International Journal of Forecasting 32 (2016) 788–803

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Evaluating predictive count data distributions in retail sales forecasting

Stephan Kolassa

SAP Switzerland, Bahnstrasse 2, 8274 Tägerwilen, Switzerland

A R T I C L E I N F O

Keywords:
Demand forecasting
Density forecasting
Error measures
Intermittent demand
Proper scoring rules

A B S T R A C T

Massive increases in computing power and new database architectures allow data to be stored and processed at finer and finer granularities, yielding count data time series with lower and lower counts. These series can no longer be dealt with using the approximative methods that are appropriate for continuous probability distributions. In addition, it is not sufficient to calculate point forecasts alone: we need to forecast entire (discrete) predictive distributions, particularly for supply chain forecasting and inventory control, but also for other planning processes. However, tools that are suitable for evaluating the quality of discrete predictive distributions are not commonly used in sales forecasting. We explore classical point forecast accuracy measures, explain why measures such as MAD, MASE and wMAPE are inherently unsuitable for count data, and use the randomized Probability Integral Transform (PIT) and proper scoring rules to compare the performances of multiple causal and noncausal forecasting models on two datasets of daily retail sales.

1. Introduction

Count data are integer-valued time series. Count data are especially important for supply chain forecasting, where most products are sold in units. A sufficiently fine granularity along any of the typical dimensions in a demand hierarchy (location, product or time) turns count data into *intermittent demand*, i.e., data that exhibit "many" zeros. For instance, most of the demand series that a typical supermarket faces are intermittent at a store × SKU × day level, though of course the series become non-intermittent with aggregation along any dimension.

High-volume count data, e.g., sufficiently highly aggregated data, can and usually will be forecasted using methods that, strictly speaking, are only valid for continuous data, since the relative error between the underlying count data and the continuous approximation will become small with higher volumes, and many statistical theorems relying on large-sample theory, such as the Central Limit Theorem, can be applied.

However, in recent years the trend has been towards "Big Data", i.e., the collection and processing of more and more data. Modern database systems allow data of finer and finer granularities to be stored (Januschowski, Kolassa, Lorenz, & Schwarz, 2013). Operational forecasting in retail, e.g., for store replenishment, is already conducted not at highly aggregated levels, but at the most fine-grained level possible; that is, for count data and intermittent demands. As a consequence, count data and intermittent demand forecasting form an active research area, and one that will become even more important in the future; new forecasting algorithms are being developed constantly, together with variations on old ones.

However, so far, the literature on count data and intermittent demand forecasting has focused on point forecasts instead of entire predictive distributions, as has been the norm for many years now in macroeconomic or financial forecasting, for example. We argue that this

E-mail address: Stephan.Kolassa@sap.com.

http://dx.doi.org/10.1016/j.ijforecast.2015.12.004
0169-2070/

emphasis on point forecasts misses the mark. Instead, we should strive to understand the entire predictive distribution, even for count data, from which we can extract point, interval or quantile forecasts as desired. For example, quantile forecasts are obviously necessary in supply chain forecasting for setting safety amounts, but also for scenario analyses in promotional or other forecasts.

The question now arises as to how different forecasting methods should be compared. For this, one needs forecast accuracy measures, or equivalently, error measures. Count data pose specific challenges for error measures. In particular, minimizing common error measures does not necessarily lead to the "best" forecasting method for count data, especially for intermittent demand series. In addition, while the use of the Probability Integral Transform (PIT) and scoring rules for evaluating continuous predictive distributions is well established, their discrete counterparts have not been used in sales forecasting so far.

This paper is organized as follows: first, we discuss classical point forecast accuracy measures and their problems for count data. Next, we examine a randomization-based modification of the standard PIT, as well as proper scoring rules that are applicable to discrete data. Then, we apply multiple discrete models to two count datasets and evaluate the predictive distributions using the randomized PIT and scoring rules. Finally, we close with possibilities for future research.

2. Point forecast accuracy measures

2.1. Measures based on absolute errors

Assume that the future data generating process that we wish to forecast follows an estimated predictive density $\widehat{f}$, and let us assume for now that $\widehat{f} = f$ is indeed the true distribution f of future actuals; i.e., our predictive density is correctly specified. It is well known that using the expected value of f as a point forecast will minimize the expected squared error. It is almost as well known (and is an easy derivation to show, cf. Hanley, Joseph, Platt, Chung, & Belisle, 2001; Schwertman, Gilks, & Cameron, 1990) that using the *median* of f as a point forecast will minimize the expected *absolute* error.

Let us turn this argument around: the forecast which minimizes (some estimator of) the expected absolute error will estimate not the expected value of f, but its median. In particular, ranking or selecting forecasting methods, or optimizing the parameters of a forecasting method, based on the mean absolute deviation (MAD) will reward point forecasts for yielding not the future expected value, but the future median realization (Morlidge, 2015).

This distinction makes no difference in the case of a *symmetric* predictive distribution f. However, the predictive distributions that are appropriate for low volume count data (Syntetos, Babai, Lengu, & Altay, 2011) are usually far from symmetric, and this distinction does make a difference in such cases. The fact that optimizing the MAD yields biased forecasts in this case, or even EMAD-optimal forecasts that are constant at zero, has been recognized before (Prestwich, Rossi, Tarim, & Hnich, 2014; Snyder, Ord, & Beaumont, 2012; Teunter & Duncan, 2009; Wallström & Segerstedt, 2010), but Morlidge (2015) appears to have been the first to explicitly note the connection to the median of f.

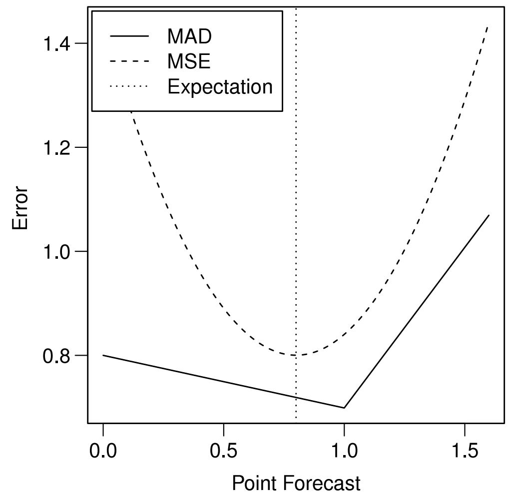

Fig. 1. MADs and MSEs of various point forecasts for $y_1, \dots, y_{10^7} \sim \text{Pois}(0.8)$. The MSE is minimized by the expectation of 0.8, while the MAD is minimized by the median of 1.

As an example, assume that $f = \text{Pois}(\lambda)$ for $\lambda < \log 2 \approx 0.693$. In this case, the median of f is 0, whereas its expectation is λ. The EMAD-optimal point forecast is 0, regardless of whether $\lambda = 0.01$, $\lambda = 0.1$ or $\lambda = 0.5$. Thus, an EMAD-optimal point forecast will be biased downward. Similarly, if $\log 2 < \lambda < \lambda_0$, where $\lambda_0 \approx 1.678$ satisfies $\lambda_0 e^{-\lambda_0} + e^{-\lambda_0} = \frac{1}{2}$, then the EMAD-optimal point forecast will be 1, which is biased upward for $\log 2 < \lambda < 1$ and downward for $1 < \lambda < \lambda_0$.

We note that the preceding example does not in any way presuppose that we *constrain* the point forecast to be integers. (Indeed, an unbiased point forecast would of course simply be λ itself.) Rather, the argument is that, as described above, the point forecast that minimizes the EMAD is the median of the future distribution—and the median of a Poisson-distributed variable turns out to be integer. Therefore, if our aim is to minimize the EMAD of a Poisson predictive distribution, this cost function automatically draws us towards an integer-valued point forecast, namely the median, which will typically be biased.

To illustrate this point, we calculated the MAD and MSE values of various (not necessarily integer) point forecasts for 10^7 simulated Poisson distributed random variables with parameter $\lambda = 0.8$. As expected, the MSE is minimized by a point forecast of 0.8, i.e., the expectation, while the MAD is minimized by a biased point forecast of 1, the median of the Pois(0.8) distribution (Fig. 1).

Exactly the same argument applies to all accuracy measures that are multiples of the MAD by a factor that does not depend on the forecast itself. For instance, the weighted mean absolute percentage error (wMAPE) is obtained by dividing the MAD by the mean of the out-of-sample realizations (Kolassa & Schütz, 2007), and the mean absolute scaled error (MASE Franses, 2016; Hyndman, 2006; Hyndman & Koehler, 2006) is obtained