A Variational EM Approach to Predicting Uncertainty in Supervised Learning

Markus Harva

Abstract— **In many applications of supervised learning, the conditional average of the target variables is not sufficient for prediction. The dependencies between the explanatory variables and the target variables can be much more complex calling for modelling the full conditional probability density. The ubiquitous problem with such methods is overfitting since due to the flexibility of the model the likelihood of any datapoint can be made arbitrarily large. In this paper a method for predicting uncertainty by modelling the conditional density is presented based on conditioning the scale parameter of the noise process on the explanatory variables. The regularisation problems are solved by learning the model using variational EM. Results with synthetic data show that the approach works well and experiments with real-world environmental data are promising.**

I. Introduction

When neural networks are used in solving regression problems, the cost function being minimised is usually the sum-of-squares. This is known to lead to modelling of the conditional average of the target variable which, in some cases, yields perfectly satisfactory results. However, sometimes the dependencies between the explanatory variables and the target variables are much more complex calling for methods beyond the sum-of-squares.

Williams [1] suggests a model where the target variables are modelled as a conditional multivariate Gaussian and a neural network is used to model *both* the mean and the covariance of the distribution. An even more general method is presented by Bishop [2], where the target variable has a mixture-of-Gaussians distribution whose means, variances and mixture weights are all conditioned on the explanatory variables and are being modelled using the multi-layer perceptron network (MLP). This combines the universal function approximation property of the MLP with the universal density function approximation property of the mixture-of-Gaussians to a theoretically sound framework.

The problem with the above mentioned works and indeed with any approach where flexible modelling of the target density is allowed, is overfitting. It is already a problem with standard neural nets and it becomes a lot worse when modelling densities. This is due to the fact that the model is able to place an infinite probability density over a datapoint. Heavy ad hoc regularisation is usually needed to circumvent this problem.

In this paper a method for predicting uncertainty is presented based on modelling the scale parameter of the target variable's distribution. Both the location and the scale parameter are conditioned on the input variables and predicted by MLPs. The regularisation problems are avoided using Bayesian machinery to learn the model; namely the variational expectation-maximisation algorithm is employed. Since a part of the parameters have distributional estimates, problems with infinite densities do not arise.

The model and the learning algorithm bear similarity to the ones in [3], where variance modelling was used in the blind source separation setting to capture higher order dependencies in the data. However, in that paper the learning tasks were unsupervised and linear, whereas here they are supervised but nonlinear.

In the next section the suggested model will be presented in detail. Section III concentrates on the learning algorithm giving the necessary ingredients needed in the variational EM framework. In Section IV, results both with synthetic and real-world environmental data are reported. Finally the paper ends with discussion.

II. Model

Since we are considering a supervised problem, we have a set of explanatory variables and a set of target variables. Without loss of generality we concentrate on the case of one single target variable. Let us denote the explanatory variables as $\mathbf{x}_t$, which is a column vector having N elements. The subindex t signifies that it is the t:th measurement from the range $1, ..., T$. The whole $N \times T$ matrix of measurements will be denoted as $\mathbf{X}$. The target variable will be denoted as y_t and the collection of all its measurements as $\mathbf{Y}$.

The model consists of two multi-layer perceptron networks, one for the data itself, and the other for the log-variance of the noise. Both of the MLPs have one hidden layer with H and K hidden nodes respectively. The probabilistic model is fully defined by the following set of equations:

$$\begin{aligned}
y_t &\sim \mathcal{N}\left(\mathbf{a}_y^T \tanh(\mathbf{C}_y \mathbf{x}_t + \mathbf{d}_y) + b_y, e^{-u_t}\right) \\
a_{yi} &\sim \mathcal{N}\left(0, 10^2\right) \\
b_y &\sim \mathcal{N}\left(0, 10^2\right) \\
u_t &\sim \mathcal{N}\left(\mathbf{a}_u^T \tanh(\mathbf{C}_u \mathbf{x}_t + \mathbf{d}_u) + b_u, \tau_u^{-1}\right) \\
a_{ui} &\sim \mathcal{N}\left(0, 10^2\right) \\
b_u &\sim \mathcal{N}\left(0, 10^2\right) \\
\tau_u &\sim \mathcal{G}\left(1, 10^{-4}\right)
\end{aligned}$$

Above, $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the Gaussian distribution with mean μ and variance σ^2 and $\mathcal{G}\left(\alpha, \beta\right)$ is the Gamma distribution with shape α and inverse scale β. The first layer

Markus Harva is with the Adaptive Informatics Research Centre, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland (phone: +358 9 451 3287; fax: +358 9 451 3277; email: markus.harva@hut.fi).

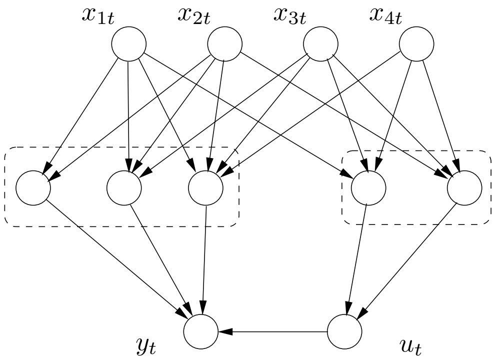

Fig. 1. Schematic illustration of the model.

parameters ($\mathbf{C}_y$, $\mathbf{d}_y$, $\mathbf{C}_u$ and $\mathbf{d}_u$) do not have priors and in the estimation process only ML estimates are obtained for them. Figure 1 shows a schematic illustration of the model. For that particular case the parameters are: $N = 4$, $H = 3$ and $K = 2$.

The introduction of the intermediate variable u_t is done for two reasons. Firstly, it gives robustness to the model by generating heavier tails to the target distribution and secondly, it makes the model estimation easier.

A. Positively Constrained Observations with Model for Variance Only

In addition to the model described above, another slightly different model will be considered for nonnegative data. Naturally, in this case the Gaussian noise process is subobtimal. Any positively supported distribution from the exponential family of distributions could be considered in place of the ordinary Gaussian. Here, the rectified Gaussian distribution is used, but exponential or rectified Laplacian (which collapses to exponential in the zero mode case) could be considered as well. This changes the prior of the target variable to

$$y_t \sim \mathcal{N}^R\left(0, e^{-u_t}\right)$$

where $\mathcal{N}^R$ denotes the rectified Gaussian distribution i.e. a Gaussian whose negative axis has been rectified and the right axis scaled appropriately. For now, only the zero mode case is considered.

III. Learning Algorithm

In this Section the learning algorithm for the proposed model is derived.

Since the conditional density is being modelled in a rather flexible manner, some regularisation is needed to avoid more or less catastrophic cases of overfitting due to infinite densities. The variational EM (V-EM) algorithm (see e.g. [4]) is one viable alternative, since a part of the parameters have distributional estimates. This helps a lot in avoiding the pitfalls of maximum likelihood estimation.

A. Variational EM

Consider a model for some data $\mathbf{Y}$ having two sets of parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Let us further assume that the model is specified as the joint density $p(\mathbf{Y}, \boldsymbol{\theta}_1 | \boldsymbol{\theta}_2) = p(\mathbf{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, p(\boldsymbol{\theta}_1)$. If the standard EM algorithm [5] was used to estimate the model it would alternate between the following two steps:

E-step: Set $Q^{(i+1)}(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2^{(i)}) = p(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2^{(i)})$

M-step: Find $\boldsymbol{\theta}_2^{(i+1)}$ such that it maximises

$$\int Q^{(i+1)}(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2^{(i)}) \log p(\mathbf{Y}, \boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(i+1)})\, d\boldsymbol{\theta}_1 \quad (1)$$

For many models for which the EM-algorithm would otherwise be preferable, the computation of the Q distribution in the E-step is not tractable. In variational EM, the first step is replaced by *fitting* a simpler distribution to $p(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2^{(i)})$ by functional minimisation of the Kullback-Leibler divergence

$$D_{KL}(Q||p) = \int Q(\boldsymbol{\theta}_1) \log \frac{Q(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2^{(i)})}\, d\boldsymbol{\theta}_1 \,. \quad (2)$$

This to be tractable, Q needs to have a suitably factorial form. The variational EM algorithm can be seen as a compromise between maximum likelihood and variational Bayesian learning [6] (VB for short). VB is the extreme case of V-EM having distributions for all the parameters i.e. the parameter set $\boldsymbol{\theta}_2$ is empty and there is no M-step. One of the biggest benefits of VB is that a lower bound for the marginal likelihood of the data is obtained, meaning that an approximation for the posterior probabilities of competing models can be computed. On the other hand, the learning algorithm can get quite complex, especially for nonlinear models, calling for precomputed lookup-tables or numerical integration [6], [7]. In V-EM, Eq. (2) yields a lower bound for $\log p(\mathbf{Y}|\boldsymbol{\theta}_2)$, but since it still depends on $\boldsymbol{\theta}_2$, trustworthy model comparison using that lower bound is not possible.

For the model of this paper the following division between parameters is made. The second layer of parameters in the model have distributional estimates and the first layer point estimates only. Most importantly, the conditional log-variance u_t and the parameters $\mathbf{a}_y$ and b_y belong to the first group, but since distributional estimates can be readily computed for the rest of the second layer variables, that is done. This arrangement seems to regularise the problem enough as to avoid overfitting, while retaining reasonable computational complexity of the learning algorithm.

B. E-step

The form of the approximate distribution Q is chosen to be fully factorial:

$$Q(\boldsymbol{\theta}_1) = \prod_{t=1}^{T} Q(u_t) \times \prod_{i=1}^{H} Q(a_{yi}) \times Q(b_y)$$
$$\times \prod_{i=1}^{K} Q(a_{ui}) \times Q(b_u) \times Q(\tau_u)$$

In the E-step the factors $Q(\cdot)$ in the approximation are updated one by one given the other approximations and the rest of the parameters $\boldsymbol{\theta}_2$.

Except for $Q(u_t)$, conjugate update rules are obtained. In the following, both the distributional form of the approximation as well as the values for its parameters are given.

The update rule for the weights of the linear mapping is

$$\begin{aligned} Q(a_{yi}) &= \mathcal{N}\left(a_{yi}|\mu, \sigma^2\right) \\ \sigma^2 &= \Big(10^{-2} + \sum_{t=1}^{T} \langle e^{u_t} \rangle f_{yit}^2\Big)^{-1} \\ \mu &= \sigma^2 \sum_{t=1}^{T} \langle e^{u_t} \rangle f_{yit} \Big(y_t - \sum_{k \neq i} \langle a_{yk} \rangle f_{ykt} - \langle b_y \rangle \Big) , \end{aligned}$$

where $f_{yit} := \tanh(\sum_{j=1}^{N} c_{yij} x_{jt} + d_i)$ and $\langle \cdot \rangle$ denotes the expectation computed over Q. The bias is updated as follows

$$\begin{aligned} Q(b_y) &= \mathcal{N}\left(b_y|\mu, \sigma^2\right) \\ \sigma^2 &= \Big(10^{-2} + \sum_{t=1}^{T} \langle e^{u_t} \rangle\Big)^{-1} \\ \mu &= \sigma^2 \sum_{t=1}^{T} \langle e^{u_t} \rangle \Big(y_t - \sum_{i=1}^{H} \langle a_{yi} \rangle f_{yit} \Big) \end{aligned}$$

The parameters $\mathbf{a}_u$ and b_u are handled in a very similar manner. The update rule for the precision parameter τ_u is

$$\begin{aligned} Q(\tau_u) &= \mathcal{G}\left(\tau_u|\alpha, \beta\right) \\ \alpha &= \frac{T}{2} + 1 \\ \beta &= \frac{1}{2} \sum_{t=1}^{T} \Big\langle \Big(u_t - \sum_{i=1}^{K} a_{ui} f_{uit} - b_u \Big)^2 \Big\rangle + 10^{-4} \end{aligned}$$

Updating the approximation for the log-variance u_t is somewhat more complicated. No easy conjugate update rule is now available. Following the scheme that was used in [3], the approximation is a priori fixed to a Gaussian i.e. $Q(u_t) = \mathcal{N}\left(u_t|m, v\right)$. In this case the KL-divergence (2) yields the following expression to be minimised w.r.t. m and v

$$F(m, v) := Mm + V(m^2 + v) + Ee^{m+v/2} - \frac{1}{2} \log v \,, \quad (3)$$

where

$$\begin{aligned} M &= -\frac{1}{2} - \langle \tau_u \rangle \left\langle \mathbf{a}_u^T \tanh(\mathbf{C}_u \mathbf{x}_t + \mathbf{d}_u) + b_u \right\rangle \\ V &= \frac{1}{2} \langle \tau_u \rangle \\ E &= \frac{1}{2} \left\langle \left(y_t - \mathbf{a}_y^T \tanh(\mathbf{C}_y \mathbf{x}_t + \mathbf{d}_y) - b_y \right)^2 \right\rangle \end{aligned}$$

In [3], the potential in Eq. (3) was minimised using a hybrid scheme of alternating fixed-point and Newton iteration. Here, it is minimised using Newton-iteration jointly for m and $\log v$. The convergence is very fast, requiring only a couple of iterations.

C. M-step

With regards to the parameters $\mathbf{C}_y$ and $\mathbf{d}_y$ the integral in Eq. (1) yields the following function to be maximised in the M-step

$$\begin{aligned} F(\mathbf{C}_y, \mathbf{d}_y) := -\sum_{t=1}^{T} \langle e^{u_t} \rangle \Bigg[\Big(\sum_{i=1}^{H} \langle a_{yi} \rangle f_{yit} + \langle b_y \rangle - y_t \Big)^2 \\ + \sum_{i=1}^{H} \mathrm{var}(a_{yi}) f_{yit}^2 \Bigg] . \end{aligned}$$

The maximisation is done by computing the gradient ∇F and then performing a line search by solving

$$\max_{\alpha} F\big(\mathbf{C}_y^{(i)} + \alpha \nabla_{\mathbf{C}_y} F \,,\ \mathbf{d}_y^{(i)} + \alpha \nabla_{\mathbf{d}_y} F\big) .$$

The (approximately) optimal step size is found by increasing (decreasing) α as long as the value of F increases. The M-step for the parameters $\mathbf{C}_u$ and $\mathbf{d}_u$ is handled in a similar manner.

D. The Predictive Density

Once the model has been learnt, it remains to compute the predictive pdf. Let us denote the parameters having priors but excluding the log-variances u_t simply as $\boldsymbol{\theta}$ i.e. $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \setminus \{u_t | t = 1, \ldots, T\}$. Now the predictive pdf is formally obtained from the integral

$$p(y_t|\mathbf{x}_t, \mathbf{X}, \mathbf{Y}) = \int p(y_t|u_t, \mathbf{x}_t, \boldsymbol{\theta}) p(u_t|\mathbf{x}_t, \boldsymbol{\theta}) Q(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \, du_t \,,$$

where $\mathbf{X}$ and $\mathbf{Y}$ denote the training data using of which the model has been learnt. The distribution of the parameters $\boldsymbol{\theta}$ can be collapsed to a delta distribution at $\langle \boldsymbol{\theta} \rangle$ without making too big of an error. However, it is not wise to neglect the integration over u_t, but it is not tractable either. Hence the following approximation is made

$$p(y_t|\mathbf{x}_t, \mathbf{X}, \mathbf{Y}) = \sum_{l=1}^{L} \pi_l \mathcal{N}\left(y_t | m_{yt}, e^{-u_{tl}}\right) \,,$$

where $m_{yt} := \left\langle \mathbf{a}_y^T \right\rangle \tanh(\mathbf{C}_y \mathbf{x}_t + \mathbf{d}_y) + \langle b_y \rangle$ and the points u_{tl} are regularly sampled from the interval $[m_{ut} - 2\sigma, m_{ut} + 2\sigma]$ in which $m_{ut} = \left\langle \mathbf{a}_u^T \right\rangle \tanh(\mathbf{C}_u \mathbf{x}_t + \mathbf{d}_u) + \langle b_u \rangle$ and $\sigma = \langle \tau_u \rangle^{-1/2}$. The mixture probabilities π_l are selected as to reflect the original density:

$$\pi_k = \frac{\mathcal{N}\left(u_{tk} | m_{ut}, \sigma^2\right)}{\sum_{l=1}^{L} \mathcal{N}\left(u_{tl} | m_{ut}, \sigma^2\right)} .$$

If the log-variance can be accurately modelled such that the precision τ_u is high, the approach is nearly equivalent to collapsing $p(u_t|\mathbf{x}_t, \boldsymbol{\theta})$ to a delta distribution at m_{ut}.

The above methodology applies to the rectified Gaussian case as well.

IV. Results

In this section results both with synthetic and real-world environmental data are presented. Both of the datasets are a part of the predictive uncertainty competition held in WCCI'06. The method proposed in this paper won the competition.

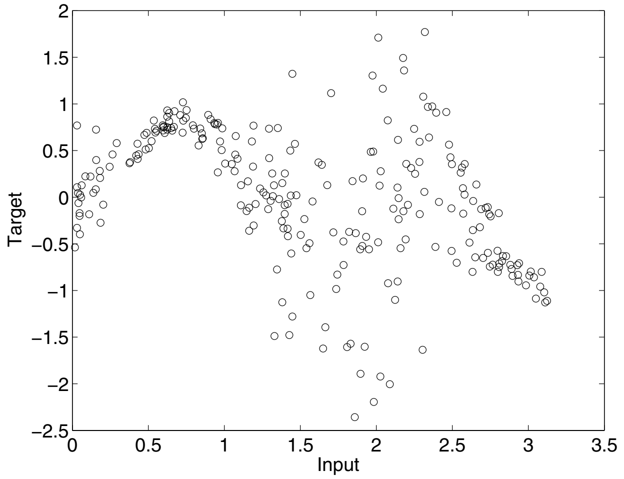

Fig. 2. Synthetic data. The input variable is plotted against the target variable.

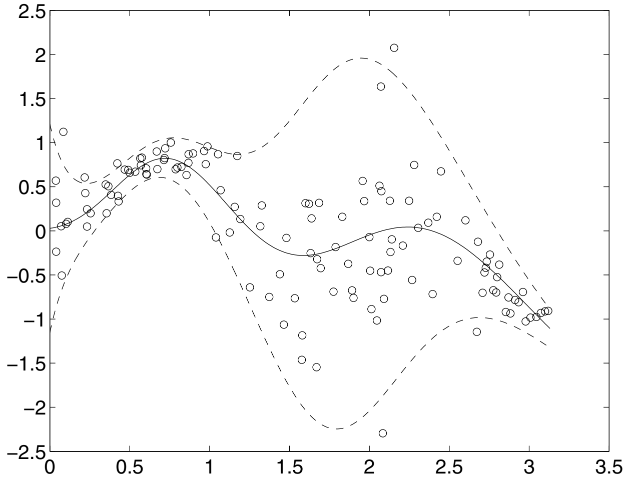

Fig. 3. The predictive pdf for the synthetic dataset. The solid line shows the mean and the dashed lines bound the area containing 95% of the predictive probability mass.

TABLE I
RESULTS (NLPD)

Model	Data Synthetic	SO_2
True	0.35	
Proposed	0.39	4.37
Regression	0.94	5.15
Histogram	1.23	4.50

A. Synthetic Data

The synthetic data consists of one explanatory variable controlling both the mean and the variance of the observations in some unknown nonlinear manner. The noise process is Gaussian, so the method presented in this paper should perform fairly well with this dataset.

The input variable versus the target variable is plotted in Figure 2. From there it is clear that there are substantial fluctuations in the variance of the noise process.

The competition organisers have split the data into training, validation, and test sets for which the targets are available only for the first two. Fifty models were learnt using the training data by iterating the learning scheme for 5000 iterations. Since point estimates are used for some of the parameters, the marginal likelihood bound obtained from Eq. (2) cannot reliably be used for model selection and hence the validation data was used to this end. The predictive performance of the models was measured by computing the average negative log-likelihood (NLPD for now on) for the validation dataset. The predictive pdf for the best model in this sense is shown in Figure 3 along with the validation datapoints, which were not used in the learning. The solid line shows the mean of the predictive density and the dashed lines show the 95% confidence interval.

Table I shows the NLPD values computed for the test set (used neither in learning nor validation). The values for the true model as well as for two baseline approaches are shown also. The Regression method stands for standard MLP learnt using back-propagation with early stopping. One hundred nets were learnt in this manner with random initialisations for the weights and the best net in the sense of validation NLPD was selected. The Histogram method was to compute the unconditional empirical histogram for the training targets, which was then used as the predictive density. Clearly the proposed model has been able to learn the structure of the data to a large degree not loosing much to the true underlying model and beating the baseline approaches with a clear margin.

B. Sulfur Dioxide Concentration Data

In this real-world dataset, the target variable is the sulfur dioxide (SO_2) concentration in an urban environment. The 27 input variables consist of the same SO_2 concentration as well as other meteorological conditions measured 24 hours earlier. Four of the input variables are plotted against the target variable in Figure 4. It is evident from the figure that this dataset is very atypical when it comes to regression tasks, since no evident correlations (linear or nonlinear) can be seen between the input and output variables. The same applies for the rest of the variables. Looking at the variables x_2 and x_8 only, it is questionable whether there are any dependencies whatsoever in the data. From the scatter plots of x_{19} and x_{25} some dependencies can be visually recognised, however. For example, the target variable y obtains clearly larger values when the input variable x_{25} obtains values between -2 and -1 than between 1 and 2.

With this data, the alternative formulation of the model was used with the target variable having heteroskedastic rectified Gaussian prior. To illustrate that something can indeed be learnt from the data, first only the variable x_{25} was used in the learning. The learning scheme was similar to that with the synthetic data. The model with the best validation data performance was chosen again. The predictive pdf for that model is shown in Figure 5. The solid line shows the threshold below which 95% of the probability mass of

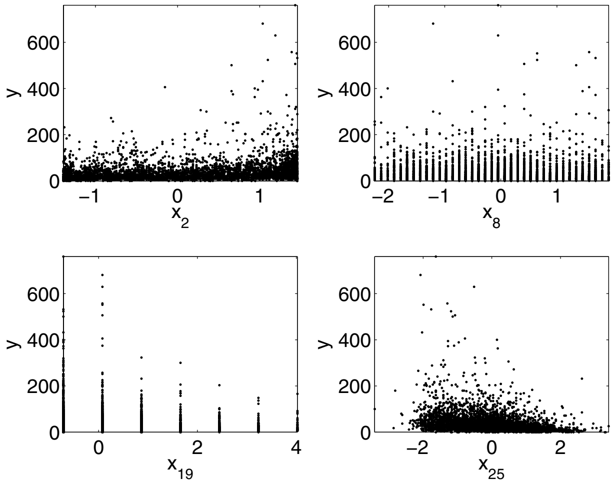

Fig. 4. SO_2 data. Four of the 27 input variables plotted against the target variable.

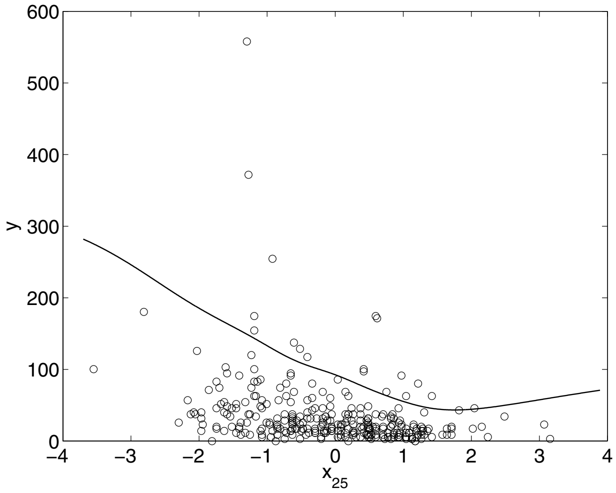

Fig. 5. The predictive pdf for one of the variables in the SO_2 dataset. 95% of the probability mass lies below the solid line.

the predictive pdf falls and the circles show (some of) the validation-set datapoints. Some conclusions about the SO_2 levels can already be made using this one input variable only.

The runs were repeated with all the input variables included. The results with the test set for the proposed and the baseline methods are shown in Table I. Now the learning task is so unsuitable for standard MLP fitting that it actually loses to the histogram approach, which doesn't use the inputs at all! The performance of the proposed approach is well above the baselines although it is probable that the margin could be made larger still by experimenting with different priors on the target variable.

V. Discussion

In this paper, a novel method for predicting uncertainty in the supervised learning setting was presented. The key idea was to explicitly model the scale parameter of the noise process, currently limited to either Gaussian or rectified Gaussian. When modelling means and variances jointly, point estimation methods, such as maximum likelihood or maximum a posteriori, can lead to catastrophic results, due to severe overfitting. This is because the likelihood can be made infinite by decreasing the variance towards smaller and smaller values. The necessary regularisation ingredient in the learning algorithm of this paper was to use the variational EM algorithm. There a part of the parameters have distributional estimates which overcomes many of the overfitting problems. By not putting distributions over all variables, the learning algorithm remains computationally almost as simple as if point estimation was used.

The results with the synthetic dataset were very good, mostly due to the fact that the used model very well matches the true one. With the real-world environmental dataset, the performance was rather good as well, but could be made better yet. It is clear, that there is a lot to be gained by choosing the most accurate prior for the target variable. The zero-mode rectified Gaussian used in the experiments is probably not the optimal choice. The obvious way to improve upon this would be to relax the zero-mode assumption. This however poses some serious technical difficulties, since the E-step of the algorithm would be much more complicated. Other rectified distributions could be appropriate too such as the rectified Laplacian having heavier tail than the Gaussian counterpart.

Acknowledgements

The author wishes to thank Antti Honkela and Tapani Raiko for useful comments and discussion.

References

[1] P. M. Williams, "Using neural networks to model conditional multivariate densities," *Neural Computation*, vol. 8, no. 4, pp. 843–854, 1996.

[2] C. M. Bishop, "Mixture density networks," Neural computing research group, Aston University, Tech. Rep. NCRG/4288, 1994.

[3] H. Valpola, M. Harva, and J. Karhunen, "Hierarchical models of variance sources," *Signal Processing*, vol. 84, no. 2, pp. 267–282, 2004.

[4] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA, USA: The MIT Press, 1999, pp. 355–368.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[6] G. E. Hinton and D. van Camp, "Keeping neural networks simple by minimizing the description length of the weights," in *Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, Santa Cruz, CA, USA, 1993, pp. 5–13.

[7] D. Barber and C. Bishop, "Ensemble learning for multi-layer networks," in *Advances in Neural Information Processing Systems 10*, M. Jordan, M. Kearns, and S. Solla, Eds. Cambridge, MA, USA: The MIT Press, 1998, pp. 395–401.