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**

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 observed realized sales and not the actual demand, in case the demand exceeds the current stock level, the data become censored. This 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 resulting in 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 in addition to the 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 from 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 probabilty 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. 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 the following 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. The latter approach is taken e.g. in [1–3] and aims to avoid estimating the underlying probability distribution. Although this approach seems preferable since it avoids determining the full demand distribution and results directly in the desired operational decision (order quantity), it faces several drawbacks. First of all, the full probability distribution for the demand of an 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 accute assessment of the forecast quality at all quantiles of the distribution, including the often extensive tails of the distribution, as well as the analysis of long-term effect as the demand predictions including all future planned advertisements or price-changes can be included into long-term forecasts. Furthermore, deriving the ordering decision directly couples the demand process with the complex delivery cycles and constraints, 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 wihtout 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.

In contrast, more traditional inventory control systems rely on the knowledge of the demand distribution in one form or another. see e.g. [4] for an overview. In (s, S) type inventory control systems [5], 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 [6], although perishable goods pose more challenges, see e.g. [7–9]. Additionally, service level constraints can be included in these kind of inventory control systems [10]. Perishable goods are well described by the ''newsvendor-problem" [11] 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. [12]. Assuming linear underage and overage costs $b, h > 0$, the optimal quantile $q_{\text{opt}} = b/(b + h)$ 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. [13]) or via sample average approximation (SAA). (see e.g. [14] for an overview over SAA). 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, advertisements, etc. are highly correlated with demand. Additionally, it is critical that we indeed reconstruct a demand distribution, hence a simple point-estimator as provided by the most common statistical techniques and machine-learning approaches will not suffice. Additionally, demand is not identically and independently distributed (i.i.d.) but depends not only on external factors such as season, weather, product life-cycle, but is also actively changed by the retailer by setting a specific price, offering rebates or running advertisements. Additionally, the demand implicitly depend 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 characersitics, pricing, advertisements, retail location details, etc. into account. This can be done in several ways: First, we can use a neural network [15] to learn the distribution from data and return a full distribution per item, store and day from which the relevant quantile can be estimated. Similarly, using quantile regression [16] this approach can be implemented in different frameworks. 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 compared to the case of the full distribution. Empirically, one can determine the best fitting distribution from data [17]. 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 [18–22]. 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 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

What we are going to show.

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. 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.

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 case of demand forecasting, each prediction is highly specific to the circumstances in which it is used (such as product ID, day 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 [23]. The major benefits of this algorithm are that it is not only extremly 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 discretisized appropriately into k bins to reflect the specific behaviour of the feature. The global mean μ is determined from all observed 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 overtraining and improve the generalization ability of the algorithm. The relative strength of the factors f_j^k are then directly interpretable in relation how important a specific feature is for each individual prediction where deviations from $f_j^k = 1$ indicate high importance.

The assigned mean and variance predictions can then be used to generate individual probability density functions according to a functional negative binomial distribution assumption for each sample.

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 [24]:

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

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

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.

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

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 $[1, \infty]$.

Hm, this part is a bit quick. We should add a sentence to the NB loss and how this is motivated and then how this is connected to Eqn. (4) and how Eqn. (1) from CB comes into play.

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

Hm, this part is not quite clear - need to add some details what is meant As described in [23], 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 presented in [25], 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.

$$\sigma^2 = \mu + \frac{\mu^2}{r} \tag{5}$$

4. Evaluation of PDF Predictions

Statistical or machine learning methods that predict full individual probability functions typically lack quantitative, or at least qualitative, evaluation of the PDF model output to assess its correctness. In the case of an estimation of the determining parameters of an assumed functional form for the PDF, assessing the correctness of the PDF

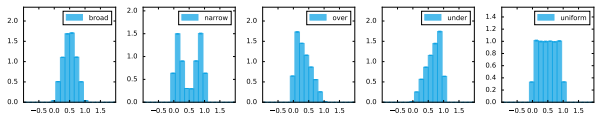

Figure 1: ...

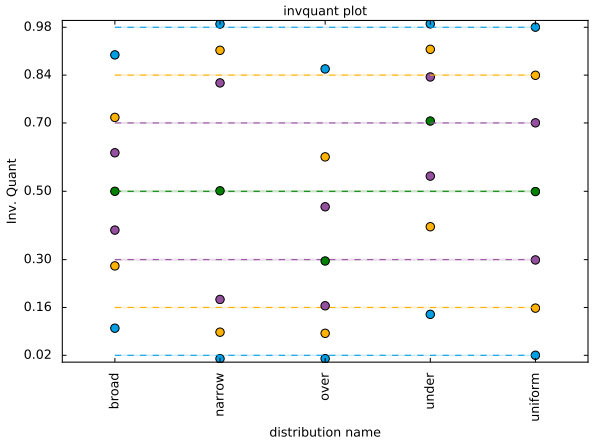

Figure 2: ...

model output refers to the evaluation of the accuracy of the prediction of the different determining parameters, e.g. mean and variance of a negative binomial distribution, and the validation of the model assumptions, i.e. the choice of the underlying PDF, by comparing the probability density estimations to observed data. However, the two methods presented in this section, namely cumulative distribution function (CDF) histograms and inverse quantile plots, are neither restricted to the evaluation of parametric PDF estimations nor to a specific functional form, like a negative binomial, but can be used generically.

4.1. Qualitative Evaluation of PDF Predictions

4.1.1. CDF Histogram

need to explain intuition a bit more, a diagram would be helpful. 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 filled to a histogram.

need to explain a bit more what we see and why we expect a uniform distribution. Figure (1) illustrates five different CDF histograms. If the PDF prediction is correct, a flat uniform distribution should occur for the CDF histogram, like illustrated by the final far-right column of shaded circles.

In the example of a parametric negative binomial PDF estimation of section (3), this would indicate 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 first two plots of Fig. 1 indicate a bias in the variance estimation and the third and fourth plot indicate a bias in the mean estimation. Need to explain why we see that. Would be good to show the respective distributions (or at least one example) of wher it doesn't fit.

4.1.2. Inverse Quantile Plot

We need to explain a lot more here. In particular, outside HEP the profile plots are not known so we need to explain that first and then explain why we want to use them here. 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.

since no reader of this paper will have seen a profile plot before or know what that is, we need to guide the audience step-by-step to what they should see and why it's correct and what each column means, how we come to the conclusion, etc. Fig. 2 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 (median?for example, the 0.50 line 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.

need to explain how one can see this In the example of a parametric negative binomial PDF estimation of section 3, this would indicate 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** [26], a distance function defined between two probability distributions on a given metric space (also known as earth movers distance), and the **Kullback-Leibler divergence** [27], a measure of how one probability distribution diverges from a second expected probability distribution. 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 ...

...

References

[1] A.-L. Beutel and S. Minner, "Safety stock planning under causal demand forecasting," *International Journal of Production Economics*, vol. 140, no. 2, pp. 637–645, 2012.

[2] G.-Y. Ban and C. Rudin, "The big data newsvendor: Practical insights from machine learning," *Operations Research*, vol. 67, no. 1, pp. 90–108, 2019.

[3] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Management Science*, vol. 66, no. 3, pp. 1025–1044, 2020.

[4] E. A. Silver, D. F. Pyke, R. Peterson, *et al.*, *Inventory management and production planning and scheduling*, vol. 3. Wiley New York, 1998.

[5] H. Scarf, *A min-max solution of an inventory problem.* Studies in The Mathematical Theory of Inventory and Production, Stanford University Press, 1958.

[6] H. Scarf, "The optimality of (s,s) policies in the dynamic inventory problem," *Mathematical Methods in the Social Sciences*, 1959.

[7] S. Nahmias and W. P. Pierskalla, "Optimal ordering policies for a product that perishes in two periods subject to stochastic demand," *Naval Research Logistics Quarterly*, vol. 20, no. 2, pp. 207–229, 1973.

[8] S. Nahmias, "Optimal ordering policies for perishable inventoryii," *Operations Research*, vol. 23, no. 4, pp. 735–749, 1975.

[9] S. Nahmias, "The fixed-charge perishable inventory problem," *Operations Research*, vol. 26, no. 3, pp. 464–481, 1978.

[10] S. Minner and S. Transchel, "Periodic review inventory-control for perishable products under service-level constraints," *OR spectrum*, vol. 32, no. 4, pp. 979–996, 2010.

[11] F. Edgeworth, "The mathematical theory of banking," *Journal of the Royal Statistical Society*, 1888.

[12] M. Khouja, "The single-period (news-vendor) problem: literature review and suggestions for future research," *Omega*, vol. 27, no. 5, pp. 537 – 553, 1999.

[13] L. C. Alwan, M. Xu, D.-Q. Yao, and X. Yue, "The dynamic newsvendor model with correlated demand," *Decision Sciences*, vol. 47, no. 1, pp. 11–30, 2016.

[14] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory.* SIAM, 2014.

[15] M. Feindt and U. Kerzel, "The neurobayes neural network package," *NIM A*, vol. 559, no. 1, pp. 190 – 194, 2006.

[16] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[17] I. Adan, M. van Eenige, and J. Resing, "Fitting discrete distributions on the first two moments," *Probability in the engineering and informational sciences*, vol. 9, no. 4, pp. 623–632, 1995.

[18] A. S. C. Ehrenberg, "The pattern of consumer purchases.," *Journal of the Royal Statistical Society Series C*, no. 1, p. 26, 1959.

[19] G. J. Goodhardt and A. Ehrenberg, "Conditional trend analysis: A breakdown by initial purchasing level," *Journal of Marketing Research*, vol. IV, pp. 155–161, May 1967.

[20] A. Ehrenberg, *Repeat-buying; theory and applications.* North-Holland Pub. Co., 1972.

[21] C. Chatfield and G. J. Goodhardt, "A consumer purchasing model with erlang inter-purchase time," *Journal of the American Statistical Association*, vol. 68, no. 344, pp. 828–835, 1973.

[22] D. C. Schmittlein, A. C. Bemmaor, and D. G. Morrison, "Technical notewhy does the nbd model work? robustness in representing product purchases, brand purchases and imperfectly recorded purchases," *Marketing Science*, vol. 4, no. 3, pp. 255–266, 1985.

[23] F. Wick, U. Kerzel, and M. Feindt, "Cyclic boosting - an explainable supervised machine learning algorithm," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, IEEE, Dec. 2019.

[24] J. Hilbe, *Negative binomial regression*. Cambridge, UK New York: Cambridge University Press, 2011.

[25] F. Wick, U. Kerzel, and M. Feindt, "Cyclic boosting - an explainable supervised machine learning algorithm," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 358–363, 2019.

[26] I. Olkin and F. Pukelsheim, "The distance between two random vectors with given dispersion matrices," *Linear Algebra Appl.*, vol. 48, pp. 257–263, 1982.

[27] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 03 1951.

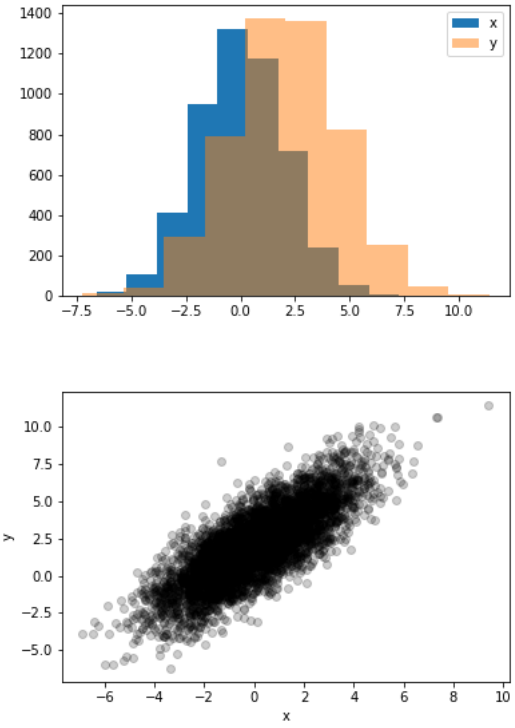

Figure 3: Marginal distribution and scatter plot of variables X and Y.

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. 3 shows the marginal distributions for x and y as well as a scatter plot of x vs. 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 shortcomming. 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. 4. 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.

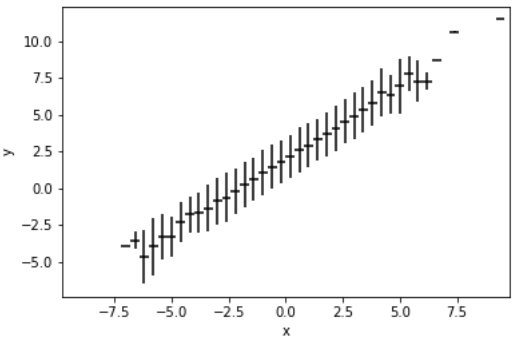

Figure 4: Profile histogram of variables X and Y.