A VARIATIONAL APPROACH TO ROBUST BAYESIAN INTERPOLATION

Michael E. Tipping
Microsoft Research
7 J J Thomson Avenue
Cambridge CB3 0FB, U.K.
mtipping@microsoft.com

Neil D. Lawrence
The University of Sheffield
Regent Court, 211 Portobello St.
Sheffield S1 4DP, U.K.
neil@dcs.shef.ac.uk

Abstract. We detail a Bayesian interpolation procedure for linear-in-the-parameter models which combines both effective complexity control and robustness to outliers. Robustness is obtained by adopting a Student-t noise distribution, defined hierarchically in terms of an inverse-Gamma prior distribution over individual Gaussian observation variances. Importantly, this hierarchical definition enables practical Bayesian variational techniques to concurrently determine both the primary model parameters *and* the form of the noise process. We show that the model is capable of flexibly inferring, from limited data, both Gaussian and more heavily-tailed Student-t noise processes as appropriate.

INTRODUCTION

We consider the classic problem of interpolation where the observation variables are assumed to be noisy. Our data comprises N input-observation pairs $\{\mathbf{x}_n, y_n\}$, and we focus on interpolation models linear in the parameters, where the interpolant $f(\mathbf{x})$ is expressed in terms of M fixed basis functions $\phi_m(\mathbf{x})$, $m = 1, \ldots, M$, weighted by some corresponding parameters θ_m:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}). \tag{1}$$

While the nonlinear functions $\phi_m(\mathbf{x})$ are fixed, $f(\mathbf{x})$ may still be very flexible if a large set of basis functions is utilised. This, of course, depends on the use of effective complexity control techniques, as exemplified by the recent and popular "support vector machine" [8] and "sparse Bayesian" [7] frameworks.

Given (1), it is conventional (and usually realistic) to assume that the observation variables deviate from the functional mapping by some additive i.i.d. noise process: $y_n = f(\mathbf{x}_n) + \epsilon_n$. Typically, this noise process might be specified as Gaussian: *i.e.* $p(\epsilon_n|\sigma^2) = \mathcal{N}(\epsilon_n|0, \sigma^2)$. While not always realistic,

this specification, along with the choice of linear predictor (1), facilitates a Bayesian treatment of the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$, since, for a Gaussian prior, the posterior and marginal likelihood are both similarly Gaussian.

An acknowledged limitation of the Gaussian noise model is that it is not *robust*, in that if the observation values are contaminated by *outliers*, the accuracy of the predictor $f(\mathbf{x})$ can be significantly compromised. The outliers may perhaps represent corrupted observations or be genuine samples from a heavy-tailed noise process. In such circumstances, one might utilise a more robust (correspondingly heavier-tailed) noise distribution, such as a zero-mean *Student-t*, conventionally defined as:

$$p(\epsilon_n|\nu,\sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}\sigma}\left\{1+\frac{1}{\nu}\left(\frac{\epsilon_n}{\sigma}\right)^2\right\}^{-(\nu+1)/2}, \tag{2}$$

where ν is the 'degrees of freedom' parameter and σ the 'scale' parameter [3]. However, the use of this distribution renders the Bayesian integration over $\boldsymbol{\theta}$ analytically intractable, and the main statistical interest in this model has focussed on stochastic approximations, *e.g.* [4, 5].

Alternatively, there has been some recent interest in *variational* deterministic approximation techniques for robust Bayesian modelling, using noise distributions that are *mixtures* of zero-mean Gaussians [6, 2]. In fact, this approach can be unified with the Student-t framework since that distribution may be realised as an *infinite* such mixture using the following *hierarchical* specification:

$$p(\epsilon_n|c,d) = \int_0^\infty p(\epsilon_n|\beta_n)p(\beta_n|c,d)\ d\beta_n, \tag{3}$$

where

$$p(\epsilon_n|\beta_n) = \mathcal{N}(\epsilon_n|0,\beta_n^{-1}), \tag{4}$$

$$p(\beta_n|c,d) = \text{Gamma}(\beta_n|c,d) = \frac{d^c}{\Gamma(c)}\beta_n^{c-1}\exp(-\beta_n d), \tag{5}$$

with $\Gamma(c)$ the 'gamma' function. In terms of the parameterisation (2), the equivalent distribution is obtained with $\nu = 2c$ and $\sigma = \sqrt{d/c}$. Effectively, equation (3) specifies that the noise model is a mixture (average) of an infinite number of Gaussians of varying precisions (inverse variances) β_n, with the mixture weight for a given β_n specified by the Gamma distribution $p(\beta_n|c,d)$.

This decomposition of the noise model into separate Gaussian and Gamma components allows convenient application of variational methods, and we describe in the next section how this leads to an effective Bayesian inference procedure. In practical terms, we variationally approximate the Bayesian integration over all model parameters, with the exception of c and d, for which we find point estimates. The power of this method is that it enables us to estimate posterior distributions over all model parameters, while simultaneously learning a flexible model of the noise process $p(\epsilon_n|c,d)$. Furthermore, as $c \to \infty$, this distribution tends to a Gaussian and so, as we subsequently illustrate, we can obtain improved interpolants when outliers are present without sacrificing the facility to treat the noise as Gaussian if supported by the data.

VARIATIONAL INFERENCE FOR THE HIERARCHICAL STUDENT-t NOISE MODEL

The Desired Bayesian Posterior

Given the observations $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$, we desire to compute the Bayesian posterior distribution over all unknowns by applying Bayes' rule:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\ p(\boldsymbol{\beta}|c, d)\ p(\boldsymbol{\theta}|\boldsymbol{\alpha})\ p(\boldsymbol{\alpha}|a, b)}{p(\mathbf{y})}, \tag{6}$$

where the likelihood term is given by:

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (2\pi)^{-N/2} \prod_{n=1}^{N} \beta_n^{1/2} \exp\left\{ -\frac{\beta_n}{2} \left[y_n - \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}_n) \right]^2 \right\}, \tag{7}$$

and the prior terms by:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \mathcal{N}(\theta_m|0, \alpha_m^{-1}), \tag{8}$$

$$p(\boldsymbol{\alpha}|a, b) = \prod_{m=1}^{M} \mathrm{Gamma}(\alpha_m|a, b), \text{ and} \tag{9}$$

$$p(\boldsymbol{\beta}|c, d) = \prod_{n=1}^{N} \mathrm{Gamma}(\beta_n|c, d), \text{ as given earlier.} \tag{10}$$

For the parameters, note that we have specified a Gaussian prior (8) such as utilised in "sparse Bayesian" models [7], where the prior parameter probabilities depend on independent hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$. These hyperparameters are in turn controlled by Gamma distributions parameterised by a and b, which may be fixed to very small values to obtain relatively flat (uninformative) hyperpriors over each α_m. The only other variables not treated probabilisitically are the key noise parameters, c and d, which will be discussed in more detail shortly.

Variational Approximation

Unfortunately, we cannot compute the posterior (6) analytically as the denominator $p(\mathbf{y})$ necessitates an intractable integration. Instead, here we adopt a variational approximation scheme as follows.

We note first that $\log p(\mathbf{y})$ can be expressed as the difference of two terms:

$$\log p(\mathbf{y}) \equiv \log p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \log p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}), \tag{11}$$

from which we write

$$\log p(\mathbf{y}) = \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \right\} - \log \left\{ \frac{p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\,\mathbf{y})}{Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \right\}, \tag{12}$$

where we have introduced an arbitrary 'approximating' distribution $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Integrating both sides of (12) with respect to $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ gives

$$\begin{aligned}\log p(\mathbf{y}) &= \int Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \right\} d\boldsymbol{\theta} d\boldsymbol{\alpha} d\boldsymbol{\beta} \\ &\quad - \int Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \log \left\{ \frac{p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y})}{Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \right\} d\boldsymbol{\theta} d\boldsymbol{\alpha} d\boldsymbol{\beta}, \\ &= \mathcal{L}\left[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right] + \mathrm{KL}\left[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) || \, p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y})\right], \end{aligned} \tag{13}$$

since $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is a distribution and integrates to one. The second term in (13) is the Kullback-Leibler divergence between the approximating distribution $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and the posterior $p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y})$ that we desire. Since $\mathrm{KL}[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) || \, p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y})] \geq 0$, it follows that $\mathcal{L}\left[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right]$ is a rigorous lower bound on $\log p(\mathbf{y})$. We can therefore obtain an approximation to the posterior indirectly by maximizing $\mathcal{L}\left[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right]$ with respect to $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, as this must simultaneously minimize the Kullback-Leibler divergence.

This then leaves the question of how to specify $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. While one can adopt some parameterised form, it has been shown [9] that if we simply assume that $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are *a posteriori* separable, such that $Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) Q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, then $\mathcal{L}\left[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right]$ is maximized by inspection:

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp \langle \log p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \rangle_{Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) Q_{\boldsymbol{\beta}}(\boldsymbol{\beta})}, \tag{14}$$

with symmetric expressions for $Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ and $Q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. The expectations in, and normalisations of, (14) are readily computed when the distributions of interest are appropriately conjugate and exponential[1], and these are given shortly. Note, however, that the solutions for $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *et al.* are mutually dependent, so in practice we must iteratively cycle through them, improving (raising) the lower bound with each such iteration.

The Q-distributions

The parameters in the interpolant.

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{15}$$

with

$$\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^{\mathrm{T}} \mathbf{B} \boldsymbol{\Phi} + \mathbf{A})^{-1} \quad \boldsymbol{\mu} = \boldsymbol{\Sigma} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{B} \mathbf{y}, \tag{16}$$

and we define $\boldsymbol{\Phi}_{nm} = \phi_m(\mathbf{x}_n)$, $\mathbf{A} = \mathrm{diag}\left(\langle \alpha_1 \rangle, \langle \alpha_2 \rangle, \ldots, \langle \alpha_M \rangle\right)$ and $\mathbf{B} = \mathrm{diag}\left(\langle \beta_1 \rangle, \langle \beta_2 \rangle, \ldots, \langle \beta_N \rangle\right)$.

[1] Further useful background on the use of variational techniques in this context, and details of an experimental software package, may be found in [1].

The hyperparameters.

$$Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \text{Gamma}(\alpha_m|\widetilde{a}, \widetilde{b}_m), \tag{17}$$

with

$$\widetilde{a} = a + \frac{1}{2} \qquad \widetilde{b}_m = b + \frac{\langle\theta_m^2\rangle}{2}. \tag{18}$$

The noise process.

$$Q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \prod_{n=1}^{N} \text{Gamma}(\beta_n|\widetilde{c}, \widetilde{d}_n), \tag{19}$$

with

$$\widetilde{c} = c + \frac{1}{2} \qquad \widetilde{d}_n = d + \frac{1}{2}\left(y_n^2 - 2y_n\boldsymbol{\phi}_n^{\mathrm{T}}\langle\boldsymbol{\theta}\rangle + \boldsymbol{\phi}_n^{\mathrm{T}}\langle\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\rangle\boldsymbol{\phi}_n\right). \tag{20}$$

Expectations following from and required to evaluate the above are:

$$\langle\boldsymbol{\theta}\rangle = \boldsymbol{\mu} \qquad \langle\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\rangle = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}, \tag{21}$$

$$\langle\alpha_m\rangle = \widetilde{a}/\widetilde{b}_m \qquad \langle\beta_n\rangle = \widetilde{c}/\widetilde{d}_n. \tag{22}$$

The Variational Lower Bound

From equations (13) and (7–10), the variational lower bound on $p(\mathbf{y})$ is:

$$\begin{aligned}\mathcal{L}\left[Q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right] =& \langle\log p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta})\rangle + \langle\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})\rangle + \langle\log p(\boldsymbol{\alpha}|a, b)\rangle + \langle\log p(\boldsymbol{\beta}|c, d)\rangle \\ &- \langle Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\rangle - \langle Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\rangle - \langle Q_{\boldsymbol{\beta}}(\boldsymbol{\beta})\rangle.\end{aligned} \tag{23}$$

where

$$\langle\log p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta})\rangle = \frac{1}{2}\sum_{n=1}^{N}\langle\log\beta_n\rangle - \langle\beta_n\rangle\left(y_n^2 - 2y_n\boldsymbol{\phi}_n^{\mathrm{T}}\langle\boldsymbol{\theta}\rangle + \boldsymbol{\phi}_n^{\mathrm{T}}\langle\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\rangle\boldsymbol{\phi}_n\right), \tag{24}$$

$$\langle\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})\rangle = \frac{1}{2}\sum_{m=1}^{M}\langle\log\alpha_m\rangle - \langle\alpha_m\rangle\langle\theta_m^2\rangle, \tag{25}$$

$$\langle\log p(\boldsymbol{\alpha}|a, b)\rangle = Ma\log b - M\log\Gamma(a) + (a-1)\sum_{m=1}^{M}\langle\log\alpha_m\rangle - b\sum_{m=1}^{M}\langle\alpha_m\rangle, \tag{26}$$

$$\langle\log p(\boldsymbol{\beta}|c, d)\rangle = Nc\log d - N\log\Gamma(c) + (c-1)\sum_{n=1}^{N}\langle\log\beta_n\rangle - d\sum_{n=1}^{N}\langle\beta_n\rangle, \tag{27}$$

$$\langle Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\rangle = -\frac{M}{2}\log(2\pi e) - \frac{1}{2}\log|\boldsymbol{\Sigma}|, \tag{28}$$

$$\langle Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\rangle = \sum_{m=1}^{M}\left[(\widetilde{a}-1)\langle\log\alpha_m\rangle - \widetilde{b}_m\langle\alpha_m\rangle + \widetilde{a}\log\widetilde{b}_m - \log\Gamma(\widetilde{a})\right], \tag{29}$$

$$\langle Q_{\boldsymbol{\beta}}(\boldsymbol{\beta})\rangle = \sum_{n=1}^{N}\left[(\widetilde{c}-1)\langle\log\beta_n\rangle - \widetilde{d}_n\langle\beta_n\rangle + \widetilde{c}\log\widetilde{d}_n - \log\Gamma(\widetilde{c})\right], \tag{30}$$

where we use the results $\langle\log\alpha_m\rangle = \psi(\widetilde{a}) - \log\widetilde{b}_m$ and $\langle\log\beta_n\rangle = \psi(\widetilde{c}) - \log\widetilde{d}_n$, with $\psi(\cdot)$ the 'psi' or 'digamma' function, defined as $\psi(x) = \partial/\partial x[\log\Gamma(x)]$.

Since the bound must always increase, it can be monitored during the update procedure as a check on the computations and on convergence.

The Noise Process

Ideally, we would prefer to specify prior distributions over c and d and include them in (6). However, there are no appropriate conjugate priors for both parameters compatible with the adopted variational framework. Instead, with the Q-distributions over all other parameters fixed, we maximise the bound $\mathcal{L}\left[Q(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})\right]$ with respect to c and d with the aim of increasing the marginal likelihood $p(\mathbf{y})$ (although this is not guaranteed). In (23), only the term (27) depends on c and d, and we obtain:

$$\frac{\partial\langle\log p(\boldsymbol{\beta}|c,d)\rangle}{\partial c} = N\log d - N\psi(c) + \sum_{n=1}^{N}\langle\log\beta_n\rangle, \tag{31}$$

$$\frac{\partial\langle\log p(\boldsymbol{\beta}|c,d)\rangle}{\partial d} = \frac{Nc}{d} - \sum_{n=1}^{N}\langle\beta_n\rangle. \tag{32}$$

Setting these gradients to zero does not lead to a joint closed-form solution, and we chose to perform a short scaled conjugate gradient (SCG) optimisation in conjunction with the Q-updates.

Estimation Procedure Summary

For our experiments, we chose a broad hyperprior with $a = b = 10^{-6}$. For the noise process, we chose $c = 0.04$ and $d = 0.01$, so as to specify a mean prior standard deviation for ϵ of $\sigma = 0.5$ (see Figure 2 for an illustration).

We then cycled through the Q-distribution updates, starting for convenience with $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Every 5 sets of updates for all distributions, we performed 10 cycles of SCG optimisation of c and d (this could be performed at every Q-update if desired, but is almost certainly wasteful of computation).

We chose to terminate when none of the changes at each update to $\langle\boldsymbol{\theta}\rangle$, $\log\langle\alpha_m\rangle$ or $\log\langle\beta_n\rangle$ was greater than some threshold, here 10^{-12}. At termination, quantities of interest are the mean interpolant, $f(\mathbf{x})$ computed with $\boldsymbol{\theta} = \langle\boldsymbol{\theta}\rangle$, and the inferred noise distribution $p(\epsilon|c,d)$ utilising the optimised values of c and d.

EXAMPLES

Synthetic Data

We first illustrate performance of the algorithm on univariate synthetic data generated from the function $\mathrm{sinc}(x) = (\sin x)/x$ with both additive Gaussian and Student-t noise. We fitted a sparse Bayesian interpolation model using both standard Gaussian and the presented variational Student-t noise models. The intention is to show that the variational procedure can recover the underlying generator in both cases. Note that in the figures which follow, converged *posterior mean* interpolants $\langle f(\mathbf{x})\rangle_{Q(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})}$ are shown.

Figures 1 and 2 show the results for $N = 100$ equally-spaced examples in $[-10, 10]$ with Gaussian noise of standard deviation $\sigma = 0.25$. 'Gaussian' basis functions, located on the data such that $\phi_m(x) = \exp\{-[(x - x_m)/2.0]^2\}$, were utilised.

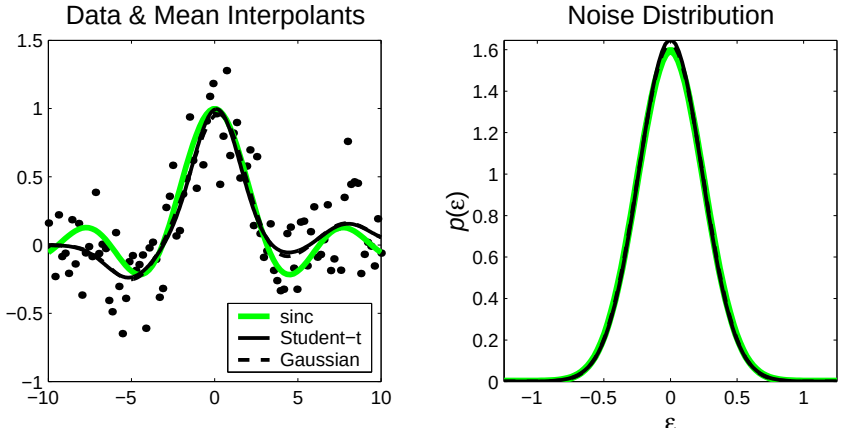

Figure 1: Left: interpolants for the two noise models. Right: the inferred noise distribution $p(\epsilon|\beta)$ for the Gaussian model, $p(\epsilon|c, d)$ for the Student-t. In this and following plots, 'truth' will be indicated in grey, a Gaussian-only model with a dashed trace, and the variational Student-t in a solid trace. Here, the Gaussian and Student-t results are practically coincident.

Figure 1 indicates that both models perform well in this case and, importantly, the adapted noise distribution $p(\epsilon|c, d)$ for the variational Student-t does appear Gaussian. Figure 2 (left) shows the inferred prior over the noise standard deviation, a re-scaling of $p(\beta|c, d)$, which should ideally be a δ-function at $\sigma = 0.25$. This plot shows a typical result: considering that the model parameters are being simultaneously estimated, the variational Student-t approach has arguably provided a very good approximation to the 'truth'. Figure 2 (right) shows the *mean* noise standard deviation estimated for multiple runs on the 'sinc' data with varying generative noise levels. The graph shows a reasonable fit to the generative noise level, though it cannot provide any further evidence of the 'Gaussianity' of $p(\epsilon|c, d)$.

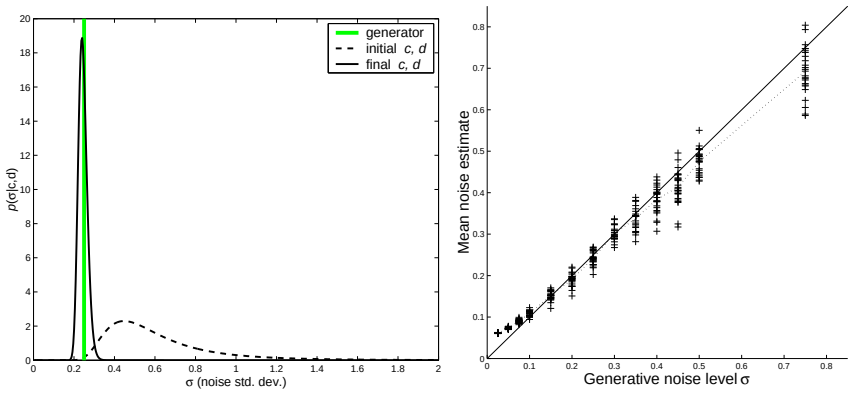

Figure 2: Left: the estimated prior over the noise level, showing also the initialisation. Right: the mean estimated noise for multiple runs with varying noise levels. The solid line shows the ideal case, and the dashed line shows a trace through the average of the mean estimates for each generative noise level.

Figures 3 and 4 show results for Student-t noise with $\nu = 4$ degrees of freedom and scale parameter $\sigma = 0.22$ (equivalent to a hierarchical formulation with $c = 2$ and $d = 0.1$). In this case, Figure 3 (left) shows that the variational Student-t model is qualitatively superior, in that it is less perturbed by outlying data (and so a better fit to the generator). Figure 3 (right) shows that we have successfully learned the change in character of the noise process, which is re-iterated by the plot of Figure 4 (left), equivalent to that of Figure 2 (left). In Figure 4 (right), the estimate of the noise distribution $p(\epsilon|c, d)$ is shown for several other values of c and d. It is notable, and we found this to be typical in other experiments, that as the noise process becomes more Gaussian, the estimate of c (and so the shape parameter ν) is considerably less accurate (although the estimate improves with more data).

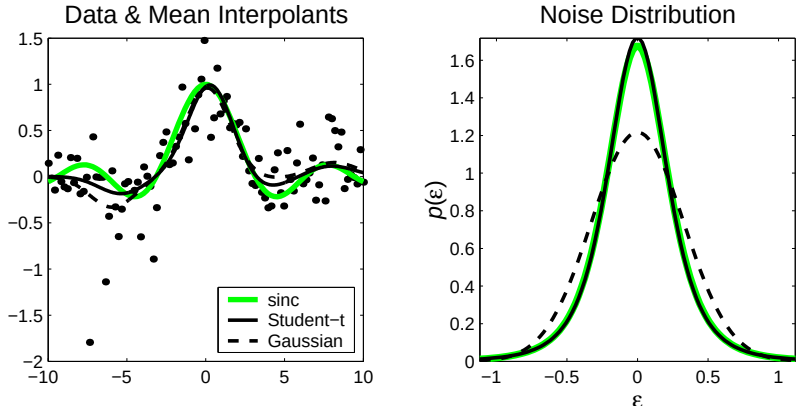

Figure 3: Left: interpolants for Student-t noise. Right: inferred noise distributions.

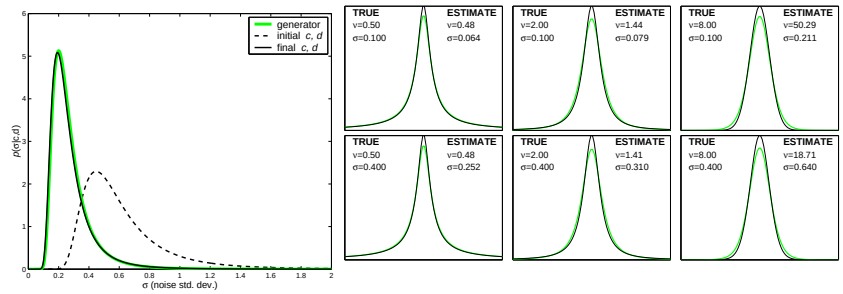

Figure 4: Left: the estimated prior over the noise standard deviation. Right: noise distributions, 'true' in gray, estimate in black, for other values of c and d. The shape parameter ν is increasing (becoming 'more Gaussian') left to right, while the upper/lower rows illustrate smaller/greater noise scales σ.

Accelerometer Data

We now illustrate the performance of the approach on a real signal where the noise process cannot be expected to be exactly Gaussian or Student-t. We consider the smoothing of an angular 'tilt' signal that is received via an AM radio link from a dual-accelerometer I.C. mounted on an 'electronic pen'. The data stream is inherently noisy, with a distribution that can be expected to deviate from Gaussian. One approach to smoothing out this noise is the use of a Bayesian interpolation model, but this is complicated by data contamination that occurs due to occasional interference on the radio link. It is these outlying values that are particularly problematic.

Figure 5 (a) illustrates one channel of data obtained by mounting the pen in a calibration rig set up to rotate the pen in a consistent circle. Although we don't know ground truth, the ideal output should be sinusoidal of constant amplitude and frequency, and plotting one channel against the second should give a perfect circle. Two outliers due to radio-link interference are evident in the portion of the data shown in Figure 5(a), and the Student-t model is seen to be less compromised by the outliers and more consistently sinusoidal.

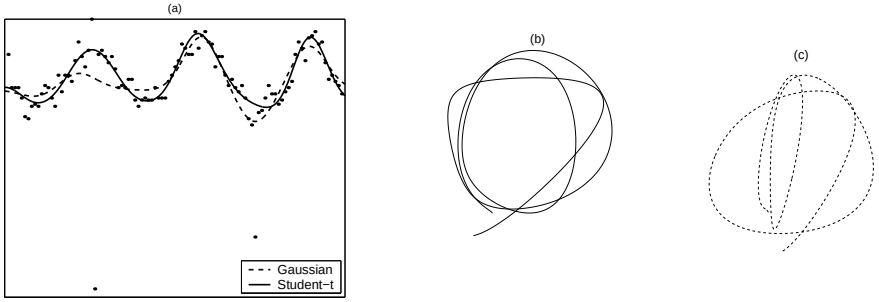

Figure 5: (a) Raw tilt data with Gaussian and Student-t interpolants. (b) XY plot using second channel, not shown in (a), for the Student-t interpolant. (c) XY plot for the Gaussian interpolant.

SUMMARY

Figures 1–4 illustrate that with even relatively limited data (100 observations here), the variational procedure is capable of recovering appropriate estimates of the underlying character of both Gaussian and Student-t noise processes, *at the same time as* performing an effective Bayesian estimation of the primary model parameters.

In practice of course, it is not expected that all significantly non-Gaussian noise processes will be exactly Student-t, as synthesized for Figures 3 and 4. However, even if the model is not an exact match, we would argue, and Figure 5 offers evidence to support this, that in many cases an adaptive Student-t model would offer superior results to a Gaussian for noise processes with significant tails or where outliers occur.

Finally, we do not exclude that for outlying data alone, mixture-based approaches [6, 2] might perform comparably well given appropriate tuning, and the algorithm presented here complements and extends those methods to offer a practical and effective Bayesian toolkit for robust interpolation.

REFERENCES

[1] C. M. Bishop and J. Winn, "Structured variational distributions in VIBES," in C. M. Bishop and B. J. Frey (eds.), **Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL, Jan 3-6**, 2003.

[2] A. Faul and M. E. Tipping, "A variational approach to robust regression," in G. Dorffner, H. Bischof and K. Hornik (eds.), **Proceedings of ICANN'01**, Springer, 2001, pp. 95–102.

[3] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, **Bayesian Data Analysis**, Chapman & Hall, 1995.

[4] J. F. Geweke, "Bayesian Treatment of the Independent Student-t Linear Model," **Journal of Applied Econometrics**, vol. 8, pp. S19–S40, 1993.

[5] R. M. Neal, "Monte Carlo implementation of Gaussian process models for Bayesian regression and classification," Techn. Report 9702, **Dept. of Statistics, University of Toronto**, 1997.

[6] W. Penny and S. J. Roberts, "Variational Bayes for non-Gaussian autoregressive models," in **Neural Networks for Signal Processing X**, IEEE, 2000, pp. 135–144.

[7] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," **Journal of Machine Learning Research**, vol. 1, pp. 211–244, 2001.

[8] V. N. Vapnik, S. E. Golowich and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing," in M. C. Mozer, M. I. Jordan and T. Petsche (eds.), **Advances in Neural Information Processing Systems 9**, MIT Press, 1997.

[9] S. Waterhouse, D. J. C. MacKay and T. Robinson, "Bayesian methods for mixtures of experts," in M. C. Mozer, D. S. Touretzky and M. E. Hasselmo (eds.), **Advances in Neural Information Processing Systems 8**, MIT Press, 1996.