Demand Forecasting of individual Probability Density Functions with Machine Learning

Felix Wick [*1], Ulrich Kerzel [†2], Trapti Singhal [‡1], and Martin Hahn [§1]

[1]Blue Yonder GmbH (Karlsruhe, Germany)
[2]IUBH Internationale Hochschule (Erfurt, Germany)

... abstract ...

Keywords: **machine learning, demand forecasting PDF**

1. Introduction

Demand forecasting is one of the main challenges for retailers and at the core of business operations. Due to its stochastic nature, demand is difficult to forecast as it depends on many influencing factors and the realized demand can be interpreted as a random variable that is described by an appropriate probability density function (PDF). In order to make operational decisions, an optimal point estimator has to be defined, that can be used to derive ordering decisions used in the replenishment process of the retailer. Demand estimation is further complicated by the fact that retailers typically only observe realized sales rather than the actual demand, and in case the demand exceeds the current stock level the data become censored. The ordering decision process is complicated by a range of factors: Even in the case of accurate demand forecasts, the decision maker has to balance conflicting metrics to reach an optimal decision: Ordering to few items may result in stockout situations leading to unrealized demand and unsatisfied customers. Ordering too many items results in excess inventory which increases transport and storage costs and, in the case of perishable goods, excessive waste, as spoilt items need to be disposed of at additional cost and potentially even environmental impact. This situation is particularly noticeable in the so called "ultra-fresh" category, which includes items such as bakery products, ready-meals, fresh diary products, or certain meat products such as ground meat. These items typically have a shelf-life of less than a business day to a few business days at most, with a continuous spectrum depending on the exact item. In many situations, additional constraints have to be considered to reach an optimal ordering decision: Delivery cycles of items may vary depending on the type of item and the wholesaler or manufacturer from which they are procured. Retailers also operate at a given service level to guarantee that a certain level of demand can be fulfilled. The exact service level typically depends on the overall business policy of the retailer and may also depend on individual products, ranging from "never-out-of-stock" items to a service level exceeding e.g. 90%.

The availability of Big Data allows capturing, storing and processing a vast amount of data associated with demand, such as historic sales records, information about promotional events or advertisements, pricing information, local weather at retail locations, seasonal information as well as a wide range of further variables. Modern machine learning algorithms can then be used to predict the per-item demand distribution, corrected for censored data, from which an optimal point estimator can be derived to be used in the subsequent ordering decision. It is important to note that demand as a random variable is not identically and independently distributed (i.i.d.). While the probability distribution describing the demand can be attributed to a given family or parametrization, the exact parameters vary: Seasonal effects, finite life cycles of products and the introduction of new products influence the demand distribution, as well as the local weather at the retail location or the retail location itself in terms of size, assortment range, customer diversity and other factors. The retailers themselves also actively influence demand by using advertisements to highlight products, offering rebates or discounts for specific products as well as pursuing an active pricing strategy. This means that while we can generally assume that demand follows a specific type of probability distribution, its parameters are unique to the instance for which an estimate is required. For example, the probability distribution governing the demand of a particular item is specific to the item, date and retail location for which the forecast is made and depends on a wide range of further influencing factors.

The remainder of the paper is organized as follows: We first review the relevant literature and existing work in sec. 2. We then describe our method to predict individual negative binomial PDFs by means of a parametric approach including two distinct machine learning models for mean of variance in sec. 3. After that we describe methods for the qualitative and quantitative evaluation of PDF predictions in sec. 4. And finally, we present a demand forecasting example to show an application of our methods in sec. 5.

---
[*]felix.wick@blueyonder.com
[†]u.kerzel@iubh-fernstudium.de
[‡]trapti.singhal@blueyonder.com
[§]martin.hahn@blueyonder.com

2. Literature Review

Inventory management offers a rich theory and the extensive body of research can be broadly grouped into two categories, where the inventory control problem is either based on some knowledge of the underlying demand distribution or an integrated approach that seeks to map directly from the available data (historic sales records and further variables) to the ordering decision. This approach is often referred to as "data-driven newsvendor" and discussed e.g. in [?, ?, ?, ?]. It aims to avoid estimating the underlying probability distribution and use the available data to derive the operational decisions directly. An overview of a range approaches can also be found in [?].

Although this approach seems preferable, since it avoids determining the full demand distribution and results directly in the desired operational decision (the order quantity), it faces several drawbacks. First of all, the full probability distribution for the demand of a specific item at a given sales location and business day includes all available information including the uncertainty of the modelled demand. This can be used to simulate the performance on a per-item level and e.g. optimize the impact on business strategy decisions on conflicting metrics such as stock out- and waste-rate. Additionally, having the full demand probability distribution available allows for an accurate assessment of the forecast quality of all quantiles of the distribution, including the often extensive tails of the distribution. Separating the demand forecast from the operational decisions (i.e. calculating the order quantities for the next delivery cycle) also allows to evaluate longer-term planning or counterfactual analysis to investigate how the demand would have been had the input variables had different values. This way, effects of promotions, advertisements or price-changes can be analyzed with the same model that is used for operational decisions. Furthermore, deriving the ordering decision directly couples the demand process with the complex delivery cycles and constraints, while keeping the steps separate allows greater operational flexibility and reduces the impact of changing manufacturers or wholesalers, as well as allowing a quick response to changes in the delivery schedule without having to re-compute the implicit demand underlying the ordering decision. Finally, long-term demand forecasts may be shared with other business units or external vendors and wholesalers to ease their planning for the production and supply-chain processes upstream of the retailer. From the perspective of industrial practice, modelling the demand separately from deriving the subsequent orders has the additional benefit that multiple retail chains can benefit from any improvement in the model description even if the concrete retailers are unrelated to each other. For example, if a particular effect was identified at some retailer A and included in the machine learning model, the improvement can be rolled out immediately or on request to all other retailers using this system on a planetary scale without having to adapt the underlying machine learning model for each retailer individually.

In contrast to the direct mapping the observed data to ordering decisions, more traditional inventory control systems rely on the knowledge of the demand distribution in one form or another, see e.g. [?] for an overview. In (s, S) type inventory control systems [?], inventory levels are monitored at regular intervals and orders are dispatched once the inventory level reaches a minimal value s. In case of linear holding and shortage costs, such policies are optimal [?], although perishable goods pose more challenges, see e.g. [?, ?, ?]. Additionally, service level constraints can be included in these kind of inventory control systems [?]. Perishable goods are well described by the "newsvendor-problem" [?], where in the simplest case all stock perishes at the end of the selling period (e.g. a business day). For a detailed review of the newsvendor problem see e.g. [?]. Assuming linear underage and overage costs $b, h > 0$, the optimal quantile $q_{\mathrm{opt}} = b/(b+h)$ of a known demand distribution $f(D)$ can be calculated exactly.

The main objective in any of these approaches is to determine the underlying demand distribution. The simplest approach is to just use the observed sales events and forecast these as a time series (see e.g. [?]) or via sample average approximation (SAA), see e.g. [?] for an overview. However, these approaches do not make use of any data apart from the sales record themselves, although we know that many variables such as price or advertisements influence, and therefore are highly correlated with, the demand. Saghafian and Tomlin [?] propose to include partial information about the distribution in the derivation of the operational decision, i.e. the calculation of the optimal order quantity.

However, in order to be able to fully optimize the operational decision, it is critical that one indeed reconstructs a full demand distribution. This also implies that a simple point-estimator, as provided by the most common statistical techniques and machine-learning approaches, will not suffice. Additionally, we need to consider that demand is not i.i.d., but depends on external factors such as season, weather, product life-cycle, and is also actively changed by the retailer by setting a specific price, offering rebates or running advertisements. Additionally, the demand implicitly depends on the location of the retail outlet as well as the specifics of that location, such as product assortment influencing the choice of possible replacement articles and many more. These complications are the main reason we cannot treat the replenishment process as n independent newsvendor-type problems.

Instead, we need to determine the full demand distribution from data, conditional on the relevant variables such as date, location, and item, taking all auxiliary data such as article characteristics, pricing, advertisements, retail location details, etc. into account. This can be done in several ways: Quantile regression [?, ?] can be implemented in various frameworks and used to estimate a range of quantiles for each predicted distribution from which an empirical probability distribution can be interpolated. Using a dedicated neural network [?], either the full probability distribution or a defined range of quantiles can be calculated directly from the data for each individual prediction without assuming an underlying model. Alternatively, one can assume a given demand model and fit the model parameters instead of reconstructing the complete distribution [?, ?]. This approach is computationally favourable, as fewer parameters need to be estimated com-

pared to the case of the full distribution. Empirically, one can determine the best fitting distribution from data [?]. However, given the stochastic nature of the demand, such an empirically determined distribution is not expected to be stable and prone to sudden changes. Instead, the choice of the demand distribution should be motivated by theoretic considerations. The discrete demand is typically modelled as a negative binomial distribution (NBD), also known as Gamma-Poisson distribution [?, ?, ?, ?, ?]. This distribution arises if the Poisson parameter μ is a random variable itself that follows a Gamma distribution. The NBD has two parameters, μ and $\sigma^2 > \mu$, and is over-dispersed compared to the Poisson distribution for which $\mu = \sigma^2$. Hence, for each ordering decision, the model parameters μ and σ need to be determined for each item at the required granularity, typically for each sales location and ordering time, depending on all auxiliary data describing article details, retail location, and influencing factors such as pricing and advertisement information.

Summary of Contributions

In the following, we will demonstrate how the explainable machine learning algorithm Cyclic Boosting [?] can be used to model the demand distribution at the granularity needed by the retailer. Typically this means that the full demand distribution has to be estimated per SKU for each sales location and opening day, conditional on a wide range of variables such as weather, prices, promotions, etc. In contrast to using a "black-box" machine learning model, using Cyclic Boosting allows to interpret how each individual prediction was made and what the most important variables are per individual prediction.

Additionally, we show how the cumulative distribution function (CDF) or inverse quantiles can be used to accurately asses the forecast quality of the full predicted demand distribution, including the tails of the distribution. This allows to verify that the predicted demand distribution accurately reflects the observed data and can hence be used both to derive operational decisions such as order quantities as well as strategic business decisions by the retailer.

3. Negative Binomial PDF Estimation

To predict an individual PDF by means of a parametric approach, one has to rely on a model assumption about the underlying distribution of the random variable to be predicted. In the following, we describe such a parametric approach with the model assumption of a negative binomial probability distribution. For a PDF prediction under a negative binomial model assumption, there is the need for the estimation of two parameters, for example its mean and its variance.

This can be done using two independent models, one to estimate the mean and the other for the variance. At least in principle, any method can be used. However, as argued above, in the case of demand forecasting, each prediction is highly specific to the circumstances in which it is used (such as product ID, date, and store location) and may depend on a multitude of describing variables

(features). Machine learning algorithms are ideally suited for this task and in the following we will use the Cyclic Boosting algorithm. The major benefits of this method are that it is not only extremely performant in practical applications of demand forecasting at large scale, but also fully explainable. Traditional machine learning algorithms are typically "black box" approaches, where the individual decision cannot be explained. Cyclic Boosting on the other hand allows to follow how each individual prediction was made. The main idea of this algorithm is the following: When used for regression, the predictions $\hat{y}_i$ of the target variable $Y \in [0, \infty)$ can be calculated in the following way:

$$\hat{y}_i = \mu \cdot \prod_{j=1}^{p} f_j^k \quad \text{with } k = \{x_{j,i} \in b_j^k\} \tag{1}$$

The parameters f_j^k are the model parameters that are determined from features j. Each feature is discretized appropriately into k bins to reflect the specific behaviour of the feature. The global mean μ is determined from all target values y observed in the data. The factors f_j^k are determined iteratively until convergence is reached and regularization techniques are applied to avoid overfitting and improve the generalization ability of the algorithm. The relative strength of the factors f_j^k are then directly interpretable in relation to how important a specific feature is for each individual prediction, where deviations from $f_j^k = 1$ indicate high importance.

Need for some details on Cyclic Boosting training process

For the PDF estimation, two subsequent, independent Cyclic Boosting models are used, the first to estimate the mean and the second to estimate the variance. The features may or may not differ between the mean and variance estimation models. But in any case, as there is a strong correlation between mean and variance, it is beneficial to include the corresponding predicted mean from the mean model as feature in the variance model. The assigned mean and variance predictions can then be used to generate individual PDFs according to a functional negative binomial distribution assumption for each sample.

While any machine learning algorithm outputting a conditional mean as prediction may be used to get an approximate estimate for the mean of an underlying negative binomial distribution, it is preferable to choose an algorithm using a negative binomial assumption for its inner regularization methods, like Cyclic Boosting in multiplicative regression mode.

In the upcoming section, we describe a method to predict the variance of a negative binomial PDF along with its mean individually for each sample by means of Cyclic Boosting.

3.1. Variance Estimation by Cyclic Boosting in Negative Binomial Width Mode

Cyclic Boosting in a modified form of its multiplicative regression mode, namely the negative binomial width mode presented in the following, can be used for the second machine learning model to predict the negative binomial variance associated with the mean predicted by the first model.

For regression, the parametrization of the negative binomial mass function can be specified as [**?**]:

$$\text{NB}(y; \mu, r) = \frac{\Gamma(r+y)}{y! \cdot \Gamma(r)} \cdot \left(\frac{r}{r+\mu}\right)^r \cdot \left(\frac{\mu}{r+\mu}\right)^y, \quad (2)$$

with mean μ and dispersion parameter r. The target variable y takes the values $y = 0, 1, 2, \dots$.

The variance estimation is achieved by minimizing the loss function defined in Eqn. (3), expressed as negative log-likelihood function of a negative binomial distribution, with respect to r_i over all input samples i, where the mean values $\hat{\mu}_i$ are fixed to the corresponding predictions from the first machine learning model outputting the conditional mean.

$$L(r) = -\mathcal{L}(r) = -\ln \sum_i \text{NB}(y_i; \hat{\mu}_i, r_i) \quad (3)$$

As stated in Eqn. (4), the values r_i are defined by the Cyclic Boosting model parameters f_j^k for each feature j and bin k. For any concrete observation i, the index k of the bin is determined by the observation of $x_{j,i}$ and the subsequent look-up into which bin this observation falls. The dispersion parameter r is bound to the interval $[1, \infty]$.

$$r_i = 1 + \prod_{j=1}^{p} f_j^k \quad \text{with } k = \{x_{j,i} \in b_j^k\} \quad (4)$$

Need for more details.

As described in [**?**], the parameter estimation in Cyclic Boosting is an iterative method corresponding to a cyclic coordinate descent, processing one feature with all its bins at a time until convergence. Unlike in the original multiplicative regression mode of Cyclic Boosting, the minimization of the loss function in Eqn. (3) cannot be solved analytically and requires a numerical method, for example a random search. All other advantages of Cyclic Boosting, like for example individual explainability of predictions, remain valid for its negative binomial width mode.

Finally, the variance $\hat{\sigma}_i^2$ can be calculated from the dispersion parameter $\hat{r}_i$ by means of Eqn. (5). And so, together with the estimated mean parameter $\hat{\mu}_i$ from the previous mean estimation model, all parameters of the negative binomial distribution NB for the prediction of an individual observation i are given.

$$\sigma^2 = \mu + \frac{\mu^2}{r} \quad (5)$$

4. Evaluation of PDF Predictions

Many statistical and most machine learning methods do not provide a full probability density distribution as result. Instead, these methods typically predict a single numerical quantity that is then compared to the observed concrete realization of the random variable using metrics such as the mean squared error (MSE), mean absolute deviation (MAD) or others. In the setting of a retailer, the observed quantity is the sales of individual products (denoted by y) and most machine learning approaches would then predict a single number $\hat{y}$ as a direct estimate of the sales. However, reducing the prediction to a single number does not allow to account for the uncertainty of the prediction or the dynamics of the system. Instead, it is imperative to predict the full PDF for each prediction to be able to optimize the subsequent operational decision. Unfortunately, most statistical or machine learning methods that predict full individual probability functions lack quantitative or at least qualitative evaluation methods to assess whether the full distribution has been forecast correctly, in particular in the tails of the distribution.

For an estimation of the determining parameters of an assumed functional form for the PDF, assessing the correctness of the PDF model output refers to the evaluation of the accuracy of the prediction of the different determining parameters. In the case of the negative binomial distribution used in this work, we have to verify that mean and variance are determined accurately, as well as checking that the choice of the underlying model can describe the observed data.

In the following, we will show how histograms of the CDF and inverse quantiles can be used to determine the accuracy of the predicted PDFs. Although we limit the following discussion to the negative binomial model, the method can be applied generally to any representation of a probability density distribution, even if the PDF is obtained empirically.

4.1. Qualitative Evaluation of PDF Predictions

As a first step, we use the inverse quantile function or probability integral transform (PIT), see e.g. [**?**, **?**]. This offers a first qualitative assessment by looking at the distribution of the quantiles of the predicted density via the CDF. In case of a correctly calibrated predicted PDF, this distribution is expected to be uniform and any deviation can be interpreted as a hint that the predicted probability function is not correct [**?**].

4.1.1. Histogram of the CDF

The CDF of a PDF $f(x)$ is defined as:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(x')dx' \quad (6)$$

Here, $F_X(x)$ is the CDF with $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_x(x) = 1$. The cumulative distribution describes the probability that a the variable has a value smaller than x and intuitively represents the area under $f(x')$ up to a point x.

If the CDF is continuous and strictly increasing, then the inverse of the CDF $F^{-1}(y)$ exists and is a unique real-valued number x for each $y \in [0, 1]$ so that we can write $F(x) = y$. The latter defines the inverse quantile function because we can define the quantile τ of the probability distribution $f(x)$ as:

$$Q_\tau = F^{-1}(\tau) \quad (7)$$

Using the example of the normal distribution with $\mathcal{N}(0, 1)$ as shown in Fig. 1, we can identify the median ($\tau = 0.5$) by first looking at the CDF in the lower part of

Figure 1: Probability distribution function and cumulative distribution funtion of a normal distrubution.

Figure 2: ...

the figure, look at $y = 0.5$ on the y-axis and then identify the point on the x axis for both the PDF $f(x)$ and the CDF $F(x)$ that correspond to the quantile τ. In the case of the normal distribution, this is of course the central value at zero.

We can then interpret the CDF as a new variable $s = F(t)$, meaning that F becomes a transformation that maps t to s, i.e. $F : t \rightarrow s$. Accordingly, $\lim_{t \to -\infty} s(t) = 0$ and $\lim_{t \to \infty} s(t) = 1$ and s can be intuitively interpreted as the fraction of the distribution of t with values smaller than t from the definition of the CDF. This implies that the probability distribution of s, $g(s)$, is constant in the interval $s \in [0, 1]$ in which s is defined and s can be interpreted as the cumulative distribution of its own probability distribution:

$$s = G(s) = \int_{-\infty}^{s} g(s')ds' \tag{8}$$

In case of discrete probability functions such as the negative binomial function, the same argument still holds but the definition of the inverse quantile function is replaced by the generalized inverse: $F^{-1}(y) = \inf \{x : F(x) > y\}$ for $y \in [0, 1]$, see e.g. [**?**, p. 54]. In order to obtain a uniform distribution for discrete PDFs that is comparable to the case of continuous distributions, the histogram holding the inverse quantiles or values of the CDF is filled using random numbers according to the intervals of the CDF. For example, if the sales of zero items accounts for 75% of the observed sales distribution for this item, the value of the CDF function that is used to fill the histogram is randomly chosen in the interval $[0, 0.75]$. Proceeding similarly for all other observed values, the resulting histogram of CDF values should again be uniform as in the case of a continuous PDF.

A histogram of the quantiles (or PIT values) of each prediction is therefore expected to be uniformly distributed in $[0, 1]$ if the predicted probability distribution $f(x)$ is correctly calibrated.

This is illustrated in Fig. 2, which shows the distribution of quantiles (or PIT values) for five different cases. If both the choice of the model and the model parameters are estimated correctly, we would expect the uniform distribution as shown in the last part of the figure. The first two plots of Fig. 2 indicate a bias in the variance estimation and the third and fourth plot indicate a bias in the mean estimation.

4.1.2. Inverse Quantile Plot

Evaluating the full probability function in the simplest case, where we have only one model that can be used to predict all observations, is straight-forward: We could fill a histogram of all observed values, such as sales records, and overlay this with the single model, such as a negative binomial with predicted parameters, that is used for all observations. Then, we could directly compare the model curve with the observations, using statistical tests such as the Kolmogorov-Smirnov test. In practical applications however, we have a large number of effective prediction models, since although we always use the same model parametrization, such as the negative binomial distribution, its parameters have to be determined at the required level of granularity. For example, for daily orders, we need to predict the parameters of the negative binomial distribution for each location, sales day, and product. Unlike the simple case discussed initially, where we had many observations to compare the prediction model to, we now have just one observation per prediction, meaning that we cannot use statistical tests directly. Instead we use the fact that a correctly calibrated, predicted PDF should return a uniform cumulative density distribution *regardless* of the shape of the predicted distribution.

We therefore proceed as follows: First, each individual predicted probability density distribution is transformed to its cumulative density distribution. In the example of the granularity from above, we would then have a CDF for each predicted PDF for each product, sales day, and sales location. The CDF values, also known as quantiles, of the corresponding truth values are then compared (in the sense of higher or lower) to expected quantile values and accordingly filled to a collection of profile plots, each of which representing one expected quantile value.

Cyclic Boosting addresses the challenge of prediction explainability on a fundamental level. Rather than relying on black box approaches, individual predictions $\hat{y}_i$ for single observations should not only be accurate but also explainable. Strength and significance of each feature variable contributing to the prediction can be explained using profile plots.
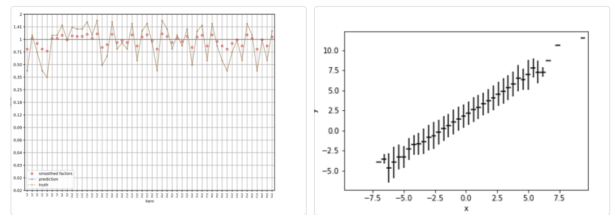

Figure 3:

Figure 3 shows analysis of the profile plots for feature variable "item" after the final iteration. The shown data points indicate mean values of prediction $\hat{y}$ and ground truth y for each bin divided by the global mean (prediction hardly visible due to good agreement). Note that the smoothed factors shown here are influenced by correlations to all other features in the model as well. This visualization directly indicates possible deviations from the optimal fit results in the different bins. Here, no significant deviations are present across the whole range of values. Furthermore, the smoothed values of the factors, i.e. the actual fitted parameters of the model, are shown. These

Figure 4: ...

Figure 5: Marginal distribution and scatter plot of variables X and Y.

differ from the mean values of the target and prediction in the different bins divided by the global mean due to correlations with other features. Right hand side plot in Figure 3 indicate values of prediction $\hat{y}$ and ground truth y where maximum points should fall on diagonal. The idea here is that the individual PDF predictions are first transformed into individual CDF predictions and the CDF values, also known as quantiles, of the corresponding truth values are then compared (in the sense of higher or lower) to expected quantile values and accordingly filled to a collection of profile plots, each of which representing one expected quantile value.

Fig. 4 illustrates five different collections of inverse quantile profile plots (each collection comparing to 7 expected quantile values), each with only one bin for the sake of visualization, for separate sets of exemplary probability density estimation and observed data combinations. Each shown line represents the percentage of probability density estimation and observed data combinations for which the observed data point should be above and below the quantile of the predicted PDF indicated by that line (for example, the 0.50 line, corresponding to the median, indicates that 50 percent of all probability density estimation and observed data combinations in a given set of data should fall above the line, and 50 percent should fall below the line), respectively. The observation of the number of samples, indicated with shaded circles, that do in fact fall above and below a particular line, then allows the evaluation of the accuracy of probability density estimations. For a set of accurate PDF predictions, one expects a uniform shape, like illustrated by the final far-right column of shaded circles. Here all shaded circles are exactly falling on the line which indicates that both the mean and the variance estimations as well as the choice of a negative binomial distribution as underlying PDF were correct. Conversely, the leftmost two columns of shaded circles illustrate a case for which the variance estimation has a bias and the center and center-right columns of shaded circles illustrate a case for which the mean estimation has a bias.

The two advantages of this method, as compared to the simpler method presented in section 4.1.1, are that an inverse quantile plot supports the qualitative evaluation of the predicted individual PDFs not only globally but (1) for different specified quantiles (potentially hinting to deviations in e.g. the tails of the distributions) and (2) in dependence of arbitrary variables of the data set (potentially hinting to deviations in e.g. specific categories of a feature). Two examples for this can be found in figures ?? and ?? in the next section.

4.2. Quantitative Evaluation of PDF Predictions

The methods described so far allow a detailed qualitative evaluation of PDF predictions. However, in order to also quantify the quality of the PDF predictions, a measure of the deviation of the PDF predictions from the optimal outcome given the observed target data is needed. To achieve this, we compare the CDF histogram of the predicted PDFs with the uniform distribution and define a prediction accuracy interval between 0 and 1.

Several different methods can be used to compute the deviance between two probability distributions. Good choices are the **Wasserstein metric** [?], a distance function defined between two probability distributions on a given metric space (also known as earth mover's distance), which is defined by

$$W_p(\mu, \nu) = \left(\inf \mathrm{E}\left[d(X, Y)^p \right] \right)^{1/p}, \qquad (9)$$

(where $E[Z]$ denotes the expected value of a random variable Z and the infimum is taken over all joint distributions of the random variables X and Y with marginals μ and ν respectively) and the **Kullback-Leibler divergence** [?], a measure of how one probability distribution diverges from a second expected probability distribution. For probability distributions P and Q defined on the same probability space, $\mathcal{X}$, the Kullback–Leibler divergence from Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)} \right) \qquad (10)$$

The first has the advantage that, unlike the second, it takes an underlying metric space into account, meaning that it depends on the distance of potential deviations.

5. Example: Demand Forecasting

... predict an individual demand volume for different product-location-date combinations ...

... example showing full prediction and evaluation chain ...

... plot showing individual PDF example ...

... plot showing cdf histo ...

... plot showing invquant profile for mean prediction on X-axis ...

... plot showing invquant profile for product groups on X-axis ...

A. Profile Histograms

In many cases, scatter plots are used to study the behaviour of two distributions or sets of data points visually. However, even for moderate amount of data, this approach becomes quickly difficult. To illustrate this, a sample of (x, y) data points was obtained in the following way: The distribution of x values was obtained by generating 5,000 samples of Gaussian distributed random numbers $X \sim \mathcal{N}(0.0, 2.0)$ and the y values are obtained via $Y \sim X + \mathcal{N}(2.0, 1.5)$. Fig. 5 shows the marginal distributions for x and y as well as a scatter plot of x vs. y.

Figure 6: Profile histogram of variables X and Y.

Although the simple linear correlation between X and Y is apparent in the scatter plot, finer details are not visible and it is easy to imagine that a more complex relationship is difficult to discern. Profile histograms are specifically designed to address this shortcoming. Intuitively, profile histograms are a one-dimensional representation of the two-dimensional scatter plot and are obtained in the following way: The variable on the x axis is discretizised into a suitable range of bins. The exact choice of binning depends on the problem at hand. One can for example choose equidistant bins in the range of the x axis or non-equidistant bins such that each bin contains the same number of observations. Then within each bin of the variable X, the a location and dispersion metric is calculated for the variable Y. This means that the bin-borders on the X axis are used as constraints on the variable Y and with these conditions applied, for example the sample mean of the selected y values as well as the standard deviation are calculated. These location and dispersion metrics in each bin of X are used to illustrate the behaviour of the variable Y as the values of the variable X change from bin to bin. The resulting profile histogram is shown in Fig. 6. This one-dimensional representation allows to understand even a complex relationship between two variables visually. Note that due to few data points at the edges of the distributions the profile histogram is expected to show visual artifacts in the corresponding regions.