Assessing predictive count data distributions

Stephan Kolassa

Swiss Data Science Day, September 16, 2016

Public

Overview

- Context: predicting large numbers of (grouped) low-count numerical target values
  - For example: retail sales at store × stock-keeping unit × day granularity – necessary for replenishment
  - Could be in a time series forecasting context, or not
- How can we assess whether our predictions are good?

Mean Squared Error	MSE	$\frac{1}{N}\sum(y_i - \hat{y}_i)^2$
Mean Absolute Error/Deviation	MAE/MAD	$\frac{1}{N}\sum\lvert y_i - \hat{y}_i\rvert$
Mean Absolute Percentage Error	MAPE	$\frac{1}{N}\sum\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$
weighted Mean Absolute Percentage Error	wMAPE	$\frac{\sum\lvert y_i - \hat{y}_i\rvert}{\sum y_i}$
Mean Absolute Scaled Error	MASE**	$\frac{\sum\lvert y_i - \hat{y}_i\rvert}{\sum\lvert y_i - y_{i-1}\rvert}$

** Note that the denominator for the MASE is calculated *in-sample* (Hyndman & Koehler, 2006)

- Please vote for your favorite point forecast accuracy measure in the event app!
- https://gameday.eu.doubledutch.me/?sessiontoken=b70fa787-7ac1-4c29-8a9f-c220f8ce6cd5&mod=polls&pollId=9441

Overview

- Context: predicting large numbers of grouped low-count numerical target values
  - For example: retail sales at store × stock-keeping unit × day granularity – necessary for replenishment
  - Could be in a time series forecasting context, or not
- How can we assess whether our predictions are good?

		"Robust"?	Defined?	Comparable*?	Intuitive?
MSE	$\frac{1}{N}\sum(y_i - \hat{y}_i)^2$	×	✓✓	××	×
MAE	$\frac{1}{N}\sum\lvert y_i - \hat{y}_i\rvert$	✓✓	✓✓	×	✓
MAPE	$\frac{1}{N}\sum\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$	✓	((✓))	✓	✓✓
wMAPE	$\frac{\sum\lvert y_i - \hat{y}_i\rvert}{\sum y_i}$	✓	(✓)	✓	✓✓
MASE**	$\frac{\sum\lvert y_i - \hat{y}_i\rvert}{\sum\lvert y_i - y_{i-1}\rvert}$	✓✓	✓	(✓)	×

* Comparability between groups/series on different levels, e.g., fast vs. slow selling products

** Note that the denominator for the MASE is calculated *in-sample* (Hyndman & Koehler, 2006)

- However, optimizing (some of) these can lead to systematically biased predictions! **Part 1**
- Better: predict & assess full densities! **Part 2**
- (Full paper here: http://dx.doi.org/10.1016/j.ijforecast.2015.12.004)

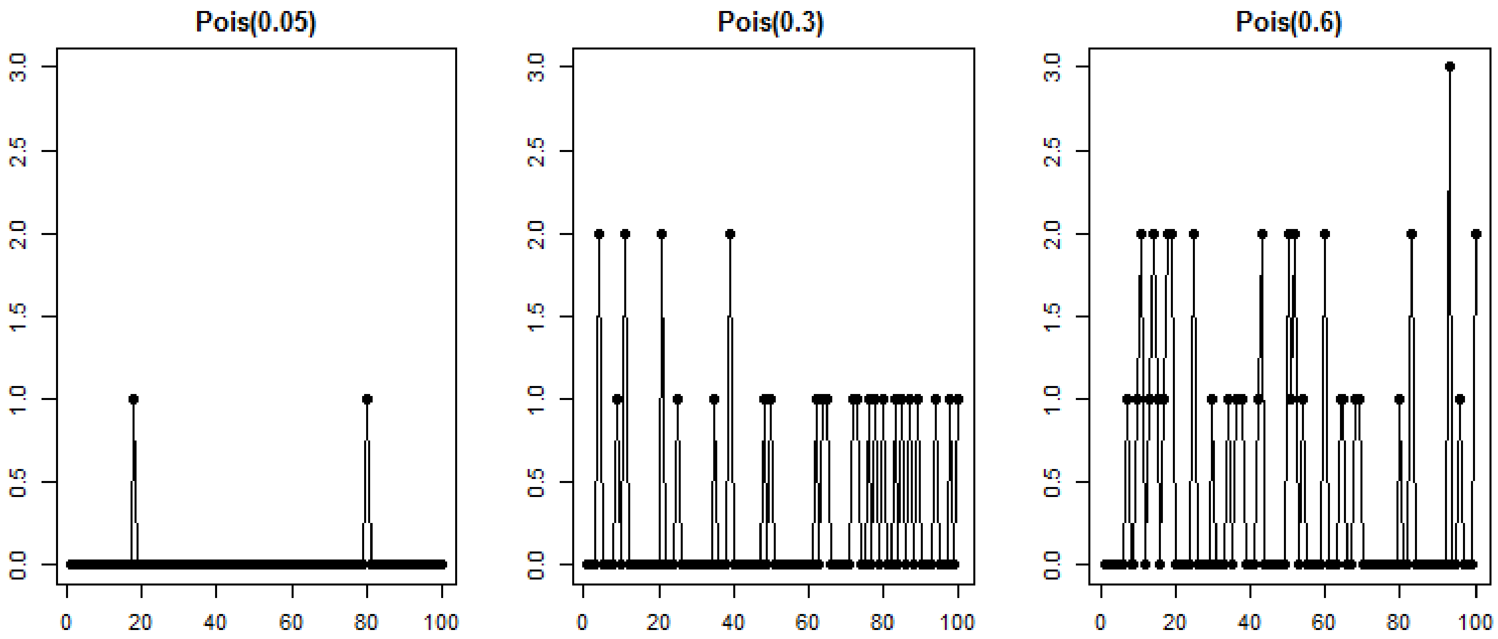
Means, medians and MADs

Part 1: Optimizing point forecast accuracy measures may yield biased predictions

- In summarizing *any* probability distribution…
  - The median minimizes the expected absolute error (Hanley et al., 2001)
  - The mean minimizes the expected squared error
- Translate this into predicting: given a (correctly specified) predictive distribution…
  - Predict the median to minimize the expected MAE/MAD
  - Predict the mean to minimize the expected MSE
- Turn this around:
  - If you optimize your forecast method or parameters to minimize MAE and the future distribution is skewed, *your forecast will be biased* (Morlidge, 2015)!
  - This is particularly relevant for intermittent series (which are usually skewed), but also for non-intermittent low volume count series
  - This also applies to the MASE (Hyndman & Kohler, 2006) and the wMAPE (Kolassa & Schütz, 2007), which are simply scalar multiples of the MAD

An example: forecasting Poisson time series

Part 1: Optimizing point forecast accuracy measures may yield biased predictions

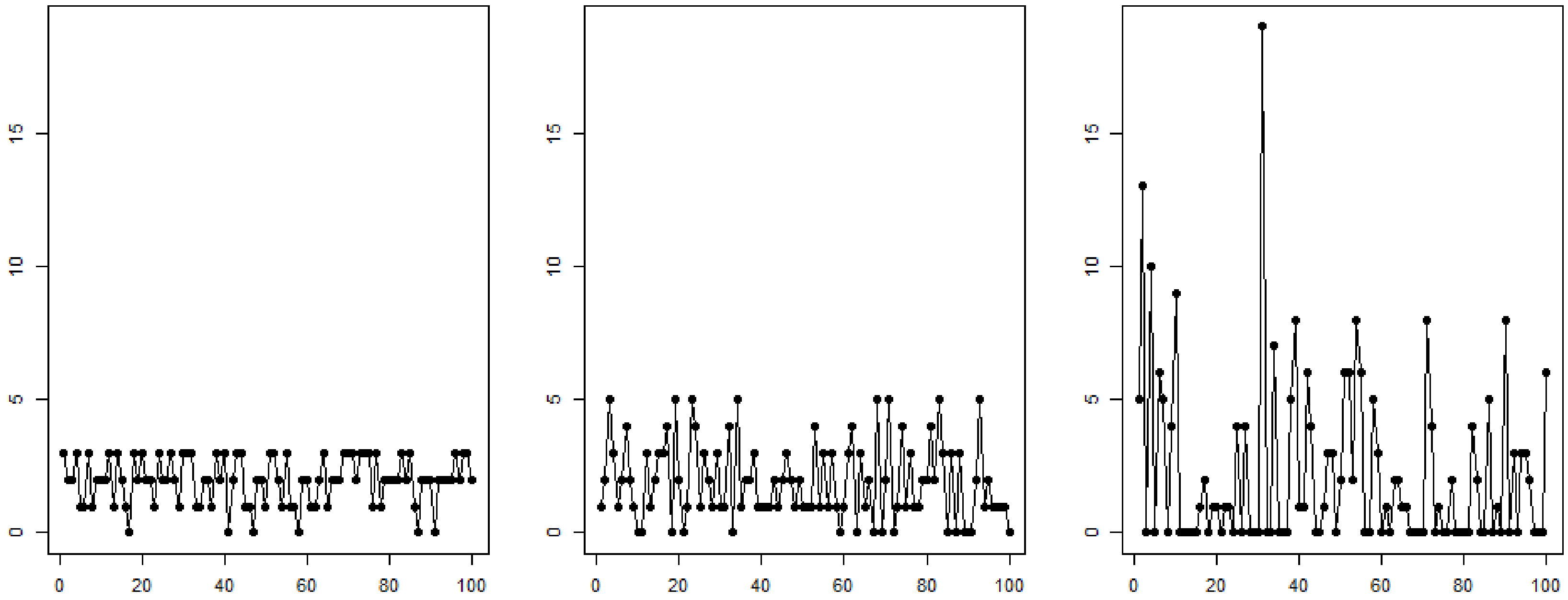


	Pois(0.05)	Pois(0.3)	Pois(0.6)
Forecast to minimize the expected MAD	0	0	0
Forecast to minimize the expected MSE	0.05	0.3	0.6

Better: forecast full predictive densities!

Part 2: Predict full predictive densities – how to assess?

- These simulated time series all have the same expectation of 2:



- Will a point forecast of 2 actually be useful?
- Not for setting safety amounts… or scenario planning…
- We need to forecast the full predictive density!
- But how do we *assess the quality* of a predictive density?

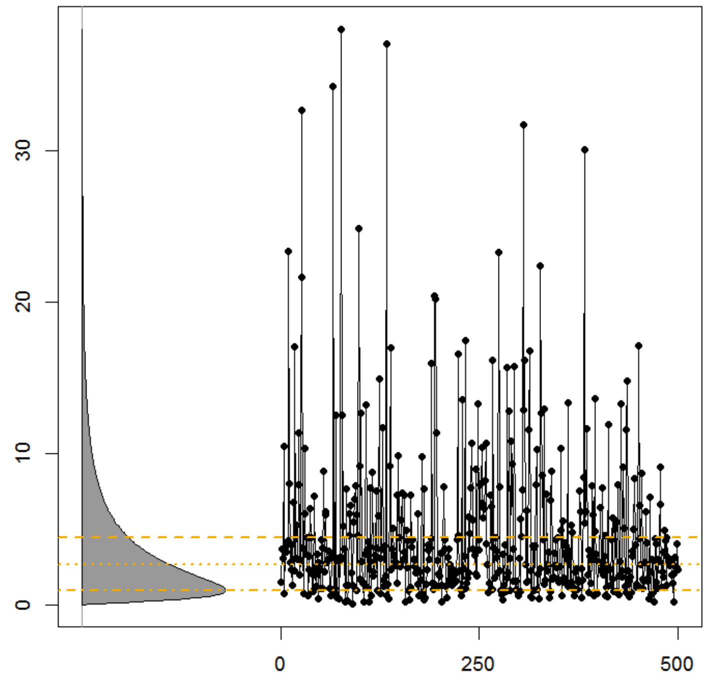
The Probability Integral Transformation (PIT)

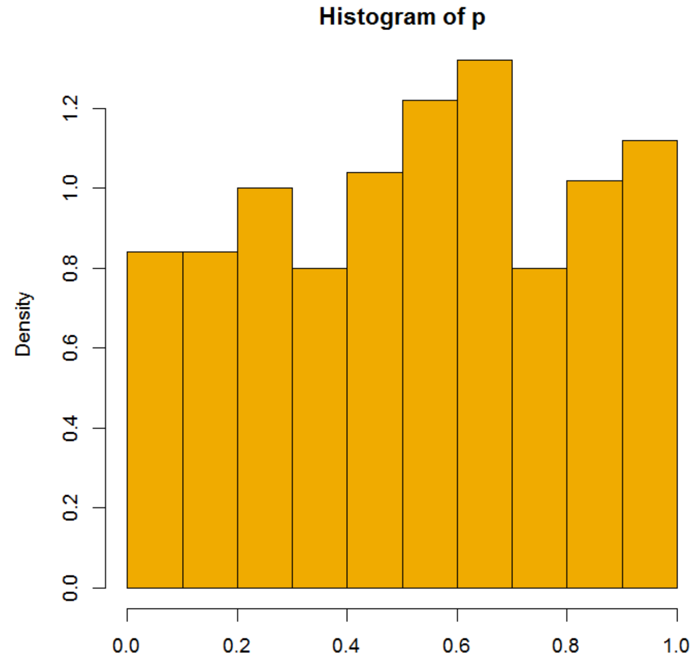
Part 2: Predict full predictive densities – how to assess?

- Assume predictive distributions with densities $\hat{f}_t$ and cumulative distribution functions $\hat{F}_t$
- Transform observations y_t:

$$p_t := \hat{P}_t(Y_t \leq y_t) = \hat{F}_t(y_t) = \int_{-\infty}^{y_t} \hat{f}_t$$

- If the predictive distributions are correct, $\hat{f}_t = f_t$ and $\hat{F}_t = F_t$, then $p_t \sim U(0,1)$ – this can be tested (e.g., Ledwina, 1994; Berkowitz, 2001; or others)





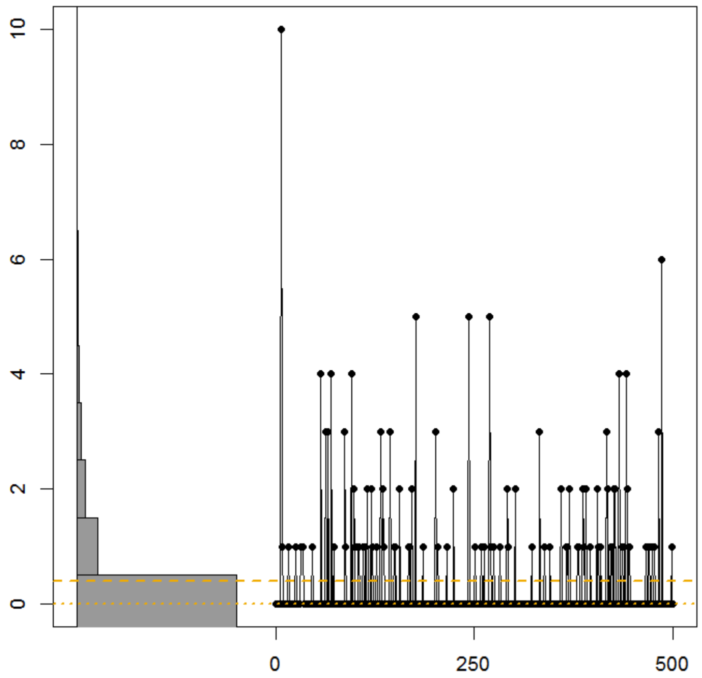
Problem: discrete predictive distributions

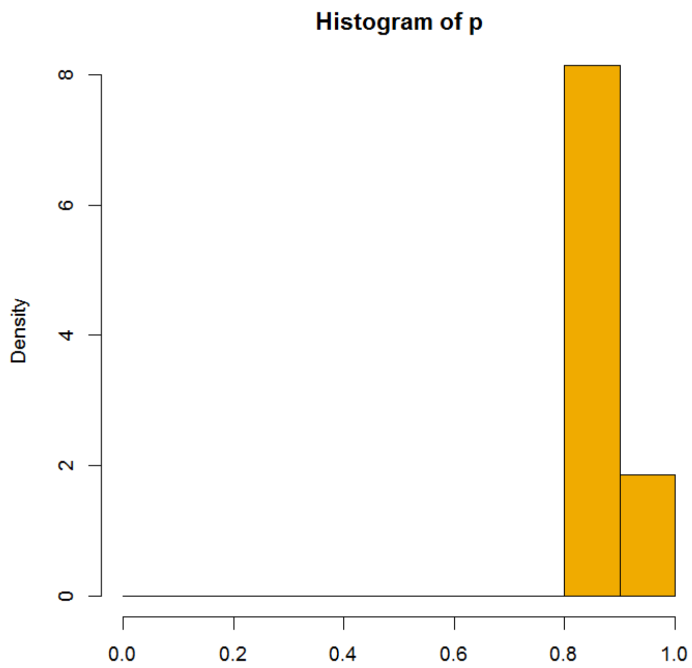
Part 2: Predict full predictive densities – how to assess?

- Transform observations y_t:

$$p_t \coloneqq \hat{P}_t(Y_t \leq y_t) = \hat{F}_t(y_t) = \int_{-\infty}^{y_t} \hat{f}_t$$

- Even if the predictive distributions are correct, p_t will have a discrete distribution if $\hat{F}_t = F$ is stationary!





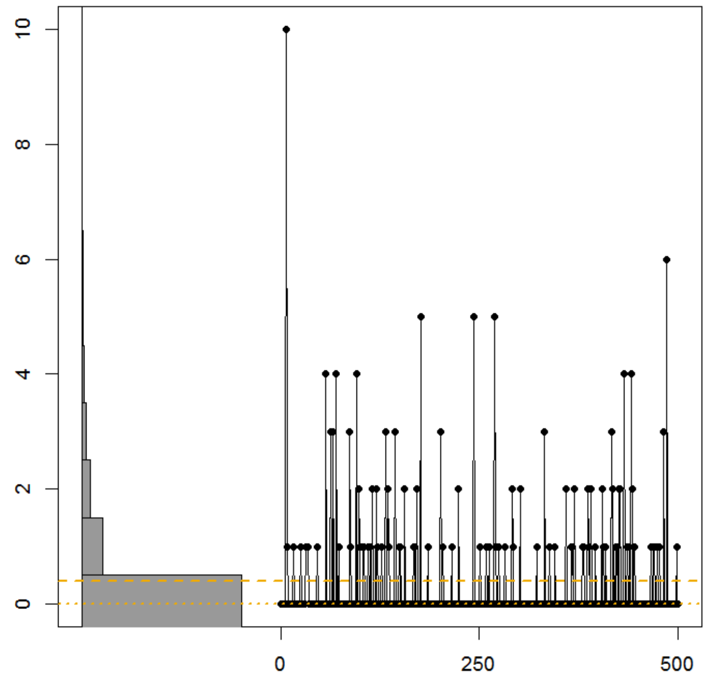
Solution: randomize the PIT

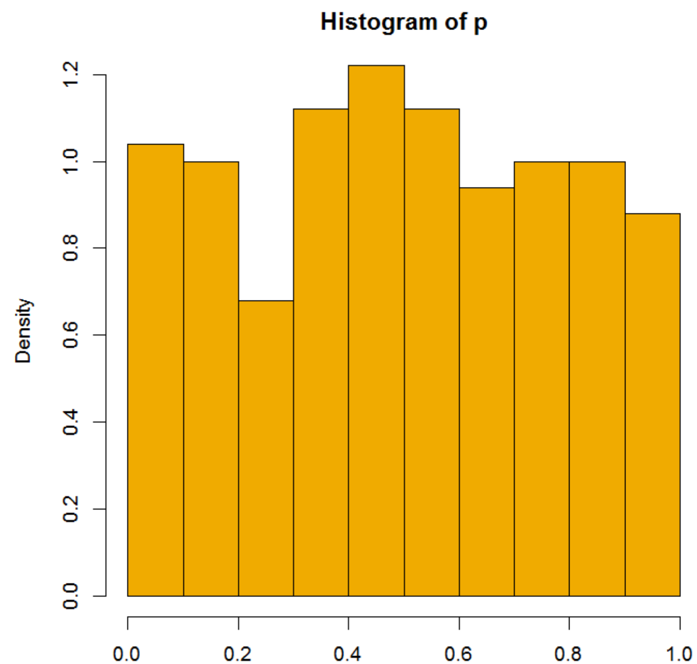
Part 2: Predict full predictive densities – how to assess?

- Brockwell (2007), Frühwirth-Schnatter (1996), Liesenfeld et al. (2006), Smith (1985)
- Set $\hat{F}_t(-1) \coloneqq 0$, and draw uniform p_t:

$$p_t \sim U\left(\hat{F}_t(y_t - 1), \hat{F}_t(y_t)\right)$$

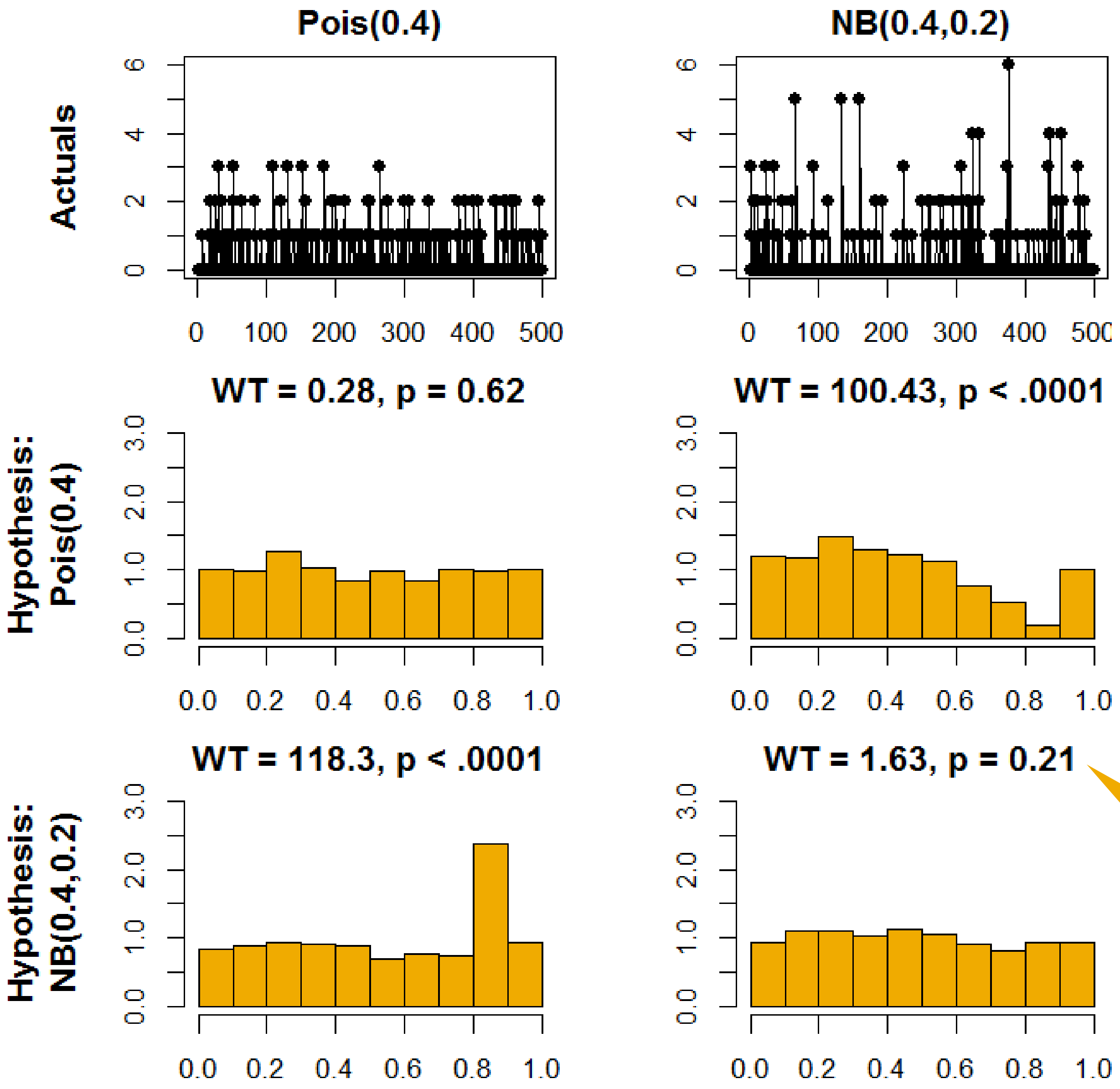
- Then $\hat{F}_t = F_t$ again implies that $p_t \sim U(0,1)$





Example: Poisson vs. negative binomial

Part 2: Predict full predictive densities – how to assess?



- Actuals and Hypotheses are either Pois(0.4) or NegBin(0.4,0.2)
  - In both cases, the median and MAD-optimal forecast is 0
  - In both cases, the expectation and MSE-optimal forecast is 0.4
  - Pois and NB differ heavily in the tails:

$$P_{\text{Pois}}(Y > 3) = 0.00078$$
$$P_{\text{NB}}(Y > 3) = 0.026$$
$$\frac{P_{\text{NB}}(Y > 3)}{P_{\text{Pois}}(Y > 3)} = 34.1$$

Data driven smooth tests for uniformity (Biecek & Ledwina, 2012):

test statistics WT and p values

How to apply this to grouped data, e.g. multiple time series?

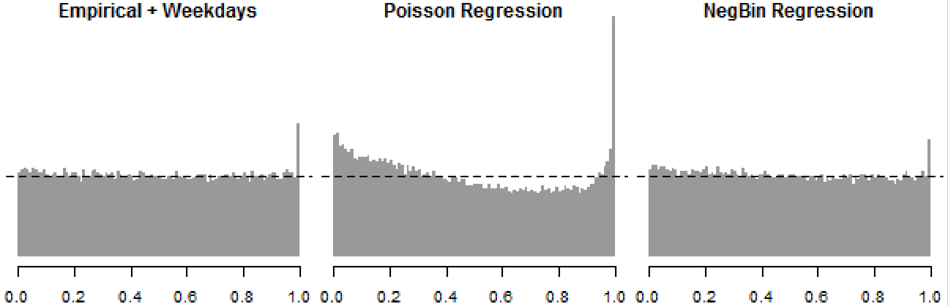
Part 2: Predict full predictive densities – how to assess?

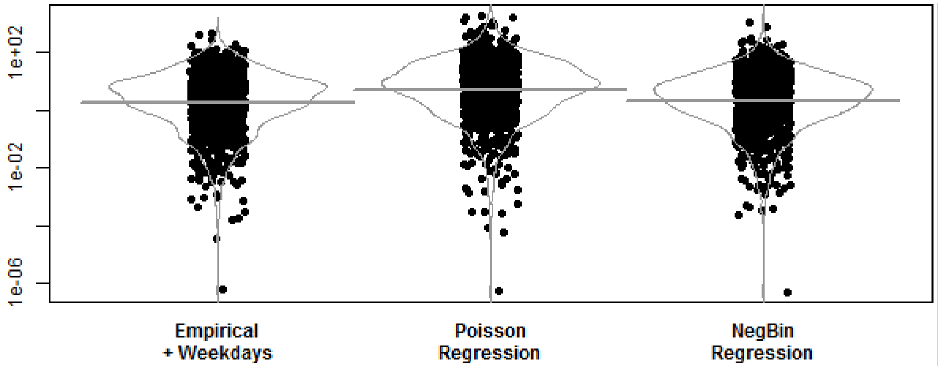
- Two possible summaries:
  - Simply “stack” all p values and test this big vector for uniformity
  - Test each series’ p values for uniformity, yielding a test statistic WT for each series – plot, summarize, compare these
- Illustration: two datasets with daily sales from European retailers
  - 1000 series each
  - Forecast horizon up to 100 days for each series
- Try multiple forecasting approaches – here, look at three:
  - Empirical + Weekday
    - Density forecast for next Friday is just the historically observed distribution of Friday sales
  - Poisson Regression
  - NegBin Regression
    - Regressions include day of week, price, trend and Christmas

Results

Part 2: Predict full predictive densities – how to assess?

Retailer A



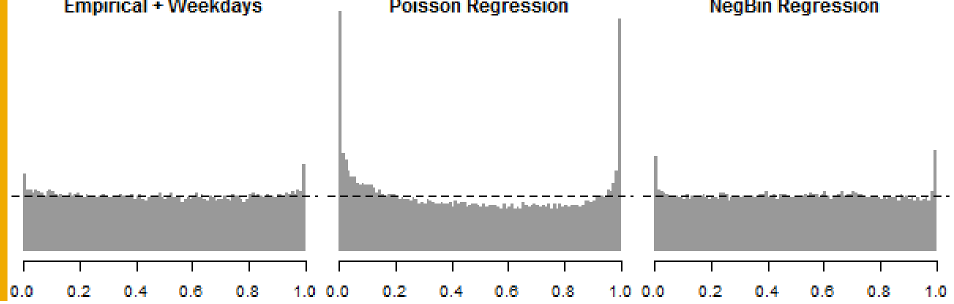


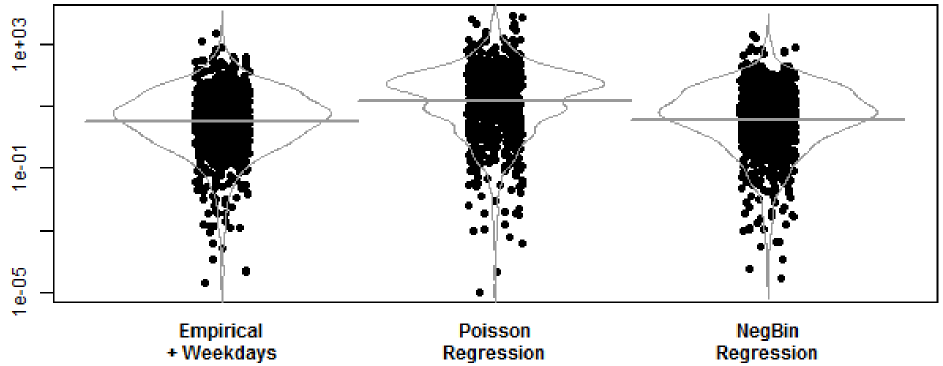
	Empirical + Weekdays	Poisson Regression	NegBin Regression
95%	48.1	117.6	43.3
Median	2.6	6.8	2.7
5%	0.0	0.1	0.0
Mean	10.9	35.9	12.2

rPIT p values

WT test statistics from data-driven smooth tests for uniformity

Retailer B





	Empirical + Weekdays	Poisson Regression	NegBin Regression
95%	101.5	512.1	110.4
Median	4.6	28.2	5.6
5%	0.0	0.1	0.0
Mean	26.9	142.2	30.8

- Poisson Regression obviously bad – does not capture overdispersion
- Empirical + Weekdays comparable to NegBin Regression
- The probability for high sales is *always* underforecast

Conclusion

- Do not rely on the MAE et al. to find an unbiased point forecast
  - If you do need to report MAE/MAPE/wMAPE/MASE, also report bias
  - For point predictions, use MSE, or RMSE, or a scaled RMSE that is comparable between scales

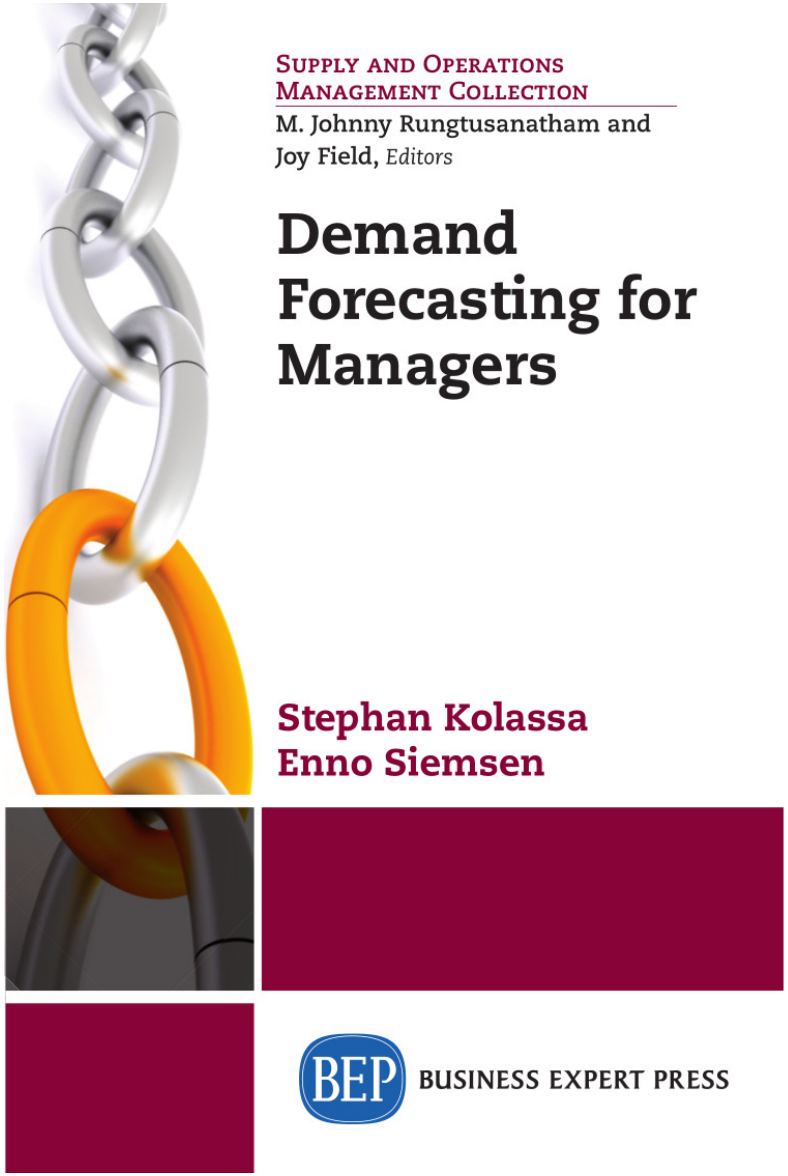
- Better: forecast and assess full predictive densities, as we did here
  - Alternative to the rPIT: proper scoring rules (see the paper)
  - Possibly assess misspecified dynamics/correlations
    - E.g., AR(7) error structure by comparing against Markov Chain alternatives
    - This is hard for low counts (low power!)

- Finally: assess the *consequences* of your forecast
  - "Cost of Forecast Error"
  - "Forecast Value Added"
  - These will usually include both interval forecasts/predictive distributions *and* subsequent processes, like logistical optimization for replenishment

References

- Berkowitz, J. (2001). Testing Density Forecasts, With Applications to Risk Management. *Journal of Business and Economic Statistics*, 19, 465-474
- Biecek, P. & Ledwina, T. (2012). ddst: Data driven smooth test. R package version 1.03. http://CRAN.R-project.org/package=ddst
- Brockwell, A. E. (2007). Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77, 1473–1478.
- Frühwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics*, 3, 291–309.
- Hanley, J. A.; Joseph, L.; Platt, R. W.; Chung, M. K. & Belisle, P. (2001). Visualizing the Median as the Minimum-Deviation Location. *The American Statistician*, 55, 150-152
- Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 679-688
- Kolassa, S. & Schütz, W. (2007). Advantages of the MAD/Mean ratio over the MAPE. *Foresight*, 6, 40-43
- Liesenfeld, R., Nolte, I., & Pohlmeier, W. (2006). Modelling financial transaction price movements: a dynamic integer count data model. *Empirical Economics*, 30, 795–825
- Ledwina, T. (1994). Data-Driven Version of Neyman's Smooth Test of Fit. *Journal of the American Statistical Association*, 89, 1000-1005
- Morlidge, S. (2015). Measuring the Quality of Intermittent Demand Forecasts: It's Worse than We've Thought! *Foresight,* 37, 37-42
- Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, 4, 283–291

Thank you!



Contact information:

Stephan Kolassa
Stephan.Kolassa@sap.com
Research Expert
Products & Innovation Suite Engineering Consumer Industries
SAP Switzerland