies are essential when the data contain outliers, that is, data points which are significantly different from the rest of the data. A formal definition of an outlier was given by Hawkins: '*An outlier is an observation which deviates so much from the other observations as to arouse suspicions*

*Corresponding Author

Email addresses: t.baldacchino@sheffield.ac.uk (Tara Baldacchino), k.worden@sheffield.ac.uk (Keith Worden), j.rowson@sheffield.ac.uk (Jennifer Rowson)

that it was generated by a different mechanism' [1]. Here, the term outlier is used to refer to a data point which is either an abnormality or the result of noise. One way of dealing with outliers is via outlier detection where outlier points are identified as being different from the underlying process, and there has been much debate in the modelling community regarding the removal of outlier data points, see [2, 3]. In this paper, the authors choose a different technique referred to as outlier accommodation achieved by using robust methods which protects the modelling process from being distorted by the presence of outliers. However, in any situation, not accounting for outliers could have severe consequences in parametric regression modelling, resulting in biased parameter estimates and an artificially inflated variance estimate (thereby masking the outliers). This may have the effect of providing incorrect results with misleading information. Hence, modelling techniques which are robust to outliers are essential.

A major drawback of many parametric regression modelling techniques is the assumption of an underlying Gaussian distribution for the innovations/residuals, and thus they are highly influenced by outliers. A commonly-used technique to ensure robustness is the use of a heavy-tailed distribution, such as the Student-t distribution. Such a distribution assumes that outliers are much more probable. This distribution has been used for robust estimation with outliers or atypical observations for many decades, see for example [4, 5]. However, it is still a topic of ongoing research and has recently been employed for robust estimation in various fields: Gaussian processes [6], time series analysis using variational Bayes [7] and reversible jump Markov chain Monte Carlo [8], mixture models [9], mixture of regression models [10], mixture of autoregressive series [11] and mixture of experts using the expectation conditional maximisation [12]. In this paper the Student-t distribution is incorporated into a mixture of experts (MoE) Bayesian modelling framework so as to provide a novel approach to robustness to outlier data in piece-wise continuous data and bifurcating processes. For an overview of robust Bayesian analysis, readers are referred to [13].

The MoE model, introduced in [14], has successfully been applied to a wide range of applications [15, 16, 17, 18, 19]. The MoE model consists of gates which probabilistically divide the input space of a system while the experts specialise on a certain part of the input space. The model parameters of a MoE model are usually estimated in one of two ways: via maximum likelihood (ML) techniques utilising the expectation-maximisation (EM) algorithm (see [20] among others), or via Bayesian inference. Within a Bayesian framework, parameter estimation is performed using either Markov chain Monte Carlo (MCMC) [21] and more recently employing variational Bayesian (VB) methods expressed in an EM-like algorithm, giving rise to the variational Bayesian EM (VBEM) algorithm, see for example [15]. The VBEM algorithm provides a deterministic technique for estimating posterior distributions, rendering VBEM less computationally demanding than MCMC methods. The main advantages of a Bayesian approach over ML is that complex models are naturally penalised, hence avoiding overfitting. It also provides a natural metric for determining the number of experts.

The gate and expert functions can take on numerous forms, and a recent

review of mixture of experts can be found in [22]. Models for the gate include: Gaussian mixture model [23], neural networks [24] and Dirichlet process [25]. Commonly used regression models for the experts include: Gaussian [20] and Gaussian process [26]. The gate and experts are decoupled during training, hence attaining a modular structure. This modularity allows the possibility of any gate model and expert model to be used together. However, despite much discussion in the literature with regards to the robustness of mixture models to outliers, see for example [27, 9, 10], there is a distinct gap when it comes to MoE models. To the authors' knowledge, the only work dealing with robust learning for MoE was tackled in [28] who applied a generalisation of the ML estimator using gradient ascent techniques. However, their method suffers from the usual drawback pertaining to ML techniques: as they increased the number of experts, the performance measure of the algorithm increased.

In this work a novel robust Bayesian MoE model is proposed by using a Student-t mixture model (SMM) for the gate, and a Student-t model for the innovations in the expert functions. This proposed model is trained via the VBEM algorithm, giving rise to closed-form analytical update equations for the model parameters. A Bayesian approach is considered here since it exhibits similar computational complexity as the ML version (see [29]), as well having several advantages over ML, as discussed previously. Inherent to the Bayesian training is the inclusion of uncertainty, via probability distributions, and hence credible bounds on predictions are obtained naturally. The novel robust Bayesian MoE model presented in this paper provides a fast and effective method for modelling bifurcating/piece-wise systems in the presence of outliers, as will be discussed in Section 4.

The layout of the paper is as follows. The robust MoE model is introduced in Section 2. Section 3 provides details of the VBEM algorithm, along with the necessary variational update equations for the model. The results of the algorithm are presented in Section 4: applied firstly to a simulated bifurcating Duffing oscillator, and secondly to the Z24 bridge data which exhibits a bilinear relationship between the modal frequencies and deck temperature.

2. Robust Mixture of Experts

In Section 2.1, the Student-t distribution is introduced as an infinite mixture of scaled Gaussians. The particular forms of the gates and experts of the MoE model used in this paper are given in Section 2.2.

2.1. Student-t distribution

The multivariate Student-t distribution for a variable $\boldsymbol{x} = [x^1, \ldots, x^{d^x}] \in \mathcal{R}^{1\times d^x}$ is given by

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}, \eta) = \frac{\Gamma(\eta/2 + d^x/2)|\Lambda|^{1/2}}{\Gamma(\eta/2)(\eta\pi)^{d^x/2}} \Big(1 + \frac{\Delta^2}{\eta}\Big)^{-(\eta+d^x)/2}, \qquad (1)$$

where the $\Gamma(\cdot)$ is the gamma function and

$$\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{2}$$

is the squared-Mahalanobis distance from $\boldsymbol{x}$ to the mean $\boldsymbol{\mu}$. $\boldsymbol{\Lambda}^{-1}$ is the covariance matrix and $\eta > 0$ is the number of degrees-of-freedom. As $\eta \to \infty$ the Student t-distribution reduces to a Gaussian distribution. At finite values of η the Student distribution has heavier tails than the corresponding Gaussian for the same $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}^{-1}$, and so the Student t-distribution represents a generalisation of the Gaussian distribution, Figure 1.

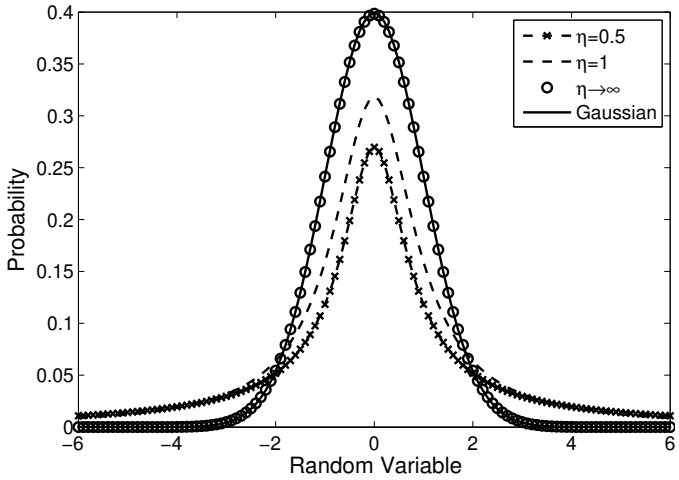

Figure 1: Univariate Student-t distribution with μ=0 and Λ=1 for different values of η. When $\eta \to \infty$ ($\circ$), the Student-t distribution corresponds to a Gaussian distribution (solid line) and the two plots coincide.

Unfortunately, no closed-form solution exists when maximising the likelihood using a Student distribution. Thus an alternative representation of the Student distribution is required and this is given as an infinite mixture of scaled Gaussians, written as

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}, \eta) = \int_0^\infty \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, (u\boldsymbol{\Lambda})^{-1}) \mathcal{G}a(u|\eta/2, \eta/2)\ du \tag{3}$$

where $\mathcal{N}(\cdot)$ is the Gaussian distribution, and $\mathcal{G}a(\cdot)$ is the Gamma distribution. The representation in (3) can be viewed as introducing an implicit continuous latent scale variable u, on which a Gamma prior is imposed, for each observation of $\boldsymbol{x}$. Equation (1) is easily obtained from (3) by marginalising over this continuous latent scale variable, since the Gamma distribution is conjugate to the Gaussian distribution. This outlook lends itself naturally to finding a maximum likelihood solution within an EM framework, as discussed in [30].

In this paper, a novel mixture of experts model is introduced by using the Student-t distribution, given in (3), for both the gate and expert functions. The gating function consists of a Student mixture model (SMM), while in the expert function the innovations take on the form of a Student distribution. This set up provides robustness to atypical data points in the dataset, both in the form of outliers in the output and non-Gaussian distributed inputs. The form of the MoE model used here is similar to the MoE model with Gaussian mixture model (GMM) at the gates, as given in [23, 16, 19]. However, as discussed in [27, 9], the GMM is susceptible to non-Gaussianity in the inputs and hence tends to select a more complex model (one which has more components) in order to capture the tails of the distribution. Naturally, this problem is inherited by the MoE when the gates take the form of GMM. Additionally, outliers in the observed variable will introduce bias in the regression, so using Student innovations helps to overcome poor regression [10].

2.2. Mixture of Experts Model

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathcal{R}^{N \times d^x}$ be a d^x dimensional input to the system of interest, consisting of N data points, such that $\boldsymbol{x}_n = [x_n^1, \ldots, x_n^{d^x}]$. Let the corresponding vector of scalar outputs be $\boldsymbol{y} = [y_1, \ldots, y_N]^\top \in \mathcal{R}^{N \times 1}$; then, a regression MoE model with M experts is given by

$$y_n = \sum_{i=1}^{M} g_i(\boldsymbol{x}_n, \theta_i^g) f_i(\boldsymbol{x}_n, \boldsymbol{w}_i) \ , \tag{4}$$

where $g_i(\cdot)$ and $f_i(\cdot)$ are the i^{th} gating and expert functions respectively. The expert is restricted to be a vector function given by $f_i(\boldsymbol{x}_n, \boldsymbol{w}_i) = \boldsymbol{w}_i^\top [\boldsymbol{x}_n \ 1]$, where $\boldsymbol{w}_i$ are the expert weights which include a bias term given by the 1 appended to the input matrix. The gating function used here is a normalised Student-t function, such that

$$g_i(\boldsymbol{x}_n, \pi_i, \theta_i^g, \eta_i) = \frac{\pi_i \mathcal{T}(\boldsymbol{x}_n | \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{-1}, \eta_i)}{\sum_{l=1}^{M} \pi_l \mathcal{T}(\boldsymbol{x}_n | \boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l^{-1}, \eta_l)} \ . \tag{5}$$

where $\theta_i^g = \{\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i\}$. The mixing coefficients are given by $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^M$, satisfying $\pi_i \geq 0$ and $\sum_{i=1}^M \pi_i = 1$. Equation (5) can also be expressed using (3), such that a latent scale variable u_{ni} is associated with each data point $\boldsymbol{x}_n$ and each component i. The likelihood for the MoE is represented as

$$p(y_n | \boldsymbol{x}_n, \boldsymbol{\Theta}) = \sum_{i=1}^{M} p(i | \boldsymbol{x}_n, \pi_i, \theta_i^g, \eta_i) p(y_n | \boldsymbol{x}_n, \theta_i^e) \ . \tag{6}$$

$p(i | \boldsymbol{x}_n, \pi_i, \theta_i^g, \eta_i) = g_i(\cdot)$ is the posterior conditional probability that $\boldsymbol{x}_n$ is assigned to the segment corresponding to the i^{th} expert. The probability distribution, $p(y_n | \boldsymbol{x}_n, \theta_i^e)$, of the i^{th} expert is taken to be a Student-t distribution

having mean f_i given by

$$\begin{aligned} p(y_n|\boldsymbol{x}_n, \theta_i^e, \kappa_i) &= \mathcal{T}(y_n|\boldsymbol{w}_i^\top[\boldsymbol{x}_n\ 1], \tau_i^{-1}, \kappa_i) \\ &= \int \mathcal{N}(y_n|\boldsymbol{w}_i^\top[\boldsymbol{x}_n\ 1], (s_{ni}\tau_i)^{-1})\mathcal{G}a(s_{ni}|\kappa_i/2, \kappa_i/2)\ ds_{ni} \end{aligned} \tag{7}$$

The parameter vector for the experts is $\boldsymbol{\theta}^e = [\boldsymbol{W}, \boldsymbol{\tau}]$, where $\boldsymbol{W} = \{\boldsymbol{w}_i\}_{i=1}^M$ is the weight vector, and $\boldsymbol{\tau}^{-1} = \{\tau_i^{-1}\}_{i=1}^M$ is the variance. The set of unknown model parameters in (6) is given by $\boldsymbol{\Theta} = [\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\theta}^g, \boldsymbol{\theta}^e]$. The alternative model developed by [23] is adopted here in order to obtain closed form solutions for the parameter updates, and the joint density is given by

$$p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\Theta}) = \prod_{n=1}^{N}\sum_{i=1}^{M} \underbrace{\pi_i\mathcal{T}(\boldsymbol{x}_n|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{-1}, \eta_i)}_{\tilde{g}_i} \mathcal{T}(y_n|\boldsymbol{w}_i^\top[\boldsymbol{x}_n\ 1], \tau_i^{-1}, \kappa_i)\ , \tag{8}$$

where the gating network $\tilde{g}_i$ is a Student Mixture Model (SMM). Maximum likelihood estimation of SMM within the EM framework was introduced by [31], while a Bayesian approach using variational Bayes was tackled by [27, 9].

Discrete latent indicator variables $\boldsymbol{Z} = \{z_{ni}\}_{i=1,n=1}^{M,N}$ are introduced such that if $(\boldsymbol{x}_n, y_n)$ was generated from the i^{th} model then $z_{ni} = 1$, otherwise it is 0. Thus the complete-data likelihood for (8) can be written as

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}|\boldsymbol{\Theta}) = \prod_{n=1}^{N}\prod_{i=1}^{M} \Big(&\pi_i\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_i, (u_{ni}\boldsymbol{\Lambda}_i)^{-1})\mathcal{G}a(u_{ni}|\eta_i/2, \eta_i/2) \times \\ &\mathcal{N}(y_n|\boldsymbol{w}_i^\top[\boldsymbol{x}_n\ 1], (s_{ni}\tau_i)^{-1})\mathcal{G}a(s_{ni}|\kappa_i/2, \kappa_i/2)\Big)^{z_{ni}}\ , \end{aligned} \tag{9}$$

where $\boldsymbol{U}$, $\boldsymbol{S}$ are $N \times M$ matrices, with $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N]^\top = \{u_{ni}\}_{i=1,n=1}^{M,N}$, where $\boldsymbol{u}_n = [u_{n1}, \ldots, u_{nM}]$, and similarly for $\boldsymbol{S}$. Marginalising (9) over all the latent variables, $\boldsymbol{\Theta}^l = [\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}]$, results in (8). Defining the likelihood in this way encourages soft competition such that only one expert is dominant in a certain region of the input space [14]. Following on from previous work in the literature, by maximising the marginal likelihood of the data, $p(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\kappa})$, updates for the mixing coefficients $\boldsymbol{\pi}$ [32, 19] and degree-of-freedom parameters $\boldsymbol{\eta}$ and $\boldsymbol{\kappa}$ [27, 9] can be obtained. These parameters are denoted as $\boldsymbol{\Theta}^{ML} = [\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\kappa}]$, where the superscript ML refers to the maximum-likelihood updating of the parameters. The rest of the parameters, $\boldsymbol{\theta}^g$ and $\boldsymbol{\theta}^e$, and the latent variables are treated as random variables, and hence Bayesian inference is used to find approximate posterior distributions for these variables. An attractive advantage of using VBEM with the model described in this section is that $[\boldsymbol{\Theta}^{VB}, \boldsymbol{\Theta}^l] = [\{\boldsymbol{\theta}^g, \boldsymbol{\theta}^e\}, \boldsymbol{\Theta}^l]$ can be derived analytically and thus have closed-form solutions (where the superscript VB refers to the variational Bayes updating of the parameters).

3. Variational Bayesian Framework

Prior distributions for the random variables of the model's parameters, $\boldsymbol{\Theta}^{VB}$, are specified in Section 3.1. Details of the VBEM algorithm for the model described in this paper are given in Section 3.2. In Section 3.3 the optimised variational distribution update equations for all the random variables are given, while in Section 3.5 updates for $\boldsymbol{\Theta}^{ML}$ are presented.

3.1. Priors

Priors from the exponential family (example, Gaussian, Gamma and Wishart, where the Wishart is a multivariate version of the Gamma distribution) are considered in this work. These prior distributions are conjugate priors for the likelihood function given in (9), thus ensuring that the posterior distributions are of the same form as the prior. In this paper, the authors use the Gaussian distribution for parameters, the Gamma distribution for variance (since the variance is strictly positive) and the Wishart distribution for covariance (since covariance is the multivariate version of variance). More information regarding conjugate priors can be found in [33].

Since both the gating mean, $\boldsymbol{\mu}$, and precision, $\boldsymbol{\Lambda}$, are assumed unknown, the conjugate prior assigned to these parameters for each Gaussian component is the Gaussian-Wishart prior,

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) =& p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ =& \prod_{i=1}^{M} \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_i)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_i|\boldsymbol{B}_0, \nu_0) \ . \end{aligned} \tag{10}$$

$\boldsymbol{B}_0$ is a $d^x \times d^x$ symmetric, positive definite matrix, and $\nu_0 > d^x - 1$ is the number of degrees-of-freedom of the Wishart distribution. The prior distribution of the joint weight and precision parameters for the expert function is given by a Gaussian-Gamma distribution,

$$p(\boldsymbol{W}, \boldsymbol{\tau}|\boldsymbol{a}) = \prod_{i=1}^{M} \mathcal{N}(\boldsymbol{w}_i|0, (\tau_i A_i)^{-1})\mathcal{G}a(\tau_i|\rho_0, \lambda_0) \ , \tag{11}$$

where $\boldsymbol{a} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M)$, and $\boldsymbol{a}_i = \{a_{i,j}\}_{j=1}^{d^x+1}$ are the parameters associated with automatic relevance determination (ARD) [34, 35]. If $a_{i,j}^{-1} = 0$, then the corresponding input $\boldsymbol{x}^j$ is irrelevant to form the distribution of the output y_i of the i^{th} expert since the corresponding weight $w_{i,j}$ will be very small. The matrix A_i is formed from $\boldsymbol{a}$ such that $A_i = \text{diag}(a_{i,1}, \ldots, a_{i,d^x+1})$. $a_{i,j}$ is the hyperparameter on which the expert weight $w_{i,j}$ depends on and is given the following hyperprior distribution

$$p(a_{i,j}) = \mathcal{G}a(a_{i,j}|c_0, d_0) \ . \tag{12}$$

The variables $\beta_0, \boldsymbol{m}_0, \boldsymbol{B}_0, \nu_0, \rho_0, \lambda_0, c_0, d_0$ are referred to as hyperparameters (parameters of the prior) and they are initialised at the start to provide broad

priors. The joint distribution of all the random variables conditioned on $\boldsymbol{\Theta}^{ML}$ can be expressed hierarchically as,

$$p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}^{l}, \boldsymbol{\Theta}^{VB}, \boldsymbol{a}|\boldsymbol{\Theta}^{ML}) = p(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\Theta}^{l}, \boldsymbol{\Theta}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{U}|\boldsymbol{Z}, \boldsymbol{\eta}) p(\boldsymbol{S}|\boldsymbol{Z}, \boldsymbol{\kappa}) \\ p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{W}, \boldsymbol{\tau}|\boldsymbol{a}) p(\boldsymbol{a}) \ , \tag{13}$$

which is shown in Figure (2).

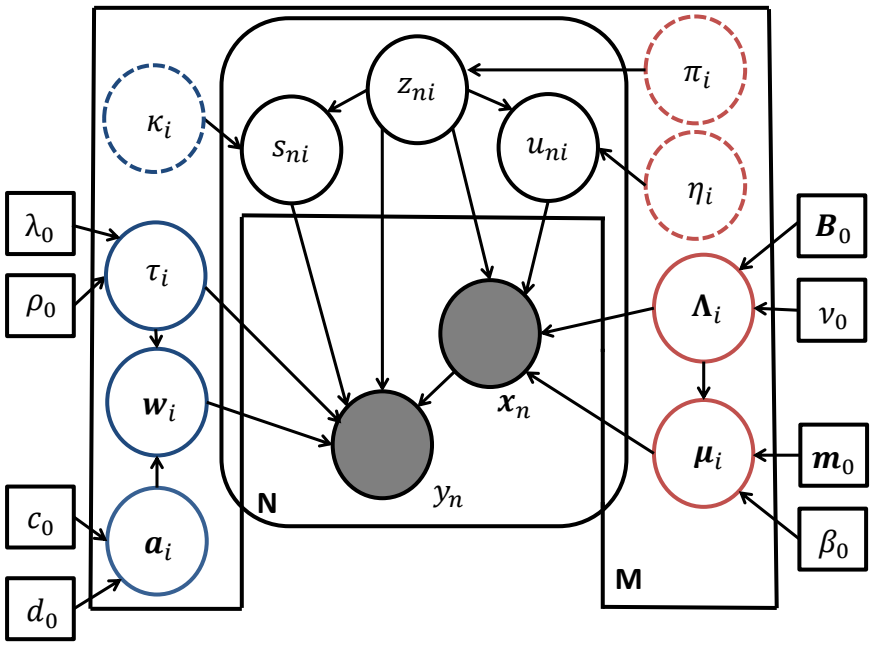

Figure 2: Graphical model for Bayesian MoE model with SMM gates and Student-t experts. The rounded plate denotes N i.i.d observations of $\boldsymbol{y}$ and $\boldsymbol{X}$ (shaded circles). The M-plate represents the M mixture components incorporating both the gate and expert parameters. The latent variables, $\boldsymbol{Z}$, $\boldsymbol{U}$ and $\boldsymbol{S}$, belong to both plates. Broken circles denote adjustable parameters, while square boxes refer to known hyperparameters. Unobserved random variables are indicated by complete circles (red corresponds to gate parameters, blue corresponds to expert parameters). The arrows represent conditional dependencies between the variables.

3.2. Variational Bayes Expectation Maximisation

Since exact inference of the Bayesian robust MoE model is not possible, an approximate Bayesian framework is needed. The choice of conjugate-exponential distributions, along with a latent variable model is elegantly accommodated by the VBEM framework. Expressing the set of all unobserved stochastic variables

by $\boldsymbol{\vartheta}$, the log-marginal likelihood (denominator in Bayes theorem) is given by [36]

$$\ln p(\boldsymbol{y}) = \mathcal{F}(q(\boldsymbol{\vartheta})) + \mathrm{KL}[q(\boldsymbol{\vartheta}) \parallel p(\boldsymbol{\vartheta}|\boldsymbol{y})] \ , \tag{14}$$

where $\mathrm{KL}[q(\boldsymbol{\vartheta}) \parallel p(\boldsymbol{\vartheta}|\boldsymbol{y})]$ is the Kullback-Leibler (KL) divergence between the variational posterior distribution $q(\boldsymbol{\vartheta})$ and the true posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$. Since the KL divergence is always positive, then $\mathcal{F}(q(\boldsymbol{\vartheta}))$ is a lower bound of the log-marginal likelihood. The main objective of variational Bayes is to maximise $\mathcal{F}(q(\boldsymbol{\vartheta}))$ with respect to $\boldsymbol{\vartheta}$ in order to get a tight bound (hence minimising the KL divergence).

For the model specified in (9), the random variables are $\boldsymbol{\vartheta} = [\boldsymbol{\Theta}^l, \{\boldsymbol{\Theta}^{VB}, \boldsymbol{a}\}]$, and thus constitutes both latent variables and model parameter variables. A constraint on the variational distribution is enforced; a factorised variational distribution is used in order to make evaluation of the lower bound tractable, that is, $q(\boldsymbol{\vartheta}) = q(\boldsymbol{\Theta}^l)q(\boldsymbol{\Theta}^{VB}, \boldsymbol{a})$. The update equations, for the E- and M-steps, are obtained by performing functional differentiation of $\mathcal{F}(q(\boldsymbol{\vartheta}))$ with respect to $q(\boldsymbol{\Theta}^l)$ and $q(\boldsymbol{\Theta}^{VB}, \boldsymbol{a})$ respectively, and equating it to zero. At the k^{th} iteration the two steps are given by [29],

$$\begin{aligned} \text{VBE-step:} \quad \ln q(\boldsymbol{\Theta}^l)_{k+1} &\propto \int q(\boldsymbol{\Theta}^{VB}, \boldsymbol{a})_k \ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}^l|\boldsymbol{\Theta}) \ d\boldsymbol{\Theta}^{VB} d\boldsymbol{a} \\ &\propto \mathbb{E}_{q(\boldsymbol{\Theta}^{VB}, \boldsymbol{a})_k}[\ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}^l|\boldsymbol{\Theta})] \ . \end{aligned} \tag{15}$$

$$\begin{aligned} \text{VBM-step:} \quad \ln q(\boldsymbol{\Theta}^{VB}, \boldsymbol{a})_{k+1} &\propto \ln p(\boldsymbol{\Theta}^{VB}, \boldsymbol{a}) + \int q(\boldsymbol{\Theta}^l)_{k+1} \ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}^l|\boldsymbol{\Theta}) \ d\boldsymbol{\Theta}^l \\ &\propto \ln p(\boldsymbol{\Theta}^{VB}, \boldsymbol{a}) + \mathbb{E}_{q(\boldsymbol{\Theta}^l)_{k+1}}[\ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}^l|\boldsymbol{\Theta})] \ , \end{aligned} \tag{16}$$

where $p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Theta}^l|\boldsymbol{\Theta})$ is given by the complete-data likelihood in (9), and $\boldsymbol{\Theta} = [\boldsymbol{\Theta}^{VB}, \boldsymbol{\Theta}^{ML}]$. $\mathbb{E}_{q(\cdot)}$ is the expectation with respect to the corresponding variational distribution, and $p(\boldsymbol{\Theta}^{VB}, \boldsymbol{a})$ is the prior over the model parameters which are specified in Section 3.1. The lower bound is maximised by iteratively using the update equations given in (15) and (16) until convergence. However, convergence to the global maximum is not guaranteed and several runs with different initial conditions need to be considered to overcome this problem.

3.3. Variational Inference

A VBEM algorithm for the SMM was considered in [27], where a factorised form was assumed between the indicator variables $\boldsymbol{Z}$ and scale variables $\boldsymbol{U}$. The restriction of having a factorised form for the latent random variables was removed in [9] where the authors considered correlations between these two random variables. The approach used here follows that given in [9] since it underestimates less the variance in the posterior distribution. The factorised variational distribution for the MoE model described in this paper, is expressed as,

$$q(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{a}) = q(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})q(\boldsymbol{W}, \boldsymbol{\tau})q(\boldsymbol{a}) \ . \tag{17}$$

The functional form of the variational distributions will be the same as the priors, and this is a consequence of adopting conjugate priors for the model structure. The optimal variational distributions are noted below, and expressed as $q^*(\cdot)$. The VBM-step uses (16) to update the variational distributions of the model parameters and these are shown in equations (18)-(24). The variational update equations for the latent variables $\boldsymbol{Z}$, $\boldsymbol{U}$ and $\boldsymbol{S}$ in the VBE-step, using (15), are shown in equations (25)-(31).

The joint variational distribution of the gate mean and covariance is a Gaussian-Wishart distribution, given by

$$q^*(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = \mathcal{N}(\boldsymbol{\mu}_i | \boldsymbol{m}_i, (\beta_i \boldsymbol{\Lambda}_i)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_i | \boldsymbol{B}_i, \boldsymbol{\nu}_i) \ , \tag{18}$$

where

$$\begin{aligned}
\boldsymbol{m}_i &= \frac{\beta_0 \boldsymbol{m}_0 + \sum_{n=1}^{N} \mathbb{E}[z_{ni}] \mathbb{E}[u_{ni}] \boldsymbol{x}_n}{\beta_i} \ , \quad \beta_i = \beta_0 + \sum_{n=1}^{N} \mathbb{E}[z_{ni}] \mathbb{E}[u_{ni}] \\
\boldsymbol{B}_i{}^{-1} &= \boldsymbol{B}_0^{-1} + \sum_{n=1}^{N} \mathbb{E}[z_{ni}] \mathbb{E}[u_{ni}] \boldsymbol{x}_n \boldsymbol{x}_n^\top + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^\top - \beta_i \boldsymbol{m}_i \boldsymbol{m}_i^\top \\
\nu_i &= \nu_0 + N_i \ , \quad N_i = \sum_{n=1}^{N} \mathbb{E}[z_{ni}] \ .
\end{aligned} \tag{19}$$

The joint variational distribution of the expert functions' mean and variance is a Gaussian-Gamma distribution having the following form

$$q^*(\boldsymbol{w}_i, \tau_i) = \mathcal{N}(\boldsymbol{w}_i | \hat{\boldsymbol{w}}_i, \Psi_i) \mathcal{G}a(\tau_i | \rho_i, \lambda_i) \ , \tag{20}$$

where

$$\begin{aligned}
\hat{\boldsymbol{w}}_i &= \boldsymbol{L}_i [\boldsymbol{X} \ \boldsymbol{1}]^\top \boldsymbol{V}_i \boldsymbol{y} \\
\boldsymbol{L}_i &= ([\boldsymbol{X} \ \boldsymbol{1}]^\top \boldsymbol{V}_i [\boldsymbol{X} \ \boldsymbol{1}] + \Upsilon_i)^{-1} \\
\boldsymbol{V}_i &= \mathrm{diag}(\mathbb{E}[z_{1i}] \mathbb{E}[s_{1i}], \ldots, \mathbb{E}[z_{Ni}] \mathbb{E}[s_{Ni}]) \\
\Psi_i &= \frac{\lambda_i}{\rho_i} \boldsymbol{L}_i \\
\rho_i &= \rho_0 + 0.5 N_i \ , \quad \lambda_i = \lambda_0 + 0.5 R_i \\
R_i &= (\boldsymbol{y} - [\boldsymbol{X} \ \boldsymbol{1}] \hat{\boldsymbol{w}}_i)^\top \boldsymbol{V}_i (\boldsymbol{y} - [\boldsymbol{X} \ \boldsymbol{1}] \hat{\boldsymbol{w}}_i) + \hat{\boldsymbol{w}}_i^\top \Upsilon_i \hat{\boldsymbol{w}}_i \ .
\end{aligned} \tag{21}$$

The term $\Upsilon_i = \mathbb{E}_{\boldsymbol{a}_i}[A_i]$ is defined in (24). The variational distribution for the ARD parameters is

$$q^*(a_{i,j}) = \mathcal{G}a(a_{i,j} | c_i, d_{i,j}) \ , \tag{22}$$

where

$$\begin{aligned}
c_i &= c_0 + 0.5 \ , \quad d_{i,j} = d_0 + 0.5 \xi_{i,j} \\
\xi_{i,j} &= \frac{\rho_i}{\lambda_i} \hat{w}_{i,j}^2 + (\boldsymbol{L}_i)_{j,j} \ ,
\end{aligned} \tag{23}$$

where $(\boldsymbol{L}_i)_{j,j}$ is the j^{th} diagonal element of $\boldsymbol{L}_i$, and $\hat{\boldsymbol{w}}_i = \{\hat{w}_{i,j}\}_{j=1}^{d^x+1}$. Using the statistic of a mean from a Gamma distribution, then

$$\Upsilon_i = \mathbb{E}_{\boldsymbol{a}_i}[A_i] = \text{diag}\left(\frac{c_i}{d_{i,1}}, \ldots, \frac{c_i}{d_{i,d^x+1}}\right) . \tag{24}$$

The VBE-step consists of updating the variational distribution of $\boldsymbol{Z}$, $\boldsymbol{U}$ and $\boldsymbol{S}$. The relevant equations are listed below, and the full derivation can be found in Appendix A. The variational distribution for the latent indicator variables follows a multinomial distribution, such that

$$\ln q^*(\boldsymbol{Z}) = \sum_{n=1}^{N}\sum_{i=1}^{M} z_{ni} \ln r_{ni} \quad \text{and} \quad r_{ni} = \frac{\gamma_{ni}}{\sum_{l=1}^{M}\gamma_{nl}} , \tag{25}$$

where $\mathbb{E}[z_{ni}] = r_{ni}$, and once the scale variables $\boldsymbol{U}$ and $\boldsymbol{S}$ have been marginalised out gives

$$\gamma_{ni} = \frac{\Gamma\left(\frac{\eta_i + d^x}{2}\right)}{\Gamma(0.5\eta_i)(\eta_i\pi)^{0.5d^x}} \pi_i \hat{\Lambda}_i^{0.5} \left(\frac{\varpi_{ni}}{\eta_i} + 1\right)^{-\frac{\eta_i + d^x}{2}} \times \frac{\Gamma\left(\frac{\kappa_i+1}{2}\right)}{\Gamma(0.5\kappa_i)(\kappa_i\pi)^{0.5}} \hat{\tau}_i^{0.5} \left(\frac{\xi_{ni}}{\kappa_i} + 1\right)^{-\frac{\kappa_i+1}{2}} . \tag{26}$$

The first part of (26) comprises the contribution of the gate, while the second part is due to the expert parameters. Both parts form individual weighted Student-t distributions. The required statistics are,

$$\begin{aligned} \ln \tilde{\Lambda}_i &= \mathbb{E}[\ln|\Lambda_i|] = \sum_{j=1}^{d^x} \psi\left(\frac{\nu_i + 1 - j}{2}\right) + \ln|\boldsymbol{B}_i| , \\ \ln \tilde{\tau}_i &= \mathbb{E}[\ln \tau_i] = \psi(\rho_i) - \ln(\lambda_i) , \end{aligned} \tag{27}$$

where $\psi(\cdot)$ is the digamma function, and ϖ_{ni} and ξ_{ni} are given in (Appendix A.3) and (Appendix A.6) respectively. The variational distribution for the scale variables $\boldsymbol{U}$ and $\boldsymbol{S}$ follow Gamma distributions. These are given by

$$q^*(u_{ni}|z_{ni} = 1) = \mathcal{G}a(u_{ni}|\alpha_{ni}^u, \epsilon_{ni}^u) , \tag{28}$$

where

$$\alpha_{ni}^u = \frac{d^x + \eta_i}{2} , \quad \epsilon_{ni}^u = \frac{\varpi_{ni} + \eta_i}{2} . \tag{29}$$

Similarly for $\boldsymbol{S}$,

$$q^*(s_{ni}|z_{ni} = 1) = \mathcal{G}a(s_{ni}|\alpha_{ni}^s, \epsilon_{ni}^s) , \tag{30}$$

where

$$\alpha_{ni}^s = \frac{1 + \kappa_i}{2} , \quad \epsilon_{ni}^s = \frac{\xi_{ni} + \kappa_i}{2} . \tag{31}$$

Thus, in this section, all the optimal variational posterior update equations have been summarised. The update equations in the VBM-step are very similar to

those obtained in the MoE with Gaussian gates and experts with the exception that now some of the equations depend on $\mathbb{E}[u_{ni}]$ and $\mathbb{E}[s_{ni}]$, and so derivations are not given here but can be found in [19]. Details regarding the VBE-step can be found in Appendix A. The equations are coupled, and therefore need to be iterated until convergence. However, some of the above equations also require the parameters $\boldsymbol{\pi}$, $\boldsymbol{\eta}$ and $\boldsymbol{g}$ to be known. Estimation of these parameters is dealt with in the following sections.

3.4. Variational Lower Bound

The quantities needed to evaluate the variational lower bound (VLB), $\mathcal{F}(q(\boldsymbol{\vartheta}))$, are obtained from the functional forms of the variational distributions calculated in the previous section. Interested readers are referred to [29] for a derivation of the VLB as expressed below. The VLB for the mixture of experts model is

$$\begin{aligned}\mathcal{F}(q) &= \mathbb{E}_q\left[\ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{a}|\boldsymbol{\Theta}^{ML})\right] - \mathbb{E}_q\left[\ln q(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{a})\right] \\ &= \mathbb{E}_q\left[\ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}|\boldsymbol{\Theta}^{ML}, \boldsymbol{\Theta}^{VB})\right] + \mathbb{E}_q\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \mathbb{E}_q\left[\ln p(\boldsymbol{W}, \boldsymbol{\tau}|\boldsymbol{a})\right] + \mathbb{E}_q\left[\ln p(\boldsymbol{a})\right] \\ &\quad - \mathbb{E}_q\left[\ln q^*(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S})\right] - \mathbb{E}_q\left[\ln q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] - \mathbb{E}_q\left[\ln q^*(\boldsymbol{W}, \boldsymbol{\tau})\right] - \mathbb{E}_q\left[\ln q^*(\boldsymbol{a})\right] \ ,\end{aligned} \tag{32}$$

where the $\mathbb{E}_q$ refers to the expectation with respect to the variational distribution $q^*(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}, \boldsymbol{\theta}^g, \boldsymbol{\theta}^e, \boldsymbol{a})$. This lower bound approximates the true marginal log-likelihood when convergence is reached. The specific expression for the lower bound is given in Appendix B.

3.5. Optimising $\boldsymbol{\pi}$, $\boldsymbol{\eta}$ and $\boldsymbol{\kappa}$ via Maximum Likelihood

By maximising the variational lower bound with respect to the parameter of interest, a corresponding update equation can be obtained. Taking the derivative of (Appendix B.1) with respect to the mixing coefficients (only this term is dependent on $\boldsymbol{\pi}$, all the other terms can be ignored) and setting it to zero, gives [32]

$$\pi_i = \frac{1}{N}\sum_{n=1}^{N} r_{ni} \ . \tag{33}$$

Maximising the mixing coefficients in this way ensures that any surplus experts will have $\pi_i \to 0$. Thus the number of experts can be set large and any excess experts can be eliminated from the model. The degree-of-freedom parameter $\boldsymbol{\eta}$ is also found by maximising the expression obtained in (32), specifically (Appendix B.2). However, this results in the nonlinear equation

$$\ln\frac{\eta_i}{2} + 1 - \psi\Big(\frac{\eta_i}{2}\Big) + \frac{1}{N_i}\sum_{n=1}^{N} r_{ni}\{\mathbb{E}[\ln u_{ni}] - \mathbb{E}[u_{ni}]\} = 0 \ , \tag{34}$$

which requires a line search algorithm to solve for η_i. In order to reduce computational complexity, an approximate closed form solution can be obtained using

Stirling's series for $\ln\Gamma(\cdot)$ [37]. This gives

$$\eta_i = \frac{1}{\frac{1}{N_i}\sum_{n=1}^{N} r_{ni}\{\mathbb{E}[u_{ni}] - \mathbb{E}[\ln u_{ni}]\} - 1} \ . \tag{35}$$

Similarly for $\boldsymbol{\kappa}$ (using Stirling's series and differentiating (Appendix B.3)) gives

$$\kappa_i = \frac{1}{\frac{1}{N_i}\sum_{n=1}^{N} r_{ni}\{\mathbb{E}[\kappa_{ni}] - \mathbb{E}[\ln \kappa_{ni}]\} - 1} \ . \tag{36}$$

Approximate solutions for updating the degree-of-freedom parameter has been applied successfully in [38] using a direct approximate formula and in [7] using Stirling's series. The derivation for (35) (and consequently (36)) is found in Appendix C. All the above equations are required for evaluation of some of the expressions obtained in Section 3.3, and so these update equations are interleaved into the iterative procedure.

3.6. Posterior Predictive Distribution

In order to perform predictions of the output to an unseen input $\boldsymbol{x}_{N+1}$, the posterior predictive distribution needs to be evaluated. The posterior predictive distribution is given by $p(y_{N+1}|\boldsymbol{x}_{N+1}, \mathcal{D})$, where $\mathcal{D} = [\boldsymbol{y}, \boldsymbol{X}]$ is the training data. This distribution is obtained by marginalising the product of the likelihood and the parameter posterior distribution with respect to the parameters. The predictive distribution is similar to that obtained for MoE with GMM gates ans experts, see [16, 19] for proofs, with the exception that now the scale variable s_{ni} also appears in the expression.

In order to obtain an analytical solution for the predictive distribution, s_{ni} cannot be marginalised out, so its maximum-a-posteriori (MAP) estimate is used instead (obtained from (30) and (31)). Letting $n' = N + 1$, the predictive distribution is given by

$$p(y_{n'}|\boldsymbol{x}_{n'}, \mathcal{D}) = \sum_{i=1}^{M} \phi_{n',i} \\ \mathcal{T}\left(y_{n'}\middle|\hat{\boldsymbol{w}}_i^{\top}[\boldsymbol{x}_{n'}\ 1], \frac{\rho_i s_{n'i}^{\mathrm{MAP}}}{\lambda_i}(1 + s_{n'i}^{\mathrm{MAP}}[\boldsymbol{x}_{n'}\ 1]\boldsymbol{L}_i[\boldsymbol{x}_{n'}\ 1]^{\top})^{-1}, 2\rho_i\right) \ , \tag{37}$$

where $\{\phi_{n',i}\}_{i=1}^{M}$ take value 1 with probabilities $\{g_i(\boldsymbol{x}_{n'}, \pi_i, \theta^g_{i\mathrm{MAP}})\}_{i=1}^{M}$ respectively (using (5) at the maximum a posteriori (MAP) estimates $\boldsymbol{\theta}^g_{\mathrm{MAP}} = \{\boldsymbol{\mu}_{\mathrm{MAP}}, \boldsymbol{\Lambda}_{\mathrm{MAP}}\}$ obtained from the posterior distribution (18), and the final value for $\boldsymbol{\pi}$). At any given time n' only one $\{\phi_{n',i}\}_{i=1}^{M}$ can be 1 (the rest are zero) corresponding to the gate with the largest probability. The relevant statistics for prediction are

$$\mathbb{E}[y_{n'}] = \sum_{i=1}^{M} \phi_{n',i}\hat{\boldsymbol{w}}_i^{\top}[\boldsymbol{x}_{n'}\ 1] \ , \tag{38}$$

and

$$\text{Var}[y_{n'}] = \sum_{i=1}^{M} \phi_{n',i} \frac{\lambda_i (1 + s_{n'i}^{\text{MAP}} [\boldsymbol{x}_{n'}\ 1] \boldsymbol{L}_i [\boldsymbol{x}_{n'}\ 1]^\top)}{s_{n'i}^{\text{MAP}} (\rho_i - 1)} \ . \tag{39}$$

In the event that no outliers are present, then the Student-t distribution at the gates and experts will reduce to a Gaussian distribution, and the posterior predictive distribution will be the same as in [19]: $\kappa_i \to \infty$ and so $s_{n'i}^{\text{MAP}} \to 1$.

Algorithm 1: **VBEM algorithm for robust MoE**

Initialise the hyperparameters for Student-t gates, $\boldsymbol{m}_0$, β_0, B_0 and ν_0.
Initialise the hyperparameters for linear experts ρ_0, λ_0, c_0, d_0 and $\Upsilon_i^{(0)} = \frac{c_0}{d_0} \mathbf{I}_{d_x+1}\ \forall\ i$.
Initialise $\kappa_{ni}^{(0)} = \eta_{ni}^{(0)} = 1$ and $\gamma_{ni}^{(0)} \sim \mathcal{U}[0,1]\ \forall\ i, n$.
for $k = 0$: stopping criteria

$$(k') = (k+1)$$

1. Evaluate mixing coefficients $\boldsymbol{\pi}^{(k')}$ via (33).
2. Update the gate parameters $\boldsymbol{m}_i^{(k')}$, $\beta_i^{(k')}$, $\nu_i^{(k')}$, $B_i^{-1(k')}$ via (19).
3. Update the expert parameters $\hat{\boldsymbol{w}}_i^{(k')}$, $\Psi_i^{(k')}$, $\rho_i^{(k')}$, $\lambda_i^{(k')}$ via (21).
4. Update ARD parameters $c_i^{(k')}$, $d_{i,j}^{(k')}$ via (23).
5. Update for $\gamma_{ni}^{(k')}$ of variational distribution of $\boldsymbol{Z}$ via (26).
6. Update for parameters of variational distribution of $\boldsymbol{U}$ via (29).
7. Evaluate degree-of-freedom parameter $\boldsymbol{\eta}^{(k')}$ via (35).
8. Update for parameters of variational distribution of $\boldsymbol{S}$ via (31).
9. Evaluate degree-of-freedom parameter $\boldsymbol{\kappa}^{(k')}$ via (36).

end for

4. Results

In this section, the MoE model with Student-t gates and experts is compared to the MoE model with Gaussian gates and experts (details of this algorithm can be found in [19]) on two datasets: a simulated Duffing oscillator and the Z24 bridge data. For brevity, the two different modelling techniques are referred to as S-MoE and G-MoE respectively. Outliers are artificially added to both datasets in order to show that when outliers are present in the training data, assuming a Gaussian form will result in a biased regression model and/or a more complex model.

The sequence of equations to be executed for the VBEM S-MoE model is given in Algorithm 1. The hyperparameters, β_0, B_0 and ν_0, were set in such a

way so as to define a large covariance (hence a low precision) with respect to the data so as to avoid confining each Gaussian gate to its local cluster, and these were set differently for the two examples (details given in the following sections). The hyperparameter $\boldsymbol{m}_0$, the centre of the gate clusters, was set to zero. Broad priors were assigned to the expert hyperparameters, given by $\rho_0 = c_0 = 0.01$ and $\lambda_0 = d_0 = 1e^{-4}$.

Convergence of the algorithm is achieved by monitoring changes in the variational lower bound (32). The algorithm is stopped when the change in VLB between iterations falls below a certain threshold. In order to overcome the problem of local maxima in the VLB distribution, Algorithm 1 was run for 100 instances with $\gamma_{ni}^{(0)}$ initialised randomly for each run: values were drawn from a Uniform distribution between 0 and 1 (represented as $\mathcal{U}[0,1]$ in Algorithm 1). The model with the largest VLB was selected as being the model that best represents the data.

As already mentioned in Section 3.5, the number of experts, M, needs to be initialised to a large number (here, the term 'large' is relative and depends on the data under investigation), and any mixing coefficient that converges to zero results in its corresponding expert not contributing to the model output. The authors investigated the effect of this initialisation by running the algorithm for different initial values for M; the final results obtained were similar for all cases. In this work, the number of experts was set to 6 for the examples considered in this section, and any $\pi_i < 10^{-5}$ resulted in the corresponding expert to be removed from the final model. The choice of M depends on the data and it is up to the modeller to set. Alternatively, rather than setting a generic threshold; the experts which contribute to the output can be determined, then remove the surplus experts. Here a generic threshold was set in advance.

4.1. The Duffing Oscillator

The nonlinear Duffing oscillator, consisting of a mass, linear and nonlinear springs and a damper, is a classic example used for system identification in dynamics. The values of parameters used here are $m = 1$, $k = 10^4$, $k_3 = 5 \times 10^9$ and $c = 20$ respectively. The differential equation, given by

$$m\ddot{y} + c\dot{y} + ky + k_3 y^3 = P\cos(\omega t) \ , \tag{40}$$

requires an initial displacement y_0 and initial velocity $\dot{y}_0$, along with a forcing amplitude (P) and frequency (ω) to be set. The variable of interest is the displacement of the mass. When the nonlinear spring stiffness constant is not zero one of three possible solutions exists at certain frequencies: one amplitude is unstable and never achieved in practice, and a high or low amplitude is then possible in steady state conditions. The initial conditions of the system determine if it ends up in a low or high amplitude region [39]. Thus, as the initial conditions vary, bifurcations in the amplitude can be seen in the response surface of the system. In this example, the range of y_0 was varied from 0 to 0.0052, while $\dot{y}_0$ was varied from 0 to 0.2, with $P = 10$ and $\omega = 170$. This range of initial conditions results in multiple bifurcations between low and high

amplitudes, including a curved bifurcation front, as seen in Figure 3. Modelling these bifurcations has been tackled using treed Gaussian processes [40] and later extended to G-MoE models in order to model splits in the data that are not parallel to the input variables [19]. In this paper, the effect of outliers on the modelling process is investigated, and outliers are artificially added to the data. The aim here is to fit a MoE model to the response surface of the system, such that it is capable of capturing all the bifurcations accurately, whilst also being insensitive to outliers.

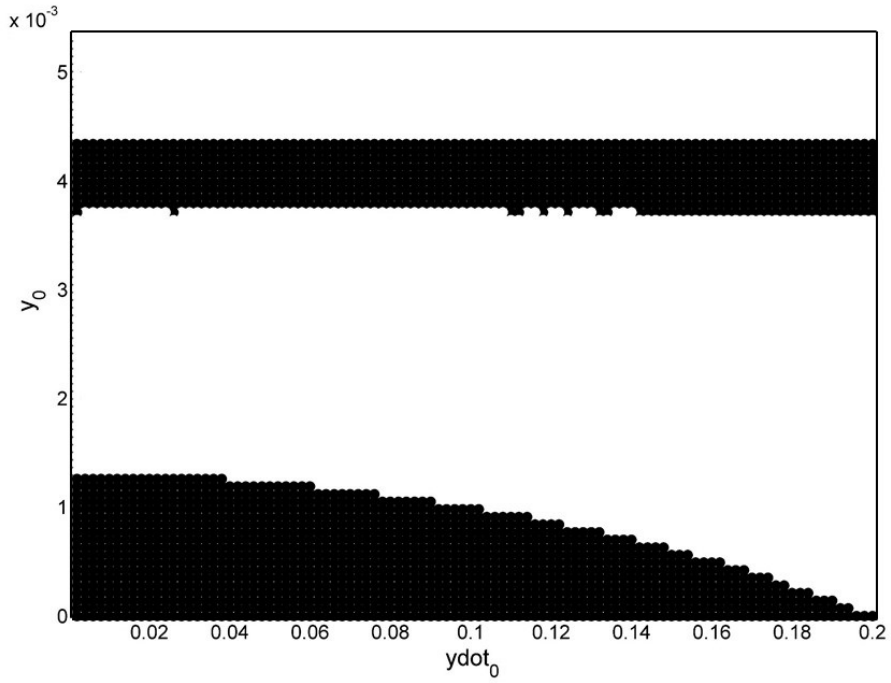

Figure 3: Top-down view representation of response surface for amplitude variation: multiple bifurcations are present between low amplitude (black) and high amplitude (white) as the initial displacement (y0) and velocity (ydot0) of the mass are varied.

100 input points were drawn from a Latin hypercube sampling process, and a Runge-Kutta method was used to simulate the Duffing oscillator for each input pair $(y_0,\dot{y}_0)$. The maximum amplitude from each run was recorded, and this measurement represented the output of the system for the purpose of response surface modelling. The dataset was standardised to zero mean and unit variance. When no outliers are present, the two algorithms perform very similarly producing similar predictive response surfaces, and the results for no outliers are reported in [19]. Atypical data points, numbering 50% of the original dataset were randomly drawn from a uniform distribution along each input and the output. Two situations were analysed: the first case consists of outliers in the

output only (these represent data points which could not have been generated by the process), while the second case deals with outliers in both the inputs and the output (representing the situation of when the true underlying system is masked by noise). The G-MoE and S-MoE models were trained using this dataset.

The results obtained from the G-MoE and S-MoE algorithms are shown in the left and right hand plots respectively in Figure 4; the top row represents the surface plot generated when outlier points (pink scatter points) are restricted to the output variable, while the bottom row shows the effect of expanding the range of outlier points. The S-MoE model (left column) provides a more accurate predictive response surface than the G-MoE model (right column) for two reasons: firstly, the model captures all of the bifurcations, and secondly, the model is capable of providing accurate predictions because each expert represents the black scatter points (system data). Thus the S-MoE model appears to be insensitive to the outliers present in the training data. The bottom right plot shows that outlier points outside the region of interest are assigned the same expert, and an accurate model for the bifurcations and system data points is still obtained. The use of a Student-t distribution is now effective since $\kappa_i < 1$ for all the dominant experts, thus the heavier tails ensure robustness in the regression analysis.

On the other hand, the G-MoE model performs additional splits to the data providing a more complex and incorrect model which does not capture the correct bifurcations. In addition, the G-MoE with outliers fails to provide accurate regression modelling since it does pass thorough the black scatter points, which represent the system. The G-MoE fails to capture the true underlying model since the predictive surface plot is highly influenced by outliers. Thus, the S-MoE algorithm is superior to the G-MoE model in the presence of outliers, providing a simpler more accurate model than the G-MoE model counterpart.

4.2. Z24 Bridge Data

The Z24 bridge was a bridge in Switzerland that prior to its demolishment in the late 1990s was under intense monitoring by the 'SIMCES project' [41]. The modal parameters of the bridge were tracked, and realistic damage scenarios were gradually introduced. Environmental factors were also measured, such as air temperature, soil temperature and humidity among several other variables. The Z24 bridge has been well studied within the structural health monitoring (SHM) community in order to establish detection of damage independently of environmental factors [42, 43].

In this paper, the relationship between the air temperature at the deck top and the second natural frequency of the Z24 bridge is investigated. Large fluctuations in the natural frequency are observed before any damage occurs, which are associated with periods of very cold temperatures. These very cold periods cause the asphalt to freeze hence causing the stiffness of the bridge to increase. Interested readers are referred to [42] regarding the change in material properties below and above freezing temperatures. The training dataset analysed here consists of the portion of data where the modal frequency is affected by

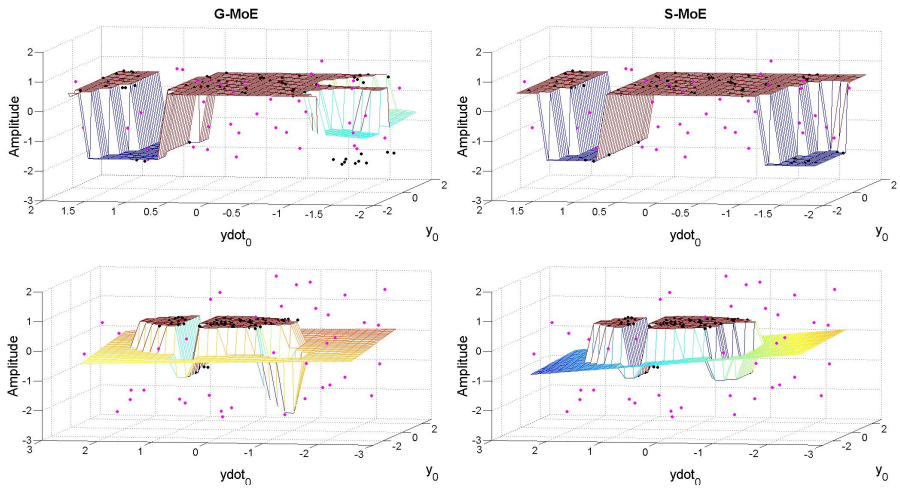

Figure 4: Comparison of the G-MoE model (left hand column) and S-MoE model (right hand column). The top row represents the surface response plot for the case when outliers are present in the output only, while the bottom row deals with the situation when outliers are present in both the inputs and output. The mesh represents the mean predictive surface plot, with the scatter points representing the training data: black points are underlying system while pink points are the artificially added outliers. The S-MoE model captures the underlying system by accurately modelling the black scatter points, whilst the G-MoE is highly influenced by outliers and hence fails to model the black scatter points. Note that the axes represent the normalised data.

temperature variation only. This dataset was analysed using treed GPs [44] and G-MoE [19] since a bilinear relationship exists between the air temperature and the natural frequency of the bridge. Using models that are capable of automatically switching between different regimes are important for modelling and understanding the underlying physics governing the system. The aim of the modelling procedure is to obtain a model that is dependent on temperature only. The model obtained using this training dataset is then tested on data that contains both temperature changes and damage effect. When predictions are performed on the damaged section, the model should be capable of giving an indication that other factors besides temperature are affecting the modal frequency. Within a Bayesian setting the variance of the predicted signal can be calculated naturally, and hence credible bounds can be computed. Damage is detected when the measured signal deviates significantly from the predictions, which can be determined when the signal moves outside the credible intervals. Switching models, on this dataset, outperformed standard GP models with respect to determining damage detection [44].

The G-MoE and S-MoE algorithms were implemented on the Z24 training data. Another input variable, the square of the temperature, was introduced so as to improve the model's predictive capabilities. Since the variance is used to establish whether damage has occurred, the exact solution for $\boldsymbol{\eta}$ and $\boldsymbol{\kappa}$, given in

(34), is used since Stirling's approximation tends to underestimate the variance (see Appendix C) causing tighter bounds. The expression obtained in (34) can be solved in Matlab® using the `fzero` function. The dataset is first run with no outliers, and the results are shown in Figure 5. The plots in the top row show the variational lower bound versus the number of experts in the final models obtained using 100 random runs. The S-MoE model achieves a tighter bound (larger VLB value) with a less complex structure (2 experts versus the 3 experts needed by the G-MoE model). Note, how for the S-MoE, models with 3 experts achieved a lower VLB than the models with 2 experts because complexity is naturally penalised within a Bayesian framework. The G-MoE has splits at 0.375°C and 13.4°C, while the S-MoE requires one split at 0.84°C. Thus both models have a split close to 0°C, however the S-MoE has a less complex structure since it combines two experts into one (and the degree-of-freedom parameters for this component have very low values for both the gate and expert). Both models are capable of detecting damage in the bridge since the second natural frequency values quickly move outside the credible bounds of the model (Figure 5, bottom row).

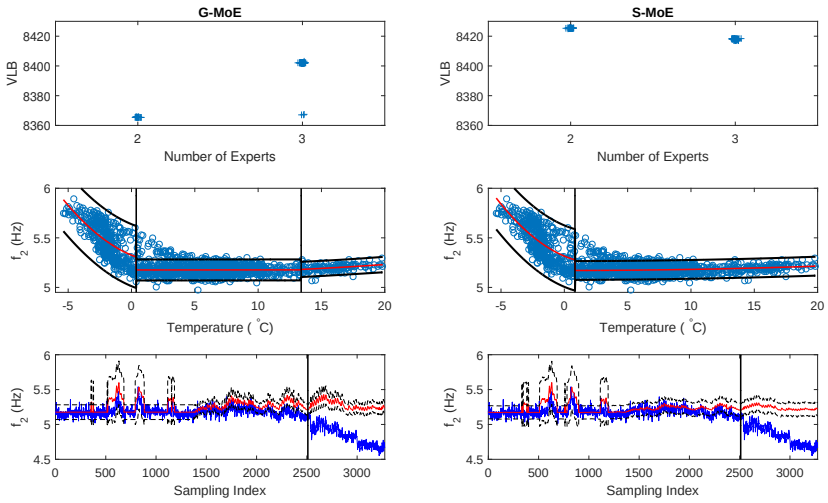

Figure 5: Comparison of the G-MoE model (left hand column) versus the S-MoE model (right hand column) on the Z24 dataset when no outliers are present. The top row shows the variational lower bound versus number of experts. In order enhance the visualization of the results, a small amount of uniform noise has been added to the horizontal position of the points. The middle row shows the relationship of the second natural frequency with temperature: blue scatter points represent the training data, the red line represents the model mean and the black lines represent ±99% credible intervals. The black vertical lines indicate the different expert regions assigned by the corresponding models. The bottom row plots are the predictions of the models (red) on the test data (blue), where the black dashed line represents the ±99% credible intervals. The black vertical line indicates start of damage.

Outlier points are artificially added to the Z24 bridge data, numbering 15% of the original training dataset. The outliers were randomly drawn from a uniform distribution along the input and the output ranges. Figure 6 shows the results obtained when the G-MoE and S-MoE algorithms are run on this new dataset. It is immediately obvious that the G-MoE model fails to represent the data very accurately since the regression is severely affected by the outliers. In particular the variance of the predicted output is very wide in order to accommodate the outliers (since it is highly influenced by outlier points). As a consequence, the G-MoE model is incapable of detecting damage to the bridge due to a very wide variance associated with the predictions, such that the credible intervals now enclose the measured modal frequency of the test set, as shown in the left bottom plot in Figure 6. On the other hand, the S-MoE successfully captures the dynamics of the system having splits at 0.21°C and 12.7°C (the model has introduced an extra split in order to accommodate differences in the variance of each section of the data). The S-MoE model is still capable of detecting damage to the bridge in the presence of outliers, since the second natural frequency values quickly move outside the credible bounds of the model (bottom right in Figure 6). So even for very few outlier points, the G-MoE model can give very biased results which in this case would lead to a wrong interpretation regarding the structural health of the bridge. On the other hand, the S-MoE algorithm provides a robust model, in the presence of outliers, that performs very similarly to the model obtained when no outliers were present.

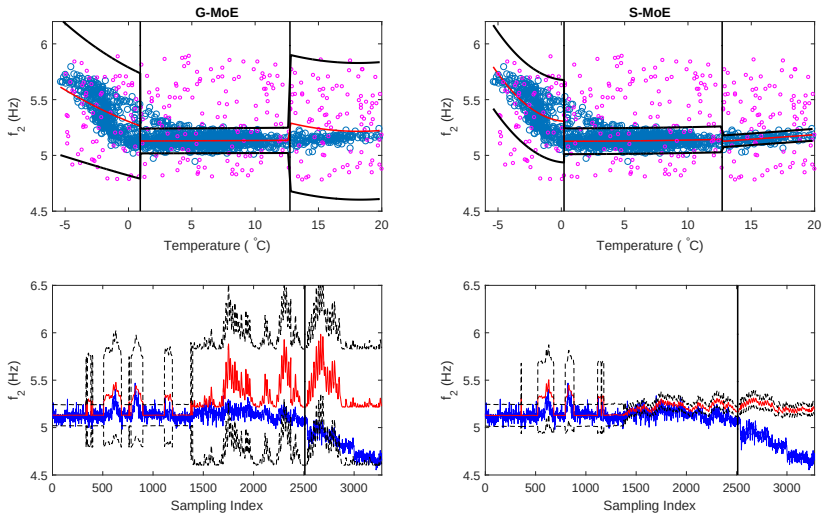

Figure 6: Comparison of the G-MoE model (left hand column) versus the S-MoE model (right hand column) on the Z24 dataset when 15% outliers are added. The top row shows the relationship of the second natural frequency with temperature: blue scatter points represent the training data, pink scatter points are the artificially added outliers, the red line represents the model mean and the black lines represent $\pm 99\%$ credible intervals. The black vertical lines indicate the different expert regions assigned by the corresponding models. The bottom row plots are the predictions of the models (red) on the test data (blue), where the black dashed line represents the $\pm 99\%$ credible intervals. The black vertical line indicates start of damage.

5. Conclusions

In this article a novel mixture of experts model using Student-t distributions was developed for robust regression and robust selection of the number of experts. The model was trained within a variational Bayesian framework using methods developed in the mixture modelling literature. The learning algorithm for this robust MoE consists of closed-form parameter update equations, hence providing very fast training times. It has been demonstrated that the S-MoE is a powerful model which can successfully capture bifurcations/discontinuities in the data in the presence of outliers. The S-MoE model proved to be insensitive to outliers, and hence predictions can be estimated with confidence. The variance of the predictions is also insensitive to outliers, unlike that of the G-MoE which gives large credible intervals in order to account for the outliers. This insensitivity to atypical points is crucial, for example, when performing structural health monitoring of the Z24 bridge. Hence, the S-MoE model leads to robust mixture and regression modelling in practice.

Acknowledgement

Author T. Baldacchino would like to thank the Leverhulme Trust Research Project Grant for financial support. The authors would like to acknowledge Dr. Elizabeth J. Cross, from the University of Sheffield, for providing access to the Z24 bridge data.

Appendix A. Derivation of $\ln q(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S})$

The joint distribution over the latent variables is given by

$$\begin{aligned}\ln q^*(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}) &= \mathbb{E}[\ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}, \boldsymbol{\Theta}^{VB}, \boldsymbol{a})|\boldsymbol{\Theta}^{ML}] \\ &= \mathbb{E}[\ln p(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}|\boldsymbol{\Theta}^{VB}, \boldsymbol{\Theta}^{ML}] + C \\ &\propto \sum_{n=1}^{N}\sum_{i=1}^{M} z_{ni}\Big(\underbrace{\mathbb{E}[\ln\{\pi_i \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_i, (u_{ni}\boldsymbol{\Lambda}_i)^{-1})\mathcal{G}a(u_{ni}|\eta_i/2, \eta_i/2)\}]}_{\text{lnP1}} \\ &\quad + \underbrace{\mathbb{E}[\ln\{\mathcal{N}(y_n|[\boldsymbol{x}_n\ 1]\boldsymbol{w}_i, (s_{ni}\tau_i)^{-1})\mathcal{G}a(s_{ni}|\kappa_i/2, \kappa_i/2)\}]}_{\text{lnP2}}\Big)\ ,\end{aligned} \tag{Appendix A.1}$$

where C contains terms independent of $\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S}$. Concentrating on the first part of the right hand side of the equation (lnP1), and expanding the terms gives,

$$\begin{aligned}\ln P1 = \ln \pi_i - 0.5d^x \ln(2\pi) + 0.5\ln\tilde{\Lambda}_i - \ln\Gamma(0.5\eta_i) + \frac{\eta_i}{2}\ln\frac{\eta_i}{2} \\ + 0.5d^x \ln u_{ni} - 0.5u_{ni}\varpi_{ni} - (\frac{\eta_i}{2} - 1)\ln u_{ni} - \frac{\eta_i}{2}u_{ni}\ ,\end{aligned} \tag{Appendix A.2}$$

where

$$\begin{aligned}\varpi_{ni} &= \mathrm{Tr}(\mathbb{E}[\boldsymbol{\Lambda}_i]\mathbb{E}[(\boldsymbol{x}_n - \boldsymbol{\mu}_i)(\boldsymbol{x}_n - \boldsymbol{\mu}_i)^\top]) \\ &= \nu_i(\boldsymbol{x}_n - \boldsymbol{m}_i)\boldsymbol{B}_i(\boldsymbol{x}_n - \boldsymbol{m}_i)^\top + \beta_i^{-1}d^x\ ,\end{aligned} \tag{Appendix A.3}$$

and the expectation $\mathbb{E}[\cdot]$ is taken according to the variational posterior distribution of that parameter. In order to marginalise out the scale variables u_{ni} from (Appendix A.2), a distribution over u_{ni} is formed, and adding any terms

in order to compensate for this marginalisation results in

$$\begin{aligned}
\ln P1 &= \ln \pi_i + 0.5 \ln \hat{\Lambda}_i + \ln \mathcal{G}a\Big(u_{ni}\Big|\frac{\eta_i + d^x}{2}, \frac{\varpi_{ni} + \eta_i}{2}\Big) - \frac{\eta_i + d^x}{2} \ln \frac{\varpi_{ni} + \eta_i}{2} \\
&\quad - 0.5 d^x \ln(2\pi) + \ln \Gamma\Big(\frac{\eta_i + d^x}{2}\Big) - \ln \Gamma(0.5\eta_i) + \frac{\eta_i}{2} \ln \frac{\eta_i}{2} \\
&= \ln \pi_i + 0.5 \ln \hat{\Lambda}_i + \ln \mathcal{G}a\Big(u_{ni}\Big|\frac{\eta_i + d^x}{2}, \frac{\varpi_{ni} + \eta_i}{2}\Big) - \frac{\eta_i + d^x}{2} \ln \frac{\eta_i}{2}\left(\frac{\varpi_{ni}}{\eta_i} + 1\right) \\
&\quad - \frac{d^x}{2} \ln(2\pi) + \ln \Gamma\Big(\frac{\eta_i + d^x}{2}\Big) - \ln \Gamma(0.5\eta_i) + \frac{\eta_i}{2} \ln \frac{\eta_i}{2} \\
&= \ln \pi_i + 0.5 \ln \hat{\Lambda}_i + \ln \mathcal{G}a\Big(u_{ni}\Big|\frac{\eta_i + d^x}{2}, \frac{\varpi_{ni} + \eta_i}{2}\Big) - \frac{\eta_i + d^x}{2} \ln \left(\frac{\varpi_{ni}}{\eta_i} + 1\right) \\
&\quad + \ln \Gamma\Big(\frac{\eta_i + d^x}{2}\Big) - \ln \Gamma(0.5\eta_i) - \frac{d^x}{2} \ln(\eta_i \pi)
\end{aligned}$$

Taking the exponential of the above equation gives

$$P1 = \frac{\Gamma\Big(\frac{\eta_i + d^x}{2}\Big)}{\Gamma(0.5\eta_i)(\eta_i \pi)^{0.5 d^x}} \pi_i \hat{\Lambda}_i^{0.5} \left(\frac{\varpi_{ni}}{\eta_i} + 1\right)^{-\frac{\eta_i + d^x}{2}} \mathcal{G}a\Big(u_{ni}\Big|\frac{\eta_i + d^x}{2}, \frac{\varpi_{ni} + \eta_i}{2}\Big) \,, \tag{Appendix A.4}$$

where the first part is a weighted Student-t distribution. Similarly, for the part contributed by the expert (P2) gives

$$P2 = \frac{\Gamma\Big(\frac{\kappa_i + 1}{2}\Big)}{\Gamma(0.5\kappa_i)(\kappa_i \pi)^{0.5}} \hat{\tau}_i^{0.5} \left(\frac{\xi_{ni}}{\kappa_i} + 1\right)^{-\frac{\kappa_i + 1}{2}} \mathcal{G}a\Big(s_{ni}\Big|\frac{\kappa_i + 1}{2}, \frac{\xi_{ni} + \kappa_i}{2}\Big) \,, \tag{Appendix A.5}$$

where

$$\begin{aligned}
\xi_{ni} &= \mathbb{E}[\tau_i](y_n - [\boldsymbol{x}_n \; 1]\mathbb{E}[\boldsymbol{w}_i])^2 \\
&= \frac{\rho_i}{\lambda_i}(y_n - [\boldsymbol{x}_n \; 1]\hat{\boldsymbol{w}}_i)^2 + [\boldsymbol{x}_n \; 1]\boldsymbol{L}_i[\boldsymbol{x}_n \; 1]^\top \,.
\end{aligned} \tag{Appendix A.6}$$

Using (Appendix A.4) and (Appendix A.5) and substituting into (Appendix A.1), gives the posterior variational distribution for $\ln q^*(\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{S})$. Hence, $q^*(\boldsymbol{Z})$ is obtained by marginalising this expression over $\boldsymbol{U}$ and $\boldsymbol{S}$, and noting that the integral of a Gamma distribution is 1 then this proves the expression for γ_{ni} in (26). This variational distribution needs to be normalised, and this is given as

$$\ln q^*(\boldsymbol{Z}) = \sum_{n=1}^{N} \sum_{i=1}^{M} z_{ni} \ln r_{ni} \,, \tag{Appendix A.7}$$

with $r_{ni} = \gamma_{ni} / \sum_l \gamma_{nl}$. Equation (Appendix A.7) follows from the fact that for each value of n, the quantities z_{ni} are binary and $\sum_i z_{ni} = 1$. Since (Appendix A.7) is a multinomial distribution, then it follows that $\mathbb{E}[z_{ni}] = r_{ni}$. The variational distribution of the scale variables u_{ni} and s_{ni} are given by the Gamma distributions obtained in (Appendix A.4) and (Appendix A.5) respectively.

Appendix B. Variational Lower Bound

The expressions for the individual terms in (32) are given here. Letting $\ln \hat{A}_i = \mathbb{E}[\ln |A_i|] = \sum_{j=1}^{d^x+1} (\psi(c_i) - \ln d_{i,j})$, $\ln \hat{u}_{ni} = \mathbb{E}[\ln u_{ni}] = \psi(\alpha^u_{ni}) - \ln \epsilon^u_{ni}$ and $\bar{u}_{ni} = \mathbb{E}[u_{ni}] = \alpha^u_{ni}/\epsilon^u_{ni}$ (and similarly for the latent variables s_{ni}), then:

$$\mathbb{E}_q\left[\ln p(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{\Theta})\right] = \frac{1}{2}\sum_{n,i}^{N,M} r_{ni}\Big\{-d^x \ln 2\pi + d^x \ln \hat{u}_{ni} + \ln \hat{\boldsymbol{\Lambda}}_i - \bar{u}_{ni}\varpi_{ni} - \ln 2\pi + \ln \hat{\tau}_i + \ln \hat{s}_{ni} - \bar{s}_{ni}\xi_{ni}\Big\}$$

where ϖ_{ni} and ξ_{ni} are given in (Appendix A.3) and (Appendix A.6) respectively.

$$\mathbb{E}_q\left[p(\boldsymbol{Z}|\boldsymbol{\pi})\right] = \sum_{n,i}^{N,M} r_{ni} \ln \pi_i \qquad \text{(Appendix B.1)}$$

$$\mathbb{E}_q\left[p(\boldsymbol{U}|\boldsymbol{Z}, \boldsymbol{\eta})\right] = \sum_{n,i}^{N,M} r_{ni}\left\{\frac{\eta_i}{2}\ln\frac{\eta_i}{2} - \ln\Gamma(\frac{\eta_i}{2}) + \left(\frac{\eta_i}{2} - 1\right)\ln \hat{u}_{ni} - \frac{\eta_i}{2}\bar{u}_{ni}\right\} \qquad \text{(Appendix B.2)}$$

$$\mathbb{E}_q\left[p(\boldsymbol{S}|\boldsymbol{Z}, \boldsymbol{g})\right] = \sum_{n,i}^{N,M} r_{ni}\left\{\frac{\kappa_i}{2}\ln\frac{\kappa_i}{2} - \ln\Gamma(\frac{\kappa_i}{2}) + \left(\frac{\kappa_i}{2} - 1\right)\ln \hat{s}_{ni} - \frac{\kappa_i}{2}\bar{s}_{ni}\right\} \qquad \text{(Appendix B.3)}$$

$$\mathbb{E}_q\left[p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{a})\right] = \sum_{i=1}^{M}\frac{1}{2}\Big\{-d^x \ln 2\pi + d^x \ln \beta_0 + \ln \hat{\boldsymbol{\Lambda}}_i - \nu_i\beta_0(\boldsymbol{m}_i - \boldsymbol{m}_0)^\top \boldsymbol{B}_i(\boldsymbol{m}_i - \boldsymbol{m}_0) - \frac{\beta_0}{\beta_i}d^x + 2\ln C_{\mathcal{W}}(\boldsymbol{B}_0, \nu_0) + (\nu_0 - d^x - 1)\ln \hat{\Lambda}_i - \nu_i \mathrm{Tr}(\boldsymbol{B}_0^{-1}\boldsymbol{B}_i)\Big\}$$

$$\mathbb{E}_q\left[p(\boldsymbol{W}, \boldsymbol{\tau})\right] = \sum_{i=1}^{M}\Bigg\{\frac{1}{2}\left[-d^x \ln 2\pi + (d^x + 1)\ln \hat{\tau}_i + \ln \hat{A}_i - \sum_{j=1}^{d^x+1}\frac{c_i}{d_{i,j}}\left(\frac{\rho_i}{\lambda_i}\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_i + (\boldsymbol{L}_i)_{j,j}\right)\right] - \ln\Gamma(\rho_0) + \rho_0 \ln \lambda_0 + (\rho_0 - 1)\ln \hat{\tau}_i - \lambda_0\frac{\rho_i}{\lambda_i}\Bigg\}$$

$$\mathbb{E}_q\left[p(\boldsymbol{a})\right] = \sum_{i,j}^{M,d^x+1}\left\{-\ln\Gamma(c_0) + c_0 \ln d_0 + (c_0 - 1)(\psi(c_i) - \ln d_{i,j}) - d_0\frac{c_i}{d_{i,j}}\right\}$$

The remaining terms in (32) are the entropies of the corresponding variational distributions, such that:

$$\mathbb{E}_q\left[q(\boldsymbol{Z})\right] = \sum_{n,i}^{N,M} r_{ni} \ln r_{ni}$$

$$\mathbb{E}_q\left[q(\boldsymbol{U}|\boldsymbol{Z})\right] = \sum_{n,i}^{N,M} r_{ni} \left\{-\ln\Gamma(\alpha^u_{ni}) + (\alpha^u_{ni}-1)\psi(\alpha^u_{ni}) + \ln\epsilon^u_{ni} - \alpha^u_{ni}\right\}$$

$$\mathbb{E}_q\left[q(\boldsymbol{S}|\boldsymbol{Z})\right] = \sum_{n,i}^{N,M} r_{ni} \left\{-\ln\Gamma(\alpha^s_{ni}) + (\alpha^s_{ni}-1)\psi(\alpha^s_{ni}) + \ln\epsilon^s_{ni} - \alpha^s_{ni}\right\}$$

$$\mathbb{E}_q\left[q(\boldsymbol{\mu},\boldsymbol{\Lambda})\right] = \sum_{i=1}^{M} \frac{1}{2}\left\{d^x \ln\beta_i - d^x(1+\ln 2\pi) + 2\ln C_{\mathcal{W}}(\boldsymbol{B}_i,\nu_i) + (\nu_i - d^x)\ln\hat{\Lambda}_i - d^x\nu_i\right\}$$

$$\begin{aligned}\mathbb{E}_q\left[q(\boldsymbol{W},\boldsymbol{\tau})\right] = \sum_{i=1}^{M} \Bigg\{ &\frac{1}{2}\left[(d^x+1)\ln\tau_i + \ln|\boldsymbol{L}_i| - (d^x+1)(1+\ln 2\pi)\right] \\ &- \ln\Gamma(\rho_i) + (\rho_i - 1)\psi(\rho_i) + \ln\lambda_i - \rho_i\Bigg\}\end{aligned}$$

$$\mathbb{E}_q\left[q(\boldsymbol{a})\right] = \sum_{i,j}^{M,d^x+1} \left\{-\ln\Gamma(c_i) + (c_i-1)\psi(c_i) + \ln d_{i,j} - c_i\right\}$$

where $C_{\mathcal{W}}(\cdot)$ is the normalisation constant associated with the Wishart distribution. The expressions above can be combined to simplify the overall variational lower bound expression.

Appendix C. Stirling's Series

The derivation for expressions (35) and (36) is given here. The Gamma function can be approximated using Stirling's series, and a truncated version of Stirling's series is given by [37]

$$\ln\Gamma(\frac{\eta}{2}) \approx \frac{1}{2}\ln 2\pi + (\frac{\eta}{2} - \frac{1}{2})\ln\frac{\eta}{2} - \frac{\eta}{2}$$

Substituting (Appendix C.1) into (Appendix B.2), and differentiating with respect to η gives

$$\frac{d}{d\eta_i}\sum_n^N r_{ni}\left\{\frac{1}{2}\ln\frac{\eta_i}{2}-\frac{1}{2}\ln 2\pi+\frac{\eta_i}{2}+\left(\frac{\eta_i}{2}-1\right)\ln\hat{u}_{ni}-\frac{\eta_i}{2}\bar{u}_{ni}\right\}=0$$

$$\frac{1}{2}\sum_n^N r_{ni}\left\{\frac{1}{\eta_i}+1+\ln\hat{u}_{ni}-\bar{u}_{ni}\right\}=0$$

$$\frac{N_i}{\eta_i}+N_i-\sum_n^N r_{ni}(\bar{u}_{ni}-\ln\hat{u}_{ni})=0$$

$$\eta_i=\frac{1}{\frac{1}{N_i}\sum_n^N r_{ni}(\bar{u}_{ni}-\ln\hat{u}_{ni})-1}$$

Comparing this equation to the exact equation given in (34), results in $\ln\frac{\eta_i}{2}-\psi(\frac{\eta_i}{2})$ being approximated by $\frac{1}{\eta_i}$. The plot of these two functions is given in Figure (C.7), and it can been seen that the function plots differ for low values of η, with the approximate solution underestimating the value of η. As η increases, the two functions converge. The same procedure is used to obtain (36).

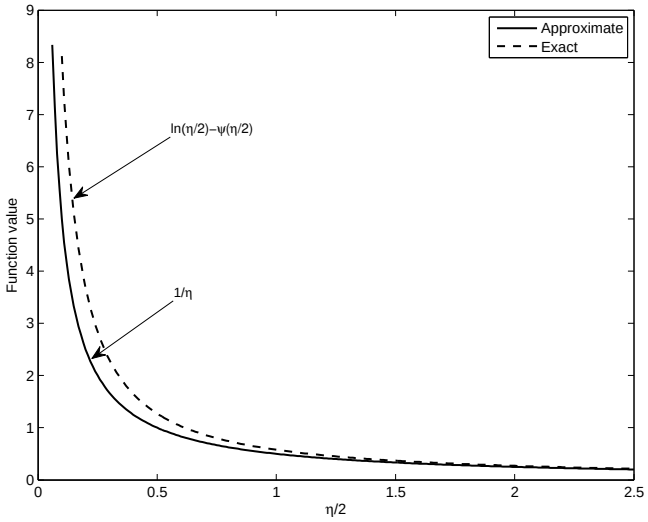

Figure C.7: A plot comparing the approximate solution using Stirling's series and the exact solution given by (34).

[1] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[2] V. Barnett and T. Lewis. *Outliers in Statistical Data (Third Edition)*. John Wiley and Sons, 1994.

[3] Jason W. Osborne and Amy Overbay. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), 2004.

[4] M. West. Robust sequential approximate Bayesian estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(2):157–166, 1981.

[5] Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[6] Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with Student-t likelihood. In *Conference on Neural Information Processing Systems (NIPS)*, 2009.

[7] Jacqueline Christmas and Richard Everson. Robust autoregression: Student-t innovations using variational Bayes. *IEEE Transactions on Signal Processing*, 59(1):48–57, 2011.

[8] Johan Dahlin, Fredrik Lindsten, Thomas B. Schön, and Adrian Wills. Hierarchical Bayesian approaches for robust inference in ARX models. In *The 16th IFAC Symposium on System Identification*, 2012.

[9] Cédric Archambeau and Michel Verleysen. Robust Bayesian clustering. *Neural Networks*, 20:129–138, 2007.

[10] Weixin Yao, Yan Wei, and Chun Yu. Robust mixture regression using the t-distribution. *Computational Statistics and Data Analysis*, 71:116–127, 2014.

[11] C. S. Wong and W. S. Chan. A Student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrica*, 96(3):751–760, 2009.

[12] F. Chamroukhi. Robust mixture of experts modeling using the t distribution. *Neural Networks*, 2016.

[13] James O. Berger, Elías Moreno, Luis Raul Pericchi, M. Jesús Bayarri, José M. Bernardo, Juan A. Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, A. Dasgupta, Paul Gustafson, Larry Wasserman, Joseph B. Kadane, Cid Srinivasan, Michael Lavine, Anthony O'Hagan, Wolfgang Polasek, Christian P. Robert, Constantinos Goutis, Fabrizio Ruggeri, Gabriella Salinetti, and Siva Sivaganesan. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994.

[14] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[15] Steve Waterhouse, David MacKay, and Tony Robinson. Bayesian methods for mixture of experts. In David S. Touretzky Michael C. Mozer and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 351–357. MIT Press, 1996.

[16] Naonori Ueda and Zoubin Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15:1223–1241, 2002.

[17] Christopher M. Bishop and Markus Svensén. Bayesian hierarchical mixture of experts. In *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference*, 2003.

[18] Alexandre X. Carvalho and Martin A. Tanner. Modeling nonlinearities with mixtures-of-experts of time series models. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006.

[19] Tara Baldacchino, Elizabeth J. Cross, Keith Worden, and Jennifer Rowson. Variational Bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. *Mechanical Systems and Signal Processing*, Under Review.

[20] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[21] Fengchun Peng, Robert A. Jacobs, and Martin A. Tanner. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960, 1996.

[22] Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. *IEEE Trans. Neural Netw. Learning Syst.*, 23(8):1177–1193, 2012.

[23] Lei Xu, Michael I. Jordan, and Geoffrey E. Hinton. An alternative model for mixtures of experts. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, pages 633–640. MIT Press, 1995.

[24] Clodoaldo A.M. Lima, André L.V. Coelho, and Fernando J. Von Zuben. Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification. *Information Sciences*, 177(10):2049–2074, 2007.

[25] Carl Edward Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 881–888, 2002.

[26] C. Yuan and C. Neubauer. Variational mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems (NIPS) 21*, pages 1897–1904, 2009.

[27] Markus Svensén and Christopher M. Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, March 2005.

[28] Héctor Allende, Romina Torres, Rodrigo Salas, and Claudio Moraga. Robust learning algorithm for the mixture of experts. In *Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*, volume 2652, pages 19–27, 2003.

[29] Matthew J. Beal and Zoubin Ghahramani. The variational bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In José M. Bernardo, M. J. Bayarri, A. Philip Dawid, James O. Berger, D. Heckerman, A. F. M. Smith, and Mike West, editors, *Bayesian Statistics 7*. Oxford University Press, 2003.

[30] Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.

[31] D. Peel and G.J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

[32] Adrian Corduneanu and Christopher M. Bishop. Variational Bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola, editors, *Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics*, pages 27–34, 2001.

[33] Andrew Gelman, John Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis. Third Edition*. CRC Press, 2014.

[34] David J. C. MacKay. Probable networks and plausible predicitons - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505, 1995.

[35] Radford M. Neal. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics)*. Springer, 1996.

[36] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[37] Chris Impens. Stirling's series made easy. *The American Mathematical Monthly*, 110(8):730–735, 2003.

[38] Shy Shoham. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, 35:1127–1142, 2002.

[39] K. Worden, G. Manson, T.M. Lord, and M.I. Friswell. Some observations on uncertainty propagation through a simple nonlinear system. *Journal of Sound and Vibration*, 288(3):601–621, 2005.

[40] W. Becker, K. Worden, and J. Rowson. Bayesian sensitivity analysis of bifurcating nonlinear models. *Mechanical Systems and Signal Processing*, 34:57–75, 2013.

[41] G. De Roeck. The state-of-the-art of damage detection by vibration monitoring: the SIMCES experience. *Journal of Structural Control,*, 10:127–134, 2003.

[42] Bart Peeters and Guido De Roeck. One-year monitoring of the Z24-Bridge: environmental effects versus damage events. *Earthquake Engineering and Structural Dynamics*, 30:149–171, 2001.

[43] Elizabeth J. Cross. *On Structural Health Monitoring in Changing Environmental and Operational Conditions*. PhD thesis, University of Sheffield, 2012.

[44] Keith Worden, Elizabeth J. Cross, and James M. W. Brownjohn. Switching response surface models for structural health monitoring of bridges. In *Surrogate-Based Modeling and Optimization*, pages 337–358, 2013.