Accurate and Calibrated Parametric Model for Variance Estimation

E. Camporeale*
Centrum Wiskunde & Informatica
Amsterdam
The Netherlands
e.camporeale@cwi.nl

Abstract

In this paper we focus on the problem of assigning uncertainties to single-point predictions. We introduce a cost function that encodes the trade-off between accuracy and reliability (calibration) in probabilistic forecast. We derive analytic formula for the case of forecasts of continuous scalar variables expressed in terms of Gaussian distributions. The new Accuracy-Reliability cost function is used to estimate the input-dependent variance, given a black-box mean function, by solving a two-objective optimization problem. The simple philosophy behind this strategy is that predictions based on the estimated variances should not only be accurate, but also reliable (i.e. statistical consistent with observations). Conversely, strategies based on the minimization of classical cost functions, such as the negative log probability density, cannot simultaneously enforce both accuracy and reliability. We show several examples both with synthetic data, where the underlying hidden noise can accurately be recovered, and with large real-world datasets.

1 Introduction

There is a growing consensus, across many fields and applications, that forecasts should have a probabilistic nature (Gneiting & Katzfuss, 2014). This is particularly true in decision-making scenarios where cost-loss analyses are designed to take into account the uncertainties associated to a given forecast (Murphy, 1977; Owens et al., 2014). In these contexts, calibration is as important as accuracy. Calibration, also known as reliability (for instance, in the meteorological literature), is the requirement that the probabilities should give an estimate of the expected frequencies of the event occurring, that is a statistical consistence between predictions and observations (Johnson & Bowler, 2009). Unfortunately, it is often the case that well established predictive models are completeley deterministic and thus provide single-point estimates only. For example, in engineering and applied physics, models often rely on computer simulations. A typical strategy to assign confidence intervals to deterministic predictions is to perform ensemble forecasting, that is to repeat the same simulation with slightly different initial or boundary conditions (Gneiting et al., 2005; Leutbecher & Palmer, 2008). However, this is rather expensive and it often requires a trade-off between computational cost and accuracy of the model, especially when there is a need for real-time predictions.
In this paper we focus on the problem of assigning uncertainties to single-point predictions, with a particular emphasis on the requirement of calibration. We restrict our attention on predictive models that output a scalar continuous variable, and whose uncertainties are in general input-dependent. For the sake of simplicity, and for its widespread use, we assume that the probabilistic forecast that we want to generate is in the form of a Gaussian distribution. Hence, the problem can be cast in terms of the estimation of the variance associated to a normal distribution centered around forecasted values

*www.cwi.nl/∼camporea

Preprint. Work in progress.

provided by a model. Because the variance can in general be input-dependent, this problem is very similar to a particular case of heteroskedastic regression. Indeed, in the regression literature it is common to assume that noisy targets are generated as $y = f(\mathbf{x}) + \varepsilon$, and here we restrict to the special case of zero-mean Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2(\mathbf{x}))$, where the heteroskedasticity is evident in the $\mathbf{x}$ dependence of the variance. However, it is important to emphasize that whilst in the classical heteroskedastic regression problem, one is interested in learning simultaneously the mean function $f(\mathbf{x})$ and the variance $\sigma^2(\mathbf{x})$, here we assume that the mean function is provided by a black-box simulation that cannot easily be improved, hence the whole attention is focused on the variance estimation. This is realistic in several applied fields, where decades of work have resulted in very accurate physics-based models, that however suffer the drawback of being completely deterministic. In the Machine Learning community, elegant and practical ways of deriving uncertainties based on non-parametric Bayesian methods are well established, either based on Bayesian neural networks (MacKay, 1992; Neal, 2012; Hernández-Lobato & Adams, 2015), deep learning (Gal & Ghahramani, 2016), or Gaussian Processes (GPs) (Rasmussen & Williams, 2006).
In this paper we embrace a complementary approach, focusing on parametric models, and consequently on the minimization of a cost function. A parametric model based on deep ensembles has recently been proposed in Lakshminarayanan et al. (2017), building on the original idea of Weigend & Nix (1994) of designing a neural network that outputs simultaneously mean and variance of a Gaussian distribution, by minimizing a proper score, namely the negative log likelihood of the predictive distribution. Lakshminarayanan et al. (2017) point out the importance of calibration of probabilistic models, even though in their work calibration is not explicitly enforced. As we mentioned, in this paper we decouple the problem of learning mean function and variance, focusing solely on the latter. Also, it is important to keep in mind that we aim at estimating the variance using a single mean function, and not an ensemble.

1.1 Summary of Contributions and Novelty

Our method is very general and does not depend on any particular choice for the black-box model that predicts the output targets (which indeed is not even required; all that is needed are the errors between predictions and real targets). The philosophy is to introduce a cost function which encodes a trade-off between the accuracy and the reliability of a probabilistic forecast. Assessing the goodness of a forecast through proper scores, such as the Negative Log Probability Density, or the Continuous Rank Probability Score, is a common practice in many applications, like weather predictions (Matheson & Winkler, 1976; Bröcker & Smith, 2007). Also, the notion that a probabilistic forecast should be well calibrated, or statistically consistent with observations, has been discussed at length in the atmospheric science literature (Murphy & Winkler, 1992; Toth et al., 2003). However, the basic idea that these two metrics (accuracy and reliability) can be combined to estimate the empirical variance from a sample of observations, and possibly to reconstruct the underlying noise as a function of the inputs has never been proposed. Moreover, as we will discuss, the two metrics are competing, when interpreted as functions of the variance only. Hence, this gives rise to a two-objective optimization problem, where one is interested in achieving a good trade-off between these two properties.
Our main contributions are the introduction of the Reliability Score (RS), that measures the discrepancy between empirical and ideal calibration, and the Accuracy-Reliability (AR) cost function. We show that for a Gaussian distribution the RS has a simple analytical formula. We also discuss the Continuous Rank Probability Score, that we argue has better numerical property than the more standard Negative Log Probability Density.

2 Loss functions for Accuracy and Reliability

The standard way of estimating the empirical variance of a Gaussian distribution is by maximizing its likelihood with respect to a set of observations. In practice, the loss function based on the negative logarithm of the probability density (NLPD) is used:

$$\text{NLPD}_i(\varepsilon_i, \sigma_i) = \frac{\log \sigma_i^2}{2} + \frac{\varepsilon_i^2}{2\sigma_i^2} + \frac{\log 2\pi}{2}, \tag{1}$$

where we define $\varepsilon_i = y_i^o - \mu_i$ as the error between the i-th observation y_i^o and prediction μ_i in a training set of size N. An important consideration is that NLPD (as any other loss function) does not enforce, per se, a correct model calibration. Calibration is the property of a probabilistic model that

measures its statistical consistence with observations. For forecasts of discrete events, it measures if an event predicted with probability p occurs, on average, with frequency p. This concept can be extended to forecasts of a continuous scalar quantity by examining the so-called reliability diagram (Anderson, 1996; Hamill, 1997, 2001). Note that in this paper we use the terms calibration and reliability interchangeably. A reliability diagram is produced in the following way. One collects the values of the probability predicted at the observed points, that is $P(y \leq y^o)$, which for a Gaussian distribution we denote with $\Phi_i = \frac{1}{2}(\mathrm{erf}(\eta_i) + 1)$, with $\eta_i = \varepsilon_i/(\sqrt{2}\sigma_i)$ being the relative errors. Its empirical cumulative distribution, defined as $C(y) = \frac{1}{N}\sum_{i=1}^{N} H(y - \Phi_i)$ (H is the Heaviside function), provides the reliability diagram, with the obvious interpretation of observed frequency as a function of the predicted probability. A perfect calibration shows in the reliability diagram as a straight diagonal line.

The motivating argument of this work is that two models with identical NLPD can have remarkably different reliability diagrams. We show an example in Figure 1 : 1000 data points have been generated as $\mathcal{N}(0, \sigma(x)^2)$, with $x \in [0, 1]$ and $\sigma(x) = x + \frac{1}{2}$, as in the synthetic dataset proposed in Goldberg et al. (1998). A model completely consistent with the data generation mechanism (i.e. with zero mean and variance σ^2) produces the blue line in the reliability diagram in the left panel, that is almost perfect calibration. However, one can generate a second model with a modified variance $\tilde{\sigma}^2$ such that $\mathrm{NLPD}_i(\varepsilon_i, \tilde{\sigma}_i) = \mathrm{NLPD}_i(\varepsilon_i, \sigma_i)$, that is

$$\frac{\log \tilde{\sigma}_i^2}{2} + \frac{\varepsilon_i^2}{2\tilde{\sigma}_i^2} = \frac{\log \sigma_i^2}{2} + \frac{\varepsilon_i^2}{2\sigma_i^2} \tag{2}$$

Eq. (2) always produces a solution with $\tilde{\sigma}_i \neq \sigma_i$, as long as $\sigma_i^2 \neq \varepsilon_i^2$ (the global minimum of NLPD, for fixed ε_i). The red line in the left panel of Figure 1 has been derived from such a modified model $\mathcal{N}(0, \tilde{\sigma}^2)$, which is obviously mis-calibrated. For this example NLPD=0.4 (equal for both cases). As a complementary argument, we show in the right panel of Figure 1 the reliability diagram of several models, with decreasing values of NLPD. One can appreciate that progressively decreasing NLPD results in a worse and worse calibration. These models have been generated again starting from the perfectly calibrated synthetic model, progressively shifting the values assigned to σ_i^2, towards the global minimum $\sigma_i^2 = \varepsilon_i^2$ (hence decreasing NLPD). Thus, minimizing a traditional cost function such as NLPD does not necessarily implies to achieve a well-calibrated model. Of course, we are not suggesting that any model generated by means of minimizing NLPD is inevitably mis-calibrated. However, unless explicitly enforced, calibration will be a by-product of other properties. For instance, neural networks and Gaussian processes enforce a certain degree of smoothness of the solution, and a smoother variance (as function of the inputs) will often tend to result in a better calibrated model (e.g., a constant variance associated to normally distributed errors always yields a well-calibrated, but probably inaccurate, model).

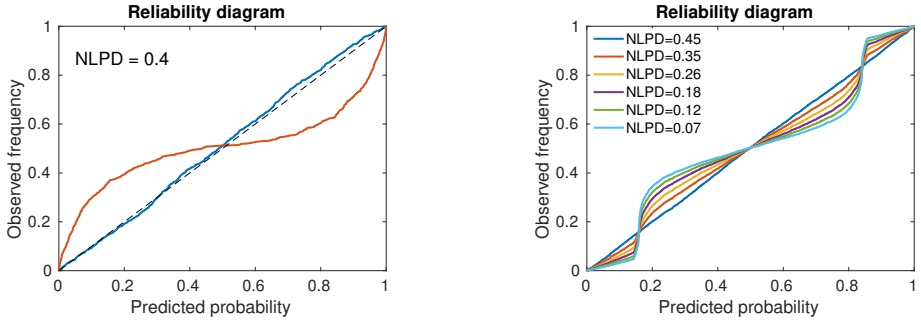

Figure 1: Left: example of two models with identical value of NLPD=0.4, and different reliability diagram. Right: Example of several models for which the NLPD decreases (from NLPD=0.45 for the blue line to NLPD=0.07 for the cyan line), at the expense of reliability. See text for details of how the synthetic data has been generated.

2.1 Continuous Rank Probability Score

In this work we propose to use the Continuous Rank Probability Score (CRPS), in lieu of the NLPD. CRPS is a generalization of the well-known Brier score (Wilks, 2011), used to assess the probabilistic

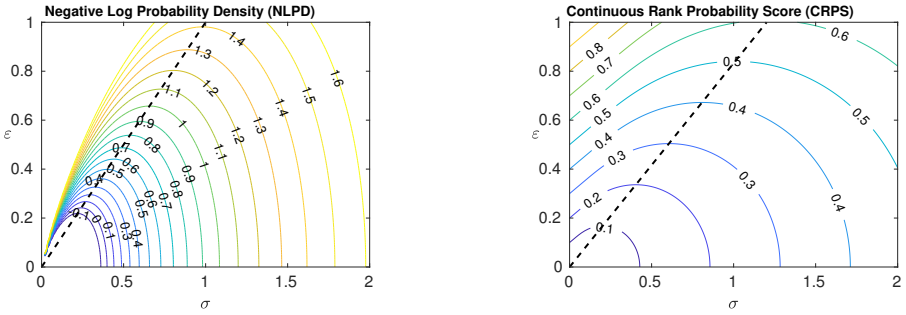

Figure 2: Isoline of constant NLPD (left) and CRPS (right) as a function of standard deviation σ, and error ε. The black dashed line indicates the minimum, as function of ε.

forecast of continuous scalar variables, when the forecast is given in terms of a probability density function, or its cumulative distribution. CRPS is defined as

$$\text{CRPS} = \int_{-\infty}^{\infty} [C(y) - H(y - y^o)]^2 \, dy \tag{3}$$

where $C(y)$ is the cumulative distribution (cdf) of the forecast, $H(y)$ is the Heaviside function, and y^o is the true (observed) value of the forecasted variable. For Gaussian distributions, the forecast is simply given by the mean value μ and the variance σ^2, and in this case the CRPS can be calculated analytically (Gneiting et al., 2005) as

$$\text{CRPS}(\mu, \sigma, y^o) = \sigma \left[\frac{y^o - \mu}{\sigma} \operatorname{erf}\left(\frac{y^o - \mu}{\sqrt{2}\sigma} \right) + \sqrt{\frac{2}{\pi}} \exp\left(-\frac{(y^o - \mu)^2}{2\sigma^2} \right) - \frac{1}{\sqrt{\pi}} \right] \tag{4}$$

Several interesting properties of the CRPS have been studied in the literature. Notably, its decomposition into reliability and uncertainty has been shown in Hersbach (2000). Even though a small value of CRPS necessarily implies a small value of NLPD, there are several reasons for preferring CRPS to NLPD. They are both negatively oriented, but CRPS is equal to zero for a perfect forecast with no uncertainty (deterministic). Indeed, the CRPS has the same unit as the variable of interest, and it collapses to the Absolute Error $|y^o - \mu|$ for $\sigma \to 0$, that is when the forecast becomes deterministic. On the other hand, the limit $\sigma \to 0$ is problematic for NLPD, being finite only for $y^o = \mu$. Figure 2 shows a graphical comparison between NLPD (left panel) and CRPS (right panel). Different curves show the isolines for the two scores, as a function of the error ε (vertical axis) and the standard deviation σ (horizontal axis). The black dashed line indicates the minimum value of the score, for a fixed value of ε. Because we are approaching the problem of variance estimation by assigning an empirical variance to single-point black-box predictions, it makes sense to minimize a score as a function of σ only, for a fixed value of the error $\varepsilon = y^0 - \mu$. By differentiating Eq.(4) with respect to σ, one obtains

$$\frac{d\text{CRPS}}{d\sigma} = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\varepsilon^2}{2\sigma^2} \right) - \frac{1}{\sqrt{\pi}} \tag{5}$$

and the minimizer is found to be $\sigma^2_{\text{min, CRPS}} = \varepsilon^2 / \log 2$. Note that the minimizer for NLPD is $\sigma^2_{\text{min, NLPD}} = \varepsilon^2$.

As it is evident from Figure 2, CRPS penalizes under- and over-confident predictions in a much more symmetric way than NLPD. Both scores are defined for a single instance of forecast and observation, hence they are usually averaged over an ensemble of predictions, to obtain the score relative to a given model: $\overline{\text{CRPS}} = \sum_k \text{CRPS}(\mu_k, \sigma_k, y^o_k)$.

2.2 Reliability Score for Gaussian forecast

Contrary to the CRPS, that is defined for a single pair of forecast-observation, it is clear that the reliability can only be defined for a large enough ensemble of such pairs, being a statistical property of a model. Here, we introduce the reliability score for normally distributed forecasts. In this case, we

expect the relative errors η calculated over a sample of N predictions-observations to have a standard normal distribution with cdf $\Phi(\eta) = \frac{1}{2}(\text{erf}(\eta) + 1)$. Hence we define the Reliability Score (RS) as:

$$\text{RS} = \int_{-\infty}^{\infty} \left[\Phi(\eta) - C(\eta)\right]^2 d\eta \tag{6}$$

where $C(\eta)$ is the empirical cumulative distribution of the relative errors η, that is

$$C(y) = \frac{1}{N}\sum_{i=1}^{N} H(y - \eta_i) \tag{7}$$

with $\eta_i = (y_i^o - \mu_i)/(\sqrt{2}\sigma_i)$. Obviously, RS measures the divergence of the empirical distribution of relative errors η from a standard normal distribution. From now on we will use the convention that the set $\{\eta_1, \eta_2, \ldots \eta_N\}$ is sorted ($\eta_i \leq \eta_{i+1}$). Obviously this does not imply that μ_i or σ_i are sorted as well. Interestingly, the integral in Eq. (6) can be calculated analytically, via expansion into a telescopic series, yielding:

$$\text{RS} = \sum_{i=1}^{N}\left[\frac{\eta_i}{N}\left(\text{erf}(\eta_i) + 1\right) - \frac{\eta_i}{N^2}(2i-1) + \frac{\exp(-\eta_i^2)}{\sqrt{\pi}N}\right] - \frac{1}{2}\sqrt{\frac{2}{\pi}} \tag{8}$$

Differentiating the i-th term of the above summation, RS_i, with respect to σ_i (for fixed ε_i), one obtains

$$\frac{d\text{RS}_i}{d\sigma_i} = \frac{\eta_i}{N\sigma_i}\left(\frac{2i-1}{N} - \text{erf}(\eta_i) - 1\right) \tag{9}$$

Hence, RS_i is minimized at the value $\sigma_{\text{min}}^{\text{RS}}$ that satisfies

$$\text{erf}(\eta_i) = \text{erf}\left(\frac{\varepsilon_i}{\sqrt{2}\sigma_{\text{min}}^{\text{RS}}}\right) = \frac{2i-1}{N} - 1 \tag{10}$$

This could have been trivially derived by realizing that the distribution of η_i that minimizes RS is the one such that the values of the empirical cumulative distribution $C(\eta)$ are uniform in the interval $[0, 1]$.

3 The Accuracy-Reliability cost function

The Accuracy-Reliability (AR) cost function introduced here follows from the simple principle that the variances σ_i^2 estimated from an ensemble of errors ε_i should result in a model that is both accurate (with respect to the CRPS score), and reliable (with respect to the RS score). Clearly, this gives rise to a two-objective optimization problem. It is trivial to verify that CRPS and RS cannot simultaneously attain their minimum value (as was evident from Figure 1). Indeed, by minimizing the former, $\eta_i = \frac{1}{2}\sqrt{\log 4}$ for any i. Obviously, a constant η_i cannot result in a minimum for RS, according to Eq. (10). Note that any cost function that is minimized (for constant ε) by a variance σ^2 that is linear in ε^2 suffer this problem. Also, notice that trying to minimize RS as a function of σ_i (for fixed errors ε_i) results in an ill-posed problem, because RS is solely expressed in terms of the relative errors η. Hence, there is no unique solution for the variances that minimizes RS. Hence, RS can be more appropriately thought of as a regularization term in the Accuracy-Reliability cost function. The simplest strategy to deal with multi-objective optimization problems is to scalarize the cost function, which we define here as

$$\text{AR} = \beta \cdot \overline{\text{CRPS}} + (1 - \beta)\text{RS}. \tag{11}$$

We choose the scaling factor β as

$$\beta = \text{RS}_{min}/(\overline{\text{CRPS}}_{min} + \text{RS}_{min}). \tag{12}$$

The minimum of $\overline{\text{CRPS}}$ is $\overline{\text{CRPS}}_{min} = \frac{\sqrt{\log 4}}{2N}\sum_{i=1}^{N}\varepsilon_i$, which is simply the mean of the errors, rescaled by a constant. The minimum of RS follows from Eqs. (8) and (10):

$$\text{RS}_{min} = \frac{1}{\sqrt{\pi}N}\sum_{i=1}^{N}\exp\left(-\left[\text{erf}^{-1}\left(\frac{2i-1}{N} - 1\right)\right]^2\right) - \frac{1}{2}\sqrt{\frac{2}{\pi}} \tag{13}$$

Notice that RS_{min} is only a function of the size of the sample N, and it converges to zero for $N \to \infty$. The heuristic choice in Eq. (12) is justified by the fact that the two scores might have different orders of magnitude, and therefore we rescale them in such a way that they are comparable in our cost function (11) (this is another reason to choose CRPS over NLPD, since the latter is not bounded from below). We believe this to be a sensible choice, although there might be applications where one would like to weigh the two scores differently. In future work, we will explore the possibility of optimizing β in a principled way, for instance constraining the difference between empirical and ideal reliability score to be within limits given by the dataset size N. Finally, in our practical implementation, we neglect the last constant term in the definition (8) so that, for sufficiently large N, $\text{RS}_{min} \simeq \frac{1}{2}\sqrt{\frac{2}{\pi}} \simeq 0.4$

4 Results

In summary, we want to estimate the input-dependent values of the empirical variances σ_i^2 associated to a sample of N observations for which we know the errors ε_i. We do so by solving an optimization problem in which the set of estimated σ_i minimizes the AR cost function defined in Eq. (11). This newly introduced cost function has a straightforward interpretation as the trade-off between accuracy and reliability, which are two essential but conflicting properties of probabilistic models. In practice, because we want to generate a model that is able to predict σ^2 as a function of the inputs $\mathbf{x}$ on any point of a domain, we introduce a structure that enforces smoothness of the unknown variance, in the form of a neural network. In the following we show some experiments on toy problems and on multidimensional real dataset to demonstrate the easiness, robustness and accuracy of the method.
For simplicity, we choose a single neural network architecture, that we use for all the tests. We use a network with 2 hidden layers, respectively with 50 and 10 neurons. The activation functions are rectified linear (ReLU) and a symmetric saturating linear function, respectively. The output is given in terms of $\log \sigma$, to enforce positivity of σ^2. For all experiments, the datasets are randomly divided into training (33%), validation (33%) and test (34%) sets. All the reported metrics are calculated on the test set only. The network is trained using a standard BFGQ quasi-Newton algorithm, and the iterations are forcefully stopped when the loss function does not decrease for 10 successive iterations on the validation set. The only inputs needed are the points x_i and the corresponding errors ε_i. Finally, in order to avoid local minima due to the random initialization of the neural network weights, we train five independent networks and choose the one that yields the smallest cost function.

4.1 Toy problems

Results are first shown on toy problems, for illustrative purposes. We choose the Yuan & Whaba (**Y**) dataset, widely used in the literature: $x \in [0, 1]$, $f(x) = 2(\exp(-30(x - 0.25)^2) + \sin(\pi x^2)) - 2$, $\sigma(x) = \exp(\sin(2\pi x))$ (Yuan & Wahba, 2004). An example of 100 points sampled from the **Y** dataset is shown in Figure 3 (circles, left-top panel), along with the true mean function $f(x)$ (red). Since in our method we assume that a mean function is provided, in this toy problem we use the result of a standard (homoskedastic) Gaussian Process regression as $f(x)$, shown in blue. In the right-top panel of Figure 3 the red line denotes the true standard deviation $\sigma(x)$ used to generate the data. The black line indicates the values of the estimated σ averaged over 100 independent runs, and the gray areas represent one and two standard deviations from the mean. A certain spread in the results is due to different training sets (in each run 100 points are sampled independently) and to random initialization. Next, we move to a multidimensional (**5D**) synthetic dataset: $\mathbf{x} \in [0, 1]^5$, $f(\mathbf{x}) = 0$, $\sigma(\mathbf{x}) = 0.45(\cos(\pi + \sum_{i=1}^{5} 5x_i) + 1.2)$ (Genz, 1984). The left-bottom panel in Figure 3 shows the distribution of σ, which ranges in the interval $[0.09, 0.99]$. The **5D** dataset is obviously more challenging, hence we use 10,000 points to train the model (note that this results in less points per dimension, compared to the one-dimensional test). It is impractical to compare graphically the real and estimated $\sigma(\mathbf{x})$ in the 5-dimensional domain. Instead, in the right-bottom panel we show the probability density of the real versus predicted values of the standard deviation. Values are normalized such that the maximum value in the colormap for any value of predicted σ is equal to one (i.e. along vertical lines). The red line shows a perfect prediction. The colormap has been generated by 10^7 points, while the model has been trained with 10,000 points only. For this case, we have used an exact mean function (equal to zero), in order to focus exclusively on the estimation of the variance. This is an excellent result for a very challenging task, given the sparsity of the training set, that shows the robustness of the method.

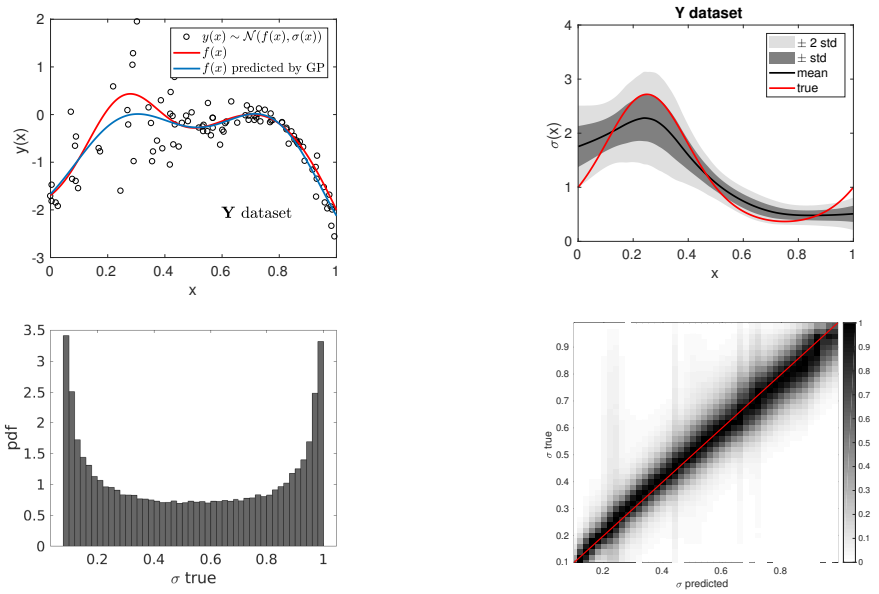

Figure 3: Top-left: 100 points sampled from the **G** dataset (circles). The blue line shows the true mean function $f(x)$, while the red one is the one predicted by the GP model. Top-right: True value of the standard deviation σ (red line) and mean value obtained averaging over 100 independent runs (black line). The gray shaded areas denote the confidence interval of one and two standard deviations calculated from the same ensemble of runs. Bottom-left: Distribution of true values of standard deviation σ for the **5D** dataset. Bottom-right: Probability density of the prediction versus real values of σ for the **5D** dataset. The red line denotes perfect prediction. The densities are normalized to have maximum value along each column equal to one. 10,000,000 samples have been used to generate the plot (with a training set of 10,000 points).

4.2 Real-World dataset

We have tested our method on the same datasets used in (Hernández-Lobato & Adams, 2015). The results reported in Table 1 are averaged over 20 independent runs. For each run, we first train a standard neural network to provide the mean function $f(\mathbf{x})$, by minimizing the mean square errors with respect to the targets. We then compare two different models: one in which the variance is estimated by minimizing NLPD (a standard method), and the other where we minimize the new AR cost function. For both models we use the same initialization and architecture of the neural network, in order to have a fair comparison. We then calculate the NLPD and the CRPS. In order to evaluate the calibration, we derive the reliability diagram (in the way described in section 2), and we compute the maximum distance to the optimal reliability (straight diagonal line). This is denoted, in Table 1, as Cal. err. (in percentage). The results obtained by using the AR cost function are always better calibrated, at the expenses of a modest increase of NLPD (not even in all cases), and essentially the same value of CRPS, with respect to the results obtained by minimizing NLPD.

5 Discussion and future work

We have presented a simple parametric model for estimating the variance of probabilistic forecasts. We assume that the data is distributed as $\mathcal{N}(f(\mathbf{x}), \sigma(\mathbf{x})^2)$, and that an approximation of the mean function $f(\mathbf{x})$ is available (the details of the model that approximates the mean function are not important). In order to generate the variance $\sigma(\mathbf{x})^2$, we propose to minimize the Accuracy-Reliability (AR) cost function, which depends only on σ, on the errors ε, and on the size of the training set N. We have shown that the classical method of minimizing the Negative Log Probability Density (NLPD) does not guarantees that the result will be well-calibrated. Indeed, we have discussed how accuracy and reliability are two conflicting metrics for a probabilistic forecast and how the latter can serve as a regularization term for the former. We have shown that by using the new AR cost function,

Table 1: Comparison between NLPD and AR on several multidimensional dataset.

Method (cost function) Score			NLPD NLPD	AR NLPD	NLPD CRPS	AR CRPS	NLPD Cal. err. (%)	AR Cal. err. (%)
Dataset	Size	Dim.						
Boston Housing	506	13	0.64	0.90	0.27	0.27	10.5	8.3
Concrete	1,030	8	0.55	0.62	0.26	0.26	6.0	5.3
Energy	768	8	-0.31	-0.23	0.13	0.13	9.6	7.7
Kin8nm	8,192	8	0.31	0.31	0.2	0.2	2.8	2.0
Naval propulsion	11,934	15	-1.52	-1.58	0.06	0.06	6.6	4.4
Power plant	9,568	4	0.04	0.04	0.15	0.15	3.1	2.6
Protein	45,730	9	1.11	1.18	0.43	0.42	7.3	7.4
Wine	1,599	11	1.24	1.29	0.47	0.47	11.9	10.6
Yacht	308	6	-0.44	-0.25	0.12	0.12	16.2	12.0
Year Prediction MSD	515,345	90	0.91	0.93	0.36	0.35	5.3	3.1

one is able to accurately discover the hidden noise function. Several tests for synthetic and real-world (large) datasets have been shown.

An important point to notice is that by setting the problem as an optimization for the Accuracy-Reliability cost function, the result for σ will be optimal, for the given approximation of the mean function $f(\mathbf{x})$. In other words, the method will inherently attempt to correct any inaccuracy in $f(\mathbf{x})$ by assigning larger variances. Hence, the agreement between predicted and true values of the standard deviation σ presented in Figure 3 must be understood within the limits of the approximation of the mean function (that in that example was provided by a Gaussian Process regression).

By decoupling the prediction of the mean function from the estimation of the variance, this method is not very expensive, and it is suitable for large datasets. Moreover, for the same reason this method is very appealing in all applications where the mean function is necessarily computed via an expensive black-box, such as computer simulations, for which the de-facto standard of uncertainty quantification is based on running a large (time-consuming and expensive) ensemble. Finally, the formulation is well suited for high-dimensional problems, since the cost function is calculated point-wise for any instance of prediction and observation.
Although very simple and highly efficient the method is still fully parametric, and hence it bears the usual drawback of possibly dealing with a large number of choices for the model selection. An interesting future direction will be to incorporate the Accuracy-Reliability cost function in a non-parametric Bayesian method for heteroskedastic regression.

References

Anderson, Jeffrey L. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7):1518–1530, 1996.

Bröcker, Jochen and Smith, Leonard A. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Genz, Alan. Testing multidimensional integration routines. In *Proc. Of International Conference on Tools, Methods and Languages for Scientific and Engineering Computation*, pp. 81–94, New York, NY, USA, 1984. Elsevier North-Holland, Inc. ISBN 0-444-87570-0.

Gneiting, Tilmann and Katzfuss, Matthias. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

Gneiting, Tilmann, Raftery, Adrian E, Westveld III, Anton H, and Goldman, Tom. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

Goldberg, Paul W, Williams, Christopher KI, and Bishop, Christopher M. Regression with input-dependent noise: A gaussian process treatment. In *Advances in neural information processing systems*, pp. 493–499, 1998.

Hamill, Thomas M. Reliability diagrams for multicategory probabilistic forecasts. *Weather and forecasting*, 12(4):736–741, 1997.

Hamill, Thomas M. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

Hernández-Lobato, José Miguel and Adams, Ryan. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pp. 1861–1869, 2015.

Hersbach, Hans. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

Johnson, Christine and Bowler, Neill. On the reliability and calibration of ensemble forecasts. *Monthly Weather Review*, 137(5):1717–1720, 2009.

Lakshminarayanan, Balaji, Pritzel, Alexander, and Blundell, Charles. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6405–6416, 2017.

Leutbecher, Martin and Palmer, Tim N. Ensemble forecasting. *Journal of Computational Physics*, 227(7):3515–3539, 2008.

MacKay, David JC. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Matheson, James E and Winkler, Robert L. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

Murphy, Allan H. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105(7):803–816, 1977.

Murphy, Allan H and Winkler, Robert L. Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455, 1992.

Neal, Radford M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Owens, Mathew J, Horbury, TS, Wicks, RT, McGregor, SL, Savani, NP, and Xiong, M. Ensemble downscaling in coupled solar wind-magnetosphere modeling for space weather forecasting. *Space Weather*, 12(6):395–405, 2014.

Rasmussen, Carl Edward and Williams, Christopher KI. *Gaussian process for machine learning*. MIT press, 2006.

Toth, Zoltan, Talagrand, Olivier, Candille, Guillem, and Zhu, Yuejian. Probability and ensemble forecasts, 2003.

Weigend, Andreas S and Nix, David A. Predictions with confidence intervals (local error bars). In *Proceedings of the international conference on neural information processing*, pp. 847–852, 1994.

Wilks, Daniel S. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

Yuan, Ming and Wahba, Grace. Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & probability letters*, 69(1):11–20, 2004.