# A Nonparametric Approach To Ensemble Forecasting

Eugene Y. Chen
Adobe Systems Incorporated
San Jose, California, USA
euchen@adobe.com

Xiaojing Dong
Santa Clara University
Santa Clara, California, USA
xdong1@scu.edu

Zhiyu Wang
Adobe Systems Incorporated
San Jose, California, USA
zhiwang@adobe.com

Zhenyu Yan
Adobe Systems Incorporated
San Jose, California, USA
wyan@adobe.com

## ABSTRACT

Ensemble forecasting has seen wide applications in social and physical sciences. Conventional methods are parametric and require the user to be knowledgeable about the types of error distributions involved. In this paper, we develop a nonparametric method that allows us to combine forecasting models with general error distributions. We applied the method to the *Wikipedia Web Traffic Time Series Forecasting* dataset, which is public and can be downloaded from Kaggle (https://www.kaggle.com/c/web-traffic-time-series-forecasting). We compare the proposed method to two of the most popular ensemble methods, Ensemble Model Output Statistics (EMOS) and Bayesian Model Averaging (BMA). We show that the proposed method yields more accurate forecasts for the page views of low traffic Wikipedia articles.

## CCS CONCEPTS

• **Applied computing → Forecasting**; • **Computing methodologies → Ensemble methods**; *Kernel methods*;

**ACM Reference Format:**
Eugene Y. Chen, Xiaojing Dong, Zhiyu Wang, and Zhenyu Yan. 2018. A Nonparametric Approach To Ensemble Forecasting. In *Proceedings of the 4th SIGKDD Workshop on Mining and Learning from Time Series (MiLeTS18).* ACM, New York, NY, USA, 8 pages.

## 1 INTRODUCTION

Forecasting plays a crucial role in planning and logistics. It is common in practice that more than one forecasting models were developed for the prediction of the quantity of interest. Bates and Granger [4] were among the first who pointed out that combining multiple forecasts with proper weights could yield a result that is superior (in the sense of accuracy and stability) to each component forecast. Since the publication of reference 3, the field of forecast combination has received much attention. Various methods have been developed and studied. For example, Regression-based methods [9, 13–15], Bayesian methods [5, 6, 17]; more recently, Artificial Neural Network (ANN)-based methods [2, 12]. We refer the readers to a few excellent reviews for an introduction to the field [3, 7, 8, 19]. In the literature, the science of combining forecasts is sometimes referred to as "ensemble forecasting" [13, 17], reminiscent of the ensemble methods in the field of Machine Learning. We will use the two terminologies interchangeably in this paper.

The existing methods of forecast combination can be encapsulated by the following formula:

$$F_T = G(\mathbf{w}, f_{1T}, f_{2T}, \ldots, f_{NT}) \tag{1}$$

where $f_{1T}, f_{2T}, \ldots, f_{NT}$ are the forecasts *for* time $T$ from the $N$ component forecasters in the ensemble and $\mathbf{w}$ is a vector of arbitrary length (it is in general of length $N$, but can be of any length in ANN-based methods). The general procedure to obtain $F_T$ can be summarized as follows:

(1) Historical actuals $\{a_t \mid \forall t \in \mathbb{N}^*_{<T}\}$ and historical forecasts $\{f_{kt} \mid \forall k \in \mathbb{N}^*_{\leq N}, \forall t \in \mathbb{N}^*_{<T}\}$ is collected.
(2) A *training scheme* is applied to determine $\mathbf{w}$.
(3) An ensemble forecast $F_T$ is generated with $\mathbf{w}$ and all of the component forecasts that is made *for* time $T$ ($\{f_{kT} \mid \forall k \in \mathbb{N}^*_{\leq N}\}$) through the use of equation 1.

We note in most cases (such as in Regression-based methods and in Bayesian methods), equation 1 is reduced to

$$F_T = \sum_{k=1}^{N} w_k f_{kT} \tag{2}$$

where $f_{kT}$, the forecast from the $k^{\text{th}}$ forecaster, can be either a real number [15] or a probabilistic distribution [17].

In general, a training scheme for $\mathbf{w}$ should not be taken as an one-size-fits-all solution for an ensemble that the user first encounters. Its effectiveness depends on the characteristics of the component forecasters, their interrelations, as well as the nature of the dataset. For example, Ordinary Least Squared-based Ensemble methods work well when the residuals are uncorrelated, having expectation zero and equal variances (Gauss-Markov Theorem) but could perform less-than-ideal otherwise. In recent implementations of parametric ensemble models (e.g., R packages for Bayesian Model Averaging (BMA) [18] and Ensemble Model Output Statistics (EMOS) [21]), users are allowed to specify the family of error distribution. However, the users are still faced with the difficulty of choosing the correct error distribution *a priori*. It is thus desirable to develop a method that does not require such knowledge.

In this paper, we develop a nonparametric method to fulfill the aforementioned purpose. We provide a general description based

*MiLeTS18, August 2018, London, United Kingdom*

on two-dimensional Gaussian error in section 2, followed by a simulation-based study in section 3. The method is applied to the Wikipedia Web Traffic dataset in section 4. Concluding remarks are provided in section 5.

## 2 DESCRIPTION OF THE PROPOSED NONPARAMETRIC METHOD

We introduce the relevant terminologies (item 1 and item 2) and assumption (item 3):

(1) The Data Generation Process (DGP) is a random process which follows a (in general time-dependent) PDF at each time-period.
(2) At any time-period, each forecaster in the ensemble has its own (random) Forecast Generation Process (FGP) that is shaped by the DGP and the forecaster's own skill/disposition. A forecast is a realization of the FGP.
(3) Given $N$ forecasters, the *differences* between the actual (i.e., a realization of DGP) and (a total of N) forecasts follow an $N$-dimensional Joint-PDF that does not change over time.

We note that item 3 is a much-relaxed assumption in comparison to that of the parametric Ensemble methods. No assumption about the error distribution is made other than that it does not evolve with time. Hence, the proposed method can be generalized to cases where the error distribution is either unknown or has no explicit mathematical expression.

The basic idea of the proposed method can be separated into three steps:

(1) Summarize the *past* errors of the base forecasters into a numerical $N$-dimensional Joint-PDF (which will be referred to as a Joint-error PDF from now on) in the Ensemble training stage. One of the means to achieve this goal is through the use of Kernel Density Estimation (KDE).
(2) In the prediction stage, when a simultaneous set of *future* forecasts are given, the Joint-error PDF is used to generate a likelihood function of the (yet unrealized) actual.
(3) Methods such as Maximum Likelihood Estimation (MLE) or minimization over a loss functional [11] are then applied to transform the likelihood function into a point forecast.

We will illustrate and explore the idea through a series of concrete examples throughout the paper. We note that the use of KDE is not a necessity when circumstances permit, as will be explored in section 4.

For simplicity, we consider the case where the DGP is a sinusoidal function of time without uncertainty, i.e., $a_t = 10\sin(\tau(t))$ ($\tau(t)$ is a linear function that maps $t \in \mathbb{N}^*$ to the equally-spaced sequence of $\tau \in \mathbb{R}$). This case is the limit where the PDF of the DGP is a Dirac delta function whose location evolves sinusoidally with time. For the ease of illustration, we set the ensemble to consist of only two forecasters whose errors are Gaussian and correlated. Specifically,

$$\begin{aligned} \epsilon_{1t} &\equiv f_{1t} - a_t \\ \epsilon_{2t} &\equiv f_{2t} - a_t \end{aligned} \tag{3}$$

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \sim N(\mu, \Sigma)$$

we refer the readers to appendix A for the parameters (i.e., $\mu$, $\Sigma$) adopted and the details of numerical results.

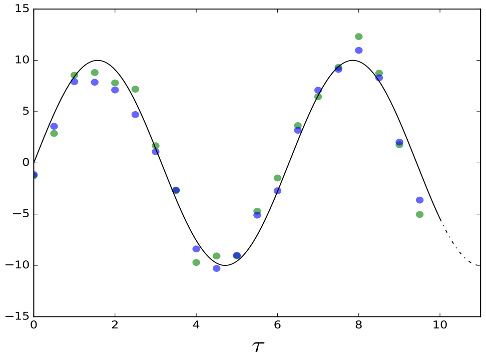

**Figure 1: Plot of the realized actual of the time-series (black solid line), the past forecasts from forecaster 1 (green dots), and that from forecaster 2. We only plot the past forecasts from 20 time-periods to make the plot easier to read. The black dashed line represents the DGP for the future.**

A pseudo-random number generator is used to simulate the forecasts from the two forecasters 5500 times in total. The first 5000 pairs of forecasts ($0 \leq \tau < 10$) are used for training and the remaining 500 pair of forecasts ($10 < \tau \leq 11$) are used for testing. The actual (solid black line) and the past forecasts from the ensemble members are plotted in figure 1. The green dots represent the forecasts from the more skillful forecaster while the blue dots represent those from the less skillful one. As expected, each pair of dots is often located at the same side of the solid black line due to its correlations.

The forecasting skills of the two forecasters and the error correlation thereof are made more apparent in figure 2. In the plot, each (translucent) blue dot represents a *pair* of simultaneous forecasts from the ensemble members. The x-coordinate of a blue dot denotes the forecast value from forecaster 1 ($f_1$, represented by green dots in figure 1) while the y-coordinate denotes that from forecaster 2 ($f_2$, represented by blue dots in figure 1).

In order to find the Joint-PDF $p(\epsilon_1, \epsilon_2)$ that the errors in the forecasting ensemble follow, we adopt a 2-dimensional Gaussian Kernel and apply KDE to the blue dots (with bandwidth selected by cross-validation least square method). The result, represented by a set of contour lines, is plotted over the blue dots in figure 2 (this concludes step 1).

With $p(\epsilon_1, \epsilon_2)$, we are able to calculate the likelihood function of the (yet unrealized) actual. Suppose the forecasting ensemble makes a pair of base forecasts $(f_{1T}, f_{2T})$, where $T$ represents a future time. We notice, by the definition of $p(\epsilon_1, \epsilon_2)$, that the likelihood of $s \in \mathbb{R}$ being the actual is $l(s) = p(f_{1T} - s, f_{2T} - s)$. Namely, the value of $p(\epsilon_1, \epsilon_2)$ taken at $(f_{1T}, f_{2T})$ when the function is re-centered to $(s,\ s)$. The process of making a nonparametric ensemble forecasting is thus visualized in figure 3. In the upper panel, the point $(f_{1T}, f_{2T})$ is denoted by the red star-shaped symbol and the point that corresponds to one of the possible $s$ ($s_1$) is denoted by the blue solid dot located at $(s_1,\ s_1)$. The re-centered Joint-error PDF is

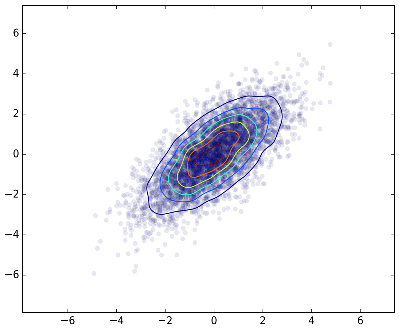

**Figure 2: Kernel Estimation of the Probability Distribution Function for Joint-errors. Each translucent blue dot represents a pair of errors in the training period. The contour lines represents the KDE found by using Cross-validation least square method.**

represented by the set of the blue contour lines. The value which the re-centered Joint-error PDF takes on $(f_{1T}, f_{2T})$ can be inferred from the two blue contour lines that brackets the red star-shaped symbol and is the y-value of the blue solid dot in the lower panel (this concludes step 2). The maximum of this likelihood function of $s$ (denoted by the orange solid dot in both panels) can be found using various optimization algorithms such as that of Nelder and Mead [16] along an auxiliary line:

$$\begin{cases} x = s \\ y = s \end{cases} \quad s \in \mathbb{R}$$

and can be served as a point forecast from the ensemble. In practice, however, one often found that the likelihood function is asymmetric around the maximum. It could result from the fact that the estimated likelihood has a plateau and (hence) the maximum is determined by noise (such as the case in this example), or simply because the likelihood function is inherently asymmetric. As a result, the point where maximum likelihood takes place does not necessarily represent an intuitively-proper point forecast. In such cases, we adopt a loss function $L(q,\ s)$ and report the point forecast as the $q$ where the action

$$a(q) = \int L(q, s)\, l(s)\, ds \tag{4}$$

is minimized [11]. The point forecast $q^*$ generated with a quadratic loss function $L(q, s) = (q - s)^2$ is denoted by the orange dash-dotted line in the lower panel (this concludes step 3). While the point denoted by the blue solid dot $(s_1, s_1)$ is used for illustrative purpose and need not bearing a specific value, we've *chosen* $s_1 = f_{1T}$ in figure 3. Thus, the y-value of the blue solid dot in the lower panel is the likelihood that the actual matches the forecast from the first forecaster. Likewise, we repeated the process for $s_2 = f_{2T}$ and denoted the associate points and contour lines in green. It is worth mentioning that, when the joint-error PDF is symmetric along the line $y = -x$, the point forecast that yields maximum likelihood is exactly the arithmetic mean of the two base forecasts, $\frac{1}{2}\,(f_{1T} + f_{2T})$ (the red solid dot in the upper panel). One such case is when the two forecasters have equal variances and no correlation. In this simplistic case, our result reduces to that of Bates and Granger [4]. We note that the red solid point is also the point on the auxiliary line that is closest to the red star-shaped symbol (the point which represents the pair of base forecasts). The arithmetic mean of the two base forecasts is denoted by the red dash-dotted line in the lower panel. One can see that it does not coincide with either the *argmax* of the likelihood (x-value of the orange solid dot) or the parameter $q^*$ that minimizes the aforementioned action (denoted by the orange dash-dotted line) in general.

The method can be easily generalized to $N$-dimensions:

(1) In the first stage, forecast errors from the past are collected for each forecaster. At each period, the predictions from all forecasters (assume there are $N$ of them in total) form an N-dimensional point. Assume there are $K$ periods in the past, we thus obtain a total of $K$ points in an $N$-dimensional space.
(2) In the second stage, an $N$-dimensional kernel is introduced to estimate the Joint-PDF of the error distribution. In this stage, the error from each forecaster and the correlation thereof are explored and summarized in a numerical PDF.
(3) In the final stage, we make an ensemble forecast as follows:
   - Define the numerical *error* PDF calculated from the second stage to be
   $$p(\epsilon_1, \epsilon_2, \ldots, \epsilon_N) \tag{5}$$
   and denote the actual value of the forecasted quantity (yet unknown) to be $s$. The joint probability function $P$ of the actual value being $s$ and the component *predictions* being $(x_1, x_2, \ldots, x_N)$ is thus
   $$P(x_1, x_2, \ldots, x_N, s) = p(x_1 - s, x_2 - s, \ldots, x_N - s). \tag{6}$$
   We note that equation 6 can be interpreted as the transportation of equation 5 along the line of unit slope in $N$-space:
   $$\{x_k = s,\ \ k \in \mathbb{N}^*_{\le N},\ s \in \mathbb{R}\} \tag{7}$$
   - Given a simultaneous set of component forecasts $f_{1T}, f_{2T}, \ldots, f_{NT}$, the likelihood function of $s$ is
   $$l(s) = P(f_{1T}, f_{2T}, \ldots, f_{NT}, s) \tag{8}$$
   where the first N-slots of $P$ are fixed. The maximum likelihood estimation of the actual value (or alternatively, the point that minimizes the action) can be found numerically.
   - Denote the marginal PDF for the $k^{\text{th}}$ forecaster's error as $p_k(\epsilon_k)$. When all errors in the ensemble are *independent* to each other, equation 5 takes the form of
   $$p(\epsilon_1, \epsilon_2, \ldots, \epsilon_N) = \prod_{k=1}^{N} p_k(\epsilon_k) \tag{9}$$

Simplifications such as equation 9 are necessary for large $N$ cases since the sample size ($\mathcal{S}$) needed to maintain a fixed mean squared error (MSE) of a nonparametric density estimator grows exponentially with $N$ [20]:

$$\mathcal{S} \propto (\frac{c}{\delta})^{N/4} \tag{10}$$

where $c$ is a constant greater than zero and $\delta$ being the MSE.

What we have described in this section are the details of how to generate an ensemble forecast, given a set of the base forecasts at a

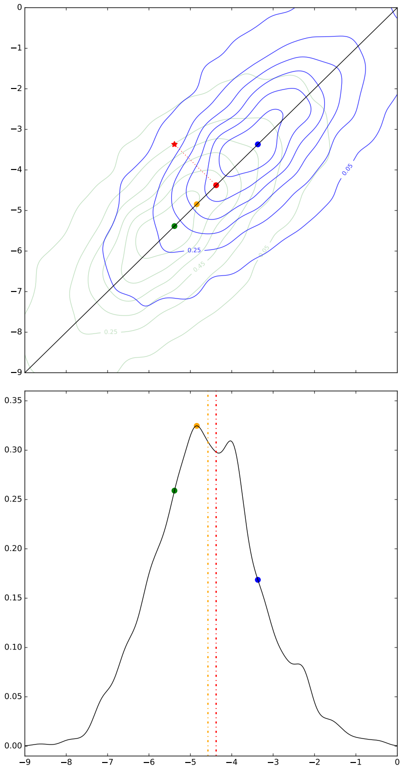

**Figure 3: An example using the proposed method. UPPER: Base forecasts are given by members in the ($N = 2$) ensemble and are denoted by the enlarged red star-shaped sign (define its coordinate to be $\mathbf{x}^*$). We evaluate the joint-PDF of errors at $\mathbf{x}^*$ when it is centered at $(s_1, s_1)$ [blue dot and contour lines] and $(s_2, s_2)$ [green dot and faded green contour lines] respectively. The value is by definition the likelihood density of $s$ when $s = s_1$ and $s = s_2$. LOWER: All points on the line of unit slope ($x_1 = x_2 = s$) in the upper panel can yield a likelihood density that is generated in the aforementioned procedure. The full likelihood density function is represented by the solid black line. The point that corresponds to the maximum likelihood estimation is marked in orange. The three colored points (blue, orange, green) in the upper and lower panels have an exact one-to-one correspondence. Alongside the blue and green points are their respective s-value and the corresponding likelihood. We note when the joint-error PDF is symmetric along the line $y = -x$, the MLE reduced to the closest point between $x^*$ and the line $x_1 = x_2$, which can be shown to be $(\frac{x_1^*+x_2^*}{2}, \frac{x_1^*+x_2^*}{2})$. In the notation of equation 2, $F = \frac{1}{2} f_{1T} + \frac{1}{2} f_{2T}$.**

specific time $T$. In section A.1, we show how the proposed method performs in comparison to EMOS and BMA on the 500 testing data points under the settings of equation 3.

## 3 SIMULATION STUDY: FORECASTING ENSEMBLE WITH UNIFORM ERROR DISTRIBUTIONS

In this section, we apply the proposed method to the case where the forecasters' error distributions are independent and uniform, i.e., where the PDF for each forecaster has the following form:

$$p_k(\epsilon_k) = \frac{1}{2\eta_k}(H(\epsilon_k + \eta_k) - H(\epsilon_k - \eta_k)) \qquad (11)$$

where $H$ is the Heaviside step function, and $\eta_k$ is the half-width of the uniform error. We construct the forecasting ensemble to consist of six members, i.e., $k = 1, 2, \ldots, 6$ and listed $\eta_k$ in section A.2. The setup of the study is otherwise kept as close as possible to that described in section 2. Namely, our DGP remains to be the sinusoidal function and the training data contains 5000 sets of past forecasts ($0 \le \tau < 10$). The proposed method is compared with other methods on 500 sets of testing forecasts ($10 < \tau \le 11$).

If we apply the proposed method directly with equation 6, we would need a total of 12.5 billion sets of past forecasts to make the estimated Joint-error pdf as accurate as that estimated in section 2 (cf. equation 10). Since a time series dataset of such order of magnitude is rare in practice, we must take simplifications such as that in equation 9 to render the task feasible.

As mentioned in the introduction, the family of error distribution needs to be specified when using a parametric model. Given that we are benchmarking the proposed approach with the popular ensemble method implementations BMA [18] and EnsembleMOS [21], we need to be consistent in this distribution assumptions. We therefore proceed with the option "gaussian" as it is considered to be the most natural choice when facing an unknown distribution. We emphasize that the proposed nonparametric method has zero knowledge about the error distribution a priori. Thus, this setup mimics the case where one deals with an ensemble of forecasters whose FGP is unknown to all of the ensemble algorithms (whereas in section 2, the BMA and EMOS procedures took the advantage of knowing the type of underlying error distribution beforehand).

An example of a nonparametric ensemble forecast construction (forecast for $\tau = 11$) in this setting is visualized in figure 4. With the simplification of equation 9, the six-dimensional PDF along the line of $\{x_k = s,\ k \in \mathbb{N}^*_{\le 6},\ s \in \mathbb{R}\}$ is reduced to the product of six one-dimensional PDFs. The estimated value of the six one-dimensional PDFs and the product thereof is denoted by six faded lines and a thick solid black line respectively. We can see that the ensemble makes a much sharper forecast than each of the base forecasters.

Details of the simulation parameters and the comparison of performance among different forecasting scheme is provided in section A.2.

## 4 THE WIKIPEDIA WEB TRAFFIC DATASET

We apply the proposed method to the Wikipedia Web Traffic Dataset (https://www.kaggle.com/c/web-traffic-time-series-forecasting/data). Each time series in this dataset represents a number of daily views

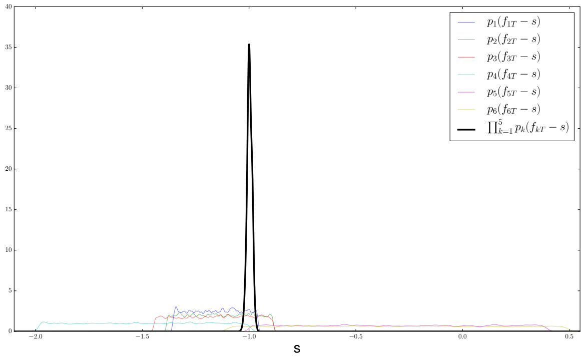

**Figure 4:** $p_k(f_{kT}-s),\ k=1,2,...,6$ **and** $\prod_{k=1}^{6} p_k(f_{kT}-s)$ **are plotted in the same figure. The marginal likelihood density function of each forecaster is centered at their point forecast $f_{kT}$ while the width of the distribution measures its past performance. The product of all of the individual density function is represented by the thick black line (cf. legend). It has been normalized such that it integrates to one.**

of a different Wikipedia article for a specific type of traffic (all, mobile, desktop, spider), starting from July 1st, 2015 up until December 31st, 2016 – a total of 550 days. We take the first 507 days (July 1st, 2015 to November 18th, 2016) as the training period (note that the first seven days of the data are only used for generating the first set of base forecasts and are not used for training the ensemble) and test different forecasting schemes on the last 43 days (November 19th, 2016 to December 31st, 2016).

We consider an ensemble that consists of two forecasters. The first forecaster forecasts the next day's traffic by looking back that of the past seven days (including the day when it conducts the forecast) and picks the median as its forecast. The second forecaster takes the traffic one week prior to the next day (i.e., six days ago) to be its forecast for the next day. Thus, all of the forecasts (and errors thereof) in this ensemble are integers. If we view the Joint-error PDF on a two-dimensional plane, the PDF only takes values on the grid points. Hence, smoothing with KDE is not needed in this setting. A two-dimensional histogram of the past errors would suffice for our purposes.

We focus on the low traffic websites in English, which are those websites with the domain name "en.wikipedia.org" and with a maximum daily view count less or equal to 50 on/before the last day of the training period (November 18th, 2016). We exclude time series of the traffic type "all" since such traffic includes both mobile and desktop traffic. In addition, we exclude the time series with one or more zero values since the dataset does not distinguish between traffic values of zero and missing values [1]. This left us with nine time series. Since the remaining nine time series are quantitatively similar, the errors that the forecasting ensemble made by forecasting them can be viewed as *i.i.d.*. In figure 5, we visualize the 2d-histogram of errors from the two forecasters (which is an estimate of the Joint-error PDF). The training data set contains 4500 points (9 time series with 500 data points each). It can be observed that the two dimensional distribution deviates rather significantly from the Gaussian distribution.

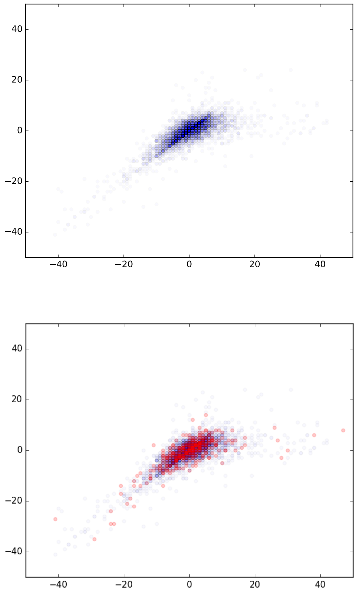

**Figure 5: UPPER: The histogram of errors made by base forecasters in the training period for the Wikipedia Web Traffic Dataset. It contains 4500 points. Darker spots in the plot implies a higher count for its associated error pair. LOWER: The histogram of the errors made by base forecaster in the testing period is over-plotted on that in the training period. The four outliers are too far away from the plotted region and is not shown. We note that "training" and "testing" in the current context here means the training and testing of ensemble methods, and should not be confused with that of base forecasters. We assumed all base forecasters are already trained and are not updated during the process of training ensemble.**

The process of making an nonparametric ensemble forecast in this case is quite similar to that described in section 2 and 3. The only difference being that the Joint-error PDF is no longer a smooth function. Rather, it is a function that only takes value on integer points. In terms of equation 4,

$$l(s) = \sum_{k=1}^{K} G(f_{1T}-s_k, f_{2T}-s_k)\,\delta(s-s_k) \tag{12}$$

where $K$ is the total number of grid points that have non-zero values, $G$ the two-dimensional histogram, and $\delta(s)$ is the Dirac Delta function. We show $G(f_{1T}-s_k, f_{2T}-s_k)$ as a function of $s_k \in \mathbb{Z}$ in figure 6.

When making the forecasts in the testing period, we found very large root mean squared error (RMSE) (∼ 15620) across *all* forecasters. Looking into the data, we found that RMSE is mainly driven by a few outlier points.

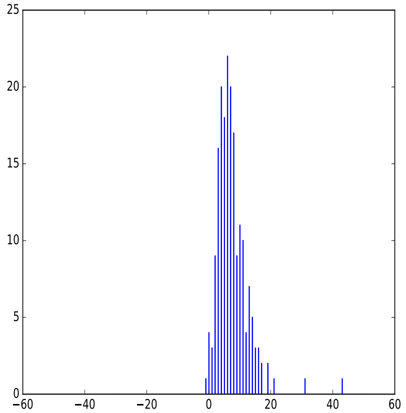

**Figure 6: $G(f_{1T} - s,\ f_{2T} - s)$ where $f_{1T} = 3$ and $f_{2T} = 7$. This shows the likelihood of the unrealized actual being $s$. Note that the PDF is the sum of some Dirac Delta functions and the histogram represents the coefficients thereof.**

For example, the webpage https://en.wikipedia.org/wiki/Bar_bet has a desktop web traffic of 301627 on December 1st, 2016 (that of the previous day is only 28). This could be traffic triggered by news events, internet memes, or could be a data issue. When we compare the different forecasting schemes, the proposed method has the best accuracy (with outliers included). After removing the four outliers, the RMSE of all forecasting scheme reached the order of 20, and the proposed method remains the most accurate (22). When we calculate the Z-score, we found the proposed method is 1.2 standard deviations from BMA and 1.85 standard deviations from EMOS. It is also the only ensemble method that performs better than both base forecasters. In the right panel of figure 5, we over-plot the test errors (excluding outliers) on the Joint-error PDF estimated with the training data. We found that they largely overlap.

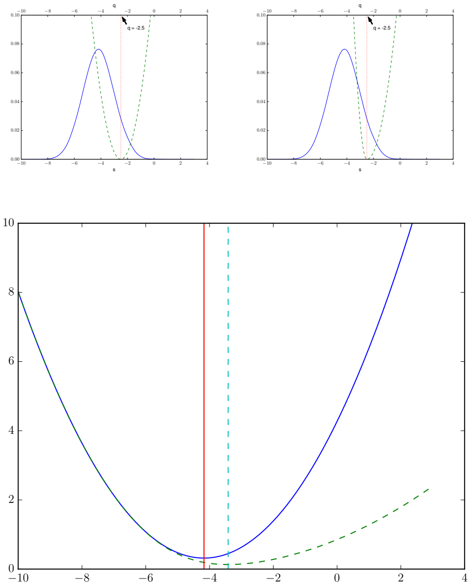

**Figure 7: The use of customized loss functions. In all plots, the blue bell-shaped line represents the likelihood function that is found with the proposed method. UPPER LEFT: The green dashed line represents the quadratic loss function which we use in this work. In the plot, the parameter $q$ (marked out with the red dotted vertical line) is set to be 2.5 for illustrative purposes. UPPER RIGHT: The green dashed line represents a customized (asymmetric) quadratic loss function that makes the forecasters prefer overestimation to underestimation. LOWER: When the customized loss function is used, the *argmax* of action shifted from the solid red vertical line to the dashed light-blue line, which implies a higher forecast.**

## 5 DISCUSSIONS AND CONCLUSIONS

In this paper, we proposed a new approach to the problem of ensemble forecasting. While the existing ensemble methods focus on finding some combination weights of the forecasted *values*, the proposed method focuses on exploring the N-dimensional PDF of the base forecast *errors* in a numerical fashion. Depending on the type of numerical numbers one forecasts, implementations differ slightly. Specifically, when the forecasts from the base forecasters are continuous numbers, KDE is used to estimate the underlying PDF. When the base forecasters produce only discrete numbers, one can also use an N-dimensional histogram for estimation.

A likelihood function of the actual is generated as an intermediate step of the proposed method, which grants one additional flexibility in forecasting. Sometimes, the purpose of forecasting is not to minimize MSE but to minimize a subjective risk of the forecaster. By introducing a customized loss function in equation 4, this purpose can be easily fulfilled (cf. Diebold [10]). For example, when planning for power supply, the forecaster might determine that the costs associated with overestimating future power usage are smaller than that associated with underestimation (since underestimation could result in an electric outage). He or she can introduce an asymmetric loss function (cf. figure 7) which would push the ensemble forecast higher, reducing the risk of power outage.

We have shown that the proposed method achieves comparable accuracy to conventional methods when the underlying error distribution is Gaussian (cf. section 2) and achieves superior accuracy when the underlying error distribution is generic (cf. section 3). In real settings such as the Wikipedia Web Traffic Data, the assumption of the proposed method does not firmly hold in the sense that the error distribution of the testing dataset is different from that of the training data set. However, the proposed method showed robustness and remains more accurate than the conventional methods.

The biggest challenge of the proposed approach is the requirement to obtain sufficient amount of data to control the error of the nonparametric density *estimators*. This is hopefully less of an issue, as the amount of data collected increases exponentially every day.

# A DETAILS OF THE NUMERICAL EXPERIMENTS

We describe the numerical details of the three use cases for easy references.

## A.1 Experiment 1: Two-dimensional Gaussian Joint-error PDF

In this experiment, the underlying mean and covariance matrix used to generate data is

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2.25 \end{bmatrix}$$

The statistics of the training sample ($\mathcal{S} = 5000$) is as follows:

$$\hat{\mu}_{in} = \begin{bmatrix} -0.0286 \\ -0.01 \end{bmatrix} \quad \hat{\Sigma}_{in} = \begin{bmatrix} 1.97 & 1.47 \\ 1.47 & 2.24 \end{bmatrix}$$

The performance of various forecasting scheme on the testing sample ($\mathcal{S} = 500$) is as follows:

| Rank | Description | MSE |
|---|---|---|
| 1 | EMOS | 1.87 |
| 2 | This work (action minimization) | 1.88 |
| 3 | BMA | 1.95 |
| 4 | This work (maximum likelihood) | 2.00 |
| 5 | $f_1$ | 2.17 |
| 6 | $f_2$ | 2.18 |

We can further conduct a paired Z-test between this work (action minimization) and other forecasting methods:

| Paired Forecasting Method | Z-score |
|---|---|
| $f_1$ | -4.92 |
| $f_2$ | -3.49 |
| BMA | -1.43 |
| EMOS | 0.24 |

We found the proposed method has comparable performance to EMOS and is better than BMA and the base forecasters.

## A.2 Experiment 2: Six Forecasters with Independent Uniform Errors

In this experiment, the underlying half-width ($\eta$) of the forecasters are:

| Parameter | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
|---|---|---|---|---|---|---|
| $\mu$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\eta$ | 0.20 | 0.25 | 0.28 | 0.50 | 0.70 | 0.80 |

The statistics of the training sample ($\mathcal{S} = 5000$) are extremely close to the underlying parameters (the deviation in $\mu$ is $\sim O(10^{-3})$ and that in $\eta$ is $\sim O(10^{-4})$) and we shall omit here. The performance of various forecasting scheme on the testing sample ($\mathcal{S} = 500$) is as follows:

| Rank | Description | MSE |
|---|---|---|
| 1 | This work | 0.003893 |
| 2 | EMOS | 0.005622 |
| 3 | BMA | 0.005704 |
| 4 | $f_1$ | 0.013043 |
| 5 | $f_2$ | 0.019991 |
| 6 | $f_3$ | 0.027044 |
| 7 | $f_4$ | 0.082629 |
| 8 | $f_5$ | 0.165969 |
| 9 | $f_6$ | 0.195908 |

When we conduct paired Z-test, the result are as follows:

| Paired Forecasting Method | Z-score |
|---|---|
| $f_1$ | -16.68 |
| $f_2$ | -18.61 |
| $f_3$ | -21.70 |
| $f_4$ | -23.08 |
| $f_5$ | -24.64 |
| $f_6$ | -23.68 |
| BMA | -8.58 |
| EMOS | -8.52 |

We conclude that the proposed method performs better than all other methods in this setting.

## A.3 Application to Wikipedia Web Traffic Dataset

For this application, we use RMSE in the tables (instead of MSE) since the MSE is very high due to the outliers (cf. section 4). The paired Z-test is still conducted on the paired difference of squared error between this work and other methods. The RMSE of various forecasting method with the outliers included are as follows:

| Rank | Description | $\sqrt{MSE}$ |
|---|---|---|
| 1 | This work | 15620.205859 |
| 2 | BMA | 15620.228509 |
| 3 | $f_2$ | 15620.244895 |
| 4 | $f_1$ | 15620.355319 |
| 5 | EMOS | 15620.496817 |

When we conduct paired Z-test, the results are as follows:

| Paired Forecasting Method | Z-score |
|---|---|
| $f_1$ | -1.057362 |
| $f_2$ | -1.037428 |
| BMA | -1.079681 |
| EMOS | -1.031325 |

After removing four outliers, RMSE becomes:

| Rank | Description | $\sqrt{MSE}$ |
|---|---|---|
| 1 | This work | 22.244917 |
| 2 | $f_2$ | 23.042732 |
| 3 | BMA | 23.572635 |
| 4 | $f_1$ | 24.324199 |
| 5 | EMOS | 24.697943 |

The results for the paired Z-test are:

| Paired Forecasting Method | Z-score |
|---|---|
| $f_1$ | -1.582394 |
| $f_2$ | -0.999291 |
| BMA | -1.204237 |
| EMOS | -1.849362 |

We conclude that the proposed method is robust under the influence of outliers and generates more accurate forecasts (in particular, when compared to EMOS).

## ACKNOWLEDGMENTS

We would like to thank the following people for helpful discussions and feedbacks: Johnny Hong (Berkeley), Yan Zheng (VISA), Anirudh Sankar (MIT). Last but not least, our wonderful colleagues at Adobe, in particular Bo Peng and Jian Li.

## REFERENCES

[1] [n. d.]. Wikipedia Web Traffic Data Description. ([n. d.]). https://www.kaggle.com/c/web-traffic-time-series-forecasting/data Online; Accessed: 2018-02-07.

[2] Cagdas Hakan Aladag, Erol Egrioglu, and Ufuk Yolcu. 2010. Forecast combination by using artificial neural networks. *Neural Processing Letters* 32, 3 (2010), 269–276.

[3] J Scott Armstrong. 2001. Combining forecasts. In *Principles of forecasting*. Springer, 417–439.

[4] John M Bates and Clive WJ Granger. 1969. The combination of forecasts. *Journal of the Operational Research Society* 20, 4 (1969), 451–468.

[5] Derek W Bunn. 1975. A Bayesian approach to the linear combination of forecasts. *Journal of the Operational Research Society* 26, 2 (1975), 325–329.

[6] Richard M Chmielecki and Adrian E Raftery. 2011. Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review* 139, 5 (2011), 1626–1636.

[7] Robert T Clemen. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5, 4 (1989), 559–583.

[8] Lilian M De Menezes, Derek W Bunn, and James W Taylor. 2000. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research* 120, 1 (2000), 190–204.

[9] Francis X Diebold. 1988. Serial correlation and the combination of forecasts. *Journal of Business & Economic Statistics* 6, 1 (1988), 105–111.

[10] Francis X Diebold. 2001. Elements of forecasting. (2001).

[11] Francis X Diebold, Todd A Gunther, and Anthony S Tay. 1998. Evaluating density forecasts. *International Economic Review* 39, 4 (1998), 863–883.

[12] Paulo SA Freitas and António JL Rodrigues. 2006. Model combination in neural-based forecasting. *European Journal of Operational Research* 173, 3 (2006), 801–814.

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

[14] Clive WJ Granger and Ramu Ramanathan. 1984. Improved methods of combining forecasts. *Journal of forecasting* 3, 2 (1984), 197–204.

[15] Gerald J Lobo. 1991. Alternative methods of combining security analysts' and statistical forecasts of annual corporate earnings. *International Journal of Forecasting* 7, 1 (1991), 57–63.

[16] John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *The computer journal* 7, 4 (1965), 308–313.

[17] Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* 133, 5 (2005), 1155–1174.

[18] Adrian E Raftery and Ian S Painter. 2005. BMA: an R package for Bayesian model averaging. *R news* 5, 2 (2005), 2–8.

[19] Allan Timmermann. 2006. Forecast combinations. *Handbook of economic forecasting* 1 (2006), 135–196.

[20] Larry Alan Wasserman. 2006. *All of nonparametric statistics: with 52 illustrations*. Springer.

[21] RA Yuen, T Gneiting, TL Thorarinsdottir, and C Fraley. 2013. ensembleMOS: Ensemble model output statistics. *R package version 0.7, http://CRAN. R-project. org/package= ensembleMOS (accessed 15 April 2016)* (2013).