Newsvendor Problems: An Integrated Method for Estimation and Optimisation

Congzheng Liu[*] Adam N. Letchford[*] Ivan Svetunkov[*]

To appear in *European Journal of Operational Research*

Abstract

Newsvendor problems (NVP) form a classical and important family of stochastic optimisation problems. In this paper, we consider a data-driven solution method proposed recently by Ban and Rudin. We first examine it from a statistical viewpoint, and establish a connection with quantile regression. We then extend the approach to nonlinear NVP. Finally, we give extensive experimental results, on both simulated and real data. The results indicate that the approach performs as well as conventional ones when applied to linear NVP, but performs better when applied to nonlinear NVP. There is also evidence that the approach is more robust with respect to model misspecification.

Keywords: inventory, forecasting, newsvendor problems, data-driven optimisation, sales and operations planning

1 Introduction

Inventory control is an important topic in Operations Research and Operations Management (see, e.g. Porteus, 2002; Silver et al., 1998; Zipkin, 2000). In this paper, we focus on *Newsvendor Problems* (NVPs), by which we mean single-period inventory control problems with stochastic demand.

In early work on NVPs (Arrow et al., 1951; Morse and Kimball, 1951), it is assumed that the demand in each time period comes from a known probability distribution. Of course, in practice, this is not the case — a fact already noted by Scarf (1958). Assuming that historical demand data is available, one can attempt to address this issue by decomposing the problem into an estimation/forecasting phase and an optimisation phase. In the first phase, one makes some assumptions (i.e. specific model form and distributional assumptions) regarding the underlying data generating process, and uses the past data to estimate the parameters of the model. In the

[*] Department of Management Science, Lancaster University, Lancaster LA1 4YX, UK. Email: {c.liu19,a.n.letchford,i.svetunkov}@lancaster.ac.uk

second phase, one determines the order quantity (or quantities) based on the estimated parameter values. Throughout this paper, we will call this two-phase approach the *disjoint* approach.

An advantage of the disjoint approach is that forecasting and optimisation experts can operate independently within an organisation. This makes things easier to manage. On the other hand, as noticed by several authors (Bertsimas and Thiele, 2005; Beutel and Minner, 2012; Karmarkar, 1994; Korpela and Tuominen, 1996), there are two disadvantages:

- The two phases use different objective functions. Indeed, in the first phase, the objective is to minimise a function of the forecasting errors, such as the root mean square error or mean absolute error. In the second phase, however, the goal is usually to maximise expected profit.
- If the forecasting model is misspecified, and/or there is substantial noise in the data, then this might impact the optimisation phase in an unexpected way, possibly leading to sub-optimal solutions. In particular, upside and downside errors may have very different effects on expected profit, due to different costs associated with over- and understocking.

An alternative to the disjoint approach is to use a single, *integrated* approach, in which the order quantities are determined directly from the data based on an assumed model or filter. In this case, an adjusted loss function is used, as in *quantile regression* (Bruzda, 2016; Huber et al., 2019) and *SPO* loss (Elmachtoub and Grigas, 2017). The advantages of these approaches are that they do not make assumptions about the demand distribution, while remaining explainable. Unfortunately, they can only be applied to relatively simple NVPs, for which the objective function is linear.

Another example of integrated approach involves *machine learning* algorithms (Liyanage and Shanthikumar, 2005; Bertsimas and Thiele, 2005; Bertsimas and Kallus, 2020). With sufficient data, they can build relationships between the optimal order quantity and exogenous variables, sidestepping the need to have two phases. This approach is called *Feature-Based NVP* in He et al. (2012).

Most Feature-Based NVP approaches are "black box" approaches, which are hard to interpret or explain. A more transparent Feature-Based NVP approach was proposed recently by Ban and Rudin (2019). In their approach, statistical parameters are selected in a way that directly attempts to minimise the expected opportunity costs.

In this paper, we consider the approach in Ban and Rudin (2019) in more detail. We begin by examining it from a statistical viewpoint, and establish a connection with quantile regression. We then extend the approach to non-linear NVP, leading to what we call an *Integrated Method for Estimation and Optimisation* (IMEO). We also provide extensive simulation experiments to examine the performance of IMEO in different settings, including the situations when the true model is known and when the underlying model is mis-specified.

The rest of the paper is organised as follows. Section 2 is a literature review. Section 3 presents the IMEO framework and the theoretical analysis. Section 4 reports experimental results on simulated demand data. In Section 5, we apply our approach to some real-life data. Some concluding remarks are made in in Section 6.

2 Literature Review

Since the literature on NVPs is vast, we mention here only works of direct relevance. In Subsections 2.1 and 2.2, we review the classical NVP and its extensions, respectively. Subsection 2.3 and 2.4 review quantile regression and the Ban and Rudin method, respectively.

2.1 The classical newsvendor problem

In the simplest NVP, as defined, for example, by Choi (2012), a company purchases goods at the beginning of a time period at a cost of v per unit, and aims to sell them by the end of the period at a price p per unit. The demand during the period is a random variable Y with known probability density function f and cumulative distribution function F. At the end of the period, any surplus goods will lead to a *holding cost* of c_h per unit. On the other hand, shortage of goods during the period will lead to a *shortage cost* of c_s per unit. The goal is to determine an *order quantity* Q, prior to the period, that maximises the expected profit.

For a given Q and a given realisation y of Y, the profit over the period is:

$$\pi(Q,y) = \begin{cases} py - vQ - c_h(Q-y), & \text{if } Q \geq y \\ pQ - vQ - c_s(y-Q), & \text{if } Q < y. \end{cases} \tag{1}$$

The expected value of $\pi(Q,y)$ is:

$$\Pi(Q) = \int_0^Q \big[py - vQ - c_h(Q-y)\big] f(y)dy + \int_Q^\infty \big[pQ - vQ - c_s(y-Q)\big] f(y)dy. \tag{2}$$

It is common to call $c_u = p - v + c_s$ the ‘underage’ cost and $c_o = v + c_h$ the ‘overage’ cost. Some calculus then shows that the order quantity that maximises $\Pi(Q)$ is (Choi, 2012):

$$Q^* = F^{-1}\left(\frac{c_u}{c_o + c_u}\right), \tag{3}$$

where F^{-1} is the inverse function of F. Thus, Q^* is the τ^{th} quantile of f, with $\tau = {c_u}/{(c_o+c_u)}$. One can think of the quantity τ as a “target service level”, since aiming for this target will bring the company maximised expected profit.

2.2 More complex newsvendor problems

Since the introduction of the NVP by Arrow et al. (1951), researchers have considered several extensions of the problem, including variants with multiple product types (Hadley and Whitin, 1963; Lau and Lau, 1996; Moon and Silver, 2000), quantity discounts (Khouja, 1995), different risk measures (Eeckhoudt et al., 1995), product substitution (Bassok et al., 1999), nonlinear cost functions (Halman et al., 2012), non-stationary demand (Kim et al., 2015), and price setting (Karlin and Carr, 1962; Mills, 1959; Petruzzi and Dada, 1999).

For the purpose of what follows, we now explain one variant, the 'Nonlinear Newsvendor Problem' (NNVP), which can be found in works of Halman et al. (2012); Khouja (1995); Kyparisis and Koulamas (2018); Pal et al. (2015) and many others. In the general NNVP, the profit function takes the form:

$$\pi(Q,y) = \begin{cases} P(Q,y) - V(Q) - C_h(Q,y), & \text{for } Q \geq y \\ P(Q,y) - V(Q) - C_s(Q,y), & \text{for } Q < y, \end{cases} \tag{4}$$

where V, P, C_h and C_s are now *functions* rather than constants.

Using the NNVP enables one to model more real-life problems. Indeed, as noticed by Pantumsinchai and Knowles (1991), Khouja (1995), non-linear costs arise frequently in practice. For example, in real life, a minor shortage may not cause large costs, but a major shortage could damage the reputation of the company. As another example, a small amount of excess stock can often be sold at a discount, but this may not be possible for a large amount, not to mention the loss of goodwill. Moreover, as a product stays on the shelf longer, the opportunity cost of the shelf space may increase over time. All these examples show that overage and underage costs can be non-linear in real life situations.

If $\pi(Q,y)$ has a particularly simple form (e.g., if it is piecewise-linear as in the classical NVP), then it may be possible to use calculus to express the optimal order quantity as a quantile (Choi, 2012). In general, however, a closed-form expression as a quantile is unlikely to exist (Halman et al., 2012; Porteus, 2002). In such cases, one must resort to numerical integration and search techniques to solve the NVP (Solis and Wets, 1981).

We now recall one specific NNVP, taken from Kyparisis and Koulamas (2018), that we are going to use in our experiments later in this paper. As second-round sales in salvage markets and proportional shortage penalties are very common in real life (Kashefi, 2016; Liberopoulos et al., 2010), it is particular necessary to check the performance of the proposed method on this NNVP. The purchase cost v and selling price p are constants, but C_h

and C_s are functions. Overstock items incur a constant unit penalty $\alpha > 0$, but they can be sold in a salvage market with fixed unit sales price β, with $0 < \beta < v$. The demand in the salvage market is itself a random variable, with known distribution, which we denote by u. That is, we have:

$$C_h(Q, y) \, = \, \alpha[Q - y]^+ - \beta \, \mathbb{E} \left[\min \left\{ [Q - y]^+, u \right\} \right]. \tag{5}$$

Moreover, the shortage penalty is proportional to the shortage quantity. That is:

$$C_s(Q, y) = \zeta \left([y - Q]^+ \right)^2 \tag{6}$$

for some constant $\zeta > 0$. This problem cannot be solved analytically, but there are known approximation methods that give adequate solutions (Kyparisis and Koulamas, 2018).

2.3 Quantile regression

Returning to the classical NVP, we now consider the (more realistic) case in which the demand distribution is unknown, but we have historical demands $y_1, y_2, \ldots, y_s$. For this case, *quantile regression* has proven to perform well, and the basic idea is as follows (Bertsimas and Thiele, 2005):

1. Compute the value of τ that maximises expected profit;
2. Use quantile regression to compute an estimate of the τ^{th} quantile of the demand in the next time period, which we denote by $\hat{y}^{(\tau)}_{s+1}$;
3. Set the order quantity $\hat{Q}_{s+1}$ equal to $\hat{y}^{(\tau)}_{s+1}$.

The biggest advantage of quantile regression is that it does not assume a specific demand distribution (Huber et al., 2019). However, it is efficient only on large samples (Ban and Rudin, 2019; Huber et al., 2019). Another drawback is that its performance depends crucially on the underlying target service level. Ban and Rudin (2019) demonstrated that the benefit of using quantile regression is limited to target service levels smaller than 0.8. If the target service level is higher, then much more data is needed in order to correctly estimate a specific quantile. Moreover, as mentioned in the previous subsection, for the more complex NVPs, there is no easy way to express the optimal order quantity as a quantile.

2.4 The Ban and Rudin approach

In the approach of Ban and Rudin (2019), a statistical model is built, in which exogenous variables are regressed against the order quantity. In more detail, we have historical data $[(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_s, y_s)]$, where $\mathbf{x}_t = [x_t^1, \ldots, x_t^p]$

represents features related to the demand, such as seasonality, price, promotions and so on. The problem now becomes that of finding the optimal function $q(\cdot)$ that maps the observed features $\mathbf{x}_{s+1} \in \mathcal{X}$ to an order $q(\mathbf{x}_{s+1}) \in \mathbb{R}$. Ban and Rudin propose to select $q(\cdot)$ from a pre-specified family of functions, in a way that minimises the empirical opportunity cost:

$$\min \sum_{t=1}^{s} \left(c_u[y_t - q(\mathbf{x}_t)]^+ + c_o[q(\mathbf{x}_t) - y_t]^+ \right), \tag{7}$$

The simplest version of their method uses linear functions, of the form:

$$q(\mathbf{x}_t) = \mathbf{x}_t^\mathsf{T} \boldsymbol{\beta} = \sum_{j=1}^{p} x_t^j \beta^j, \tag{8}$$

With this choice, one can determine the β^j easily with linear programming. Ban and Rudin also mention more complex variants, that use polynomials of the original features, and/or a quadratic regularisation term.

3 Analysing and Extending the Ban and Rudin Method

In this section, we analyse the approach of Ban and Rudin in more depth, and extend it to a more general family of problems.

3.1 Analysis

To begin, we attempt to provide an intuition behind the method. Suppose that, for each historical period $t \in [1, s]$, the observed demand y_t is a realisation of a random variable Y_t. Then, in principle, there exists an order quantity, say Q_t^*, that maximises the expected profit given Y_t and the function Π. Thus, if we had set Q_t to Q_t^* prior to observing the true demand y_t, we would have maximised our expected profit in period t. Putting it another way, if we could somehow uncover the structure of the unobservable time series of orders $\{Q_1^*, \ldots, Q_s^*\}$, we would be able to estimate Q_{s+1}^* directly.

Of course, in practice, the distribution of demand Y_t is unknown, and the values Q_t^* are not observable. So the method approximates the Q_t^* values with the $q(\mathbf{x}_t)$ values. One therefore solves an optimisation problem to find the choice of $q(\cdot)$ that minimises the expected opportunity cost. Once that is done, the order quantity for the next time period can be set to $\hat{Q}_{s+1} = q(\mathbf{x}_{s+1})$.

Ban and Rudin did not give a statistical analysis of the behaviour of the estimates of the model parameters that arise from their method. A partial answer is that, in the case of a linear profit function, their method is equivalent to quantile regression (see A for a proof). Therefore, in the linear case, their method immediately inherits all of the desirable properties of quantile regression, such as consistency (Koenker, 2005), efficiency (Koenker and Machado, 1999) and asymptotic normality (Kocherginsky et al., 2005) of the estimates of parameters.

3.2 Extension to the nonlinear case

We now consider how to extend the approach in Ban and Rudin (2019) to the NNVP. The key issue here is that there is no simple formula for the opportunity cost in the nonlinear case. Indeed, in the literature on the NNVP, authors work directly with the function (4), rather than attempting to derive explicit functions for c_o and c_u.

To get around this difficulty, we propose to maximise the expected profit instead of minimising the expected opportunity cost (7). More precisely, we propose to compute the function $q(\cdot)$ that maximises the function

$$\max \sum_{t=1}^{s} \pi(q(\mathbf{x}_t), y_t), \tag{9}$$

where π can be a profit function of any level of complexity.

Maximising (9) is a continuous nonlinear optimisation problem. Under reasonable assumptions on the functions P, V, C_h and C_s in (4), and the function $q(\cdot)$ itself, the profit function (9) will be concave. Unfortunately, it is unlikely to be everywhere differentiable. As a result, general-purpose algorithms for nonlinear optimisation are not guaranteed to converge to global maxima. Fortunately, the experiments in the next section indicate that this does not cause serious problems.

In what follows, we call our approach an *Integrated Method for Estimation and Optimisation* or IMEO. We remark that, in the case of a linear profit function, IMEO is equivalent to the method of Ban and Rudin (see B for a proof). Thus, in that case, it is again equivalent to quantile regression.

4 Computational Experiments

In order to assess the performance of IMEO and to understand its strengths and weakness, we conduct a simulation experiment. We start the discussion with Subsection 4.2.1, where the simplest case is studied, in which the profit function is linear and the underlying data generation process (DGP) is known. The case in which the profit function is nonlinear is discussed in Subsection 4.2.2. Finally, we discuss more complicated scenarios, with misspecified models, in Subsections 4.3.1 and 4.3.2. Given the discussion in Section 3, the analysis in Subsections 4.2.1 and 4.3.1 will also tell us how the "Feature-based NVP" approach by Ban and Rudin (2019) compares with the conventional approaches in different situations.

4.1 Experimental setup

In our experiments, we consider NVPs with quarterly demand data, and we generate data from a seasonal ARIMA$(1,0,0)(1,0,0)_4$ process with $\theta = 0.3$, $\Theta = 0.5$, and constant level 500. We also assume that the error term of the DGP follows the normal distribution $\mathcal{N}(0, 200^2)$. We choose a seasonal ARIMA model since it is one of the most popular statistical models in the literature (for example, see the review paper on supply chain forecasting by Syntetos et al. (2016)). We have also experimented with other models and parameters for the DGP. The scripts of extended experiments have been made available on Github (Congzheng, 2020). The proposed method showed strong robustness and the results were very similar to the ones presented below.

For each of the cases, we simulate 20,000 sets of demand data, each consisting of 4800 observations. From each set, we extract sub-sequences of lengths 40, 120, 480 and 1200. These are used to explore how the amount of available data affects the performance of each method. The first three data lengths can help us to simulate real-life circumstances, while the data lengths of 1200 and 4800 allow us to explore the asymptotic behaviour of the approach. For each set of demand data and each method, we compute $\hat{Q}_{s+1}$, the 1-step ahead forecast of orders. We then compute the following three quantities (and aggregate them using mean over 20,000 sets):

1. Percentage Profit Loss: $PPL = \frac{\pi(y_{s+1}, y_{s+1}) - \pi(\hat{Q}_{s+1}, y_{s+1})}{\pi(y_{s+1}, y_{s+1})}$, which shows the percentage of profit that would be lost due to using each method instead of knowing the true demand. In the ideal situation $MPPL$ (mean PPL) should be equal to zero.

2. Service Level: $SL = \frac{\sum \mathbb{I}(\hat{Q}_{s+1} > y_{s+1})}{20{,}000}$, where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 if the condition inside it is satisfied. This measure shows the achieved service level. In the ideal situation, SL should correspond to the target service level.

3. Fill Rate: $FR = \frac{\min\{\hat{Q}_{s+1}, y_{s+1}\}}{y_{s+1}}$, which shows how the demand is serviced. In the ideal situation, MFR (mean FR) should be equal to one.

In the proposed method ("IMEO"), we use the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) for the estimation of parameters of the model. The L-BFGS algorithm is very popular in the nonlinear programming community (see Liu and Nocedal, 1989), and is commonly used for ARIMA fitting in software packages, for example in R (R Core Team, 2020).

Besides the proposed method, three benchmark methods are also considered:

- A disjoint method ("DJ") that estimates the parameters of the model in the first phase, and then determines the optimal order quantity in the second phase.
- An integrated method that uses Quantile Regression ("QR") in order to determine the order quantity. We choose it as a rival since it is one of the most widely used statistical approaches, which has proved to work well for the NVP.
- Finally, for the purposes of benchmarking, we use DJ with the exact model and parameters from the DGP to perform a forecast, and determine the order quantity. This last method is "idealised", since, in real life, one would not know the true model or true parameters of demand. But it allows us to see, how far approaches are from the ideal one. We call this last method "DGP".

4.2 When the true model is known

In this scenario, we assume that it is known that the true model is ARIMA $(1,0,0)(1,0,0)_4$ with constant, but its parameters need to be estimated. In all the methods, it is also assumed that the error term follows normal distribution with unknown variance.

4.2.1 Linear case

We conduct several experiments with combination of costs that give service levels of 0.3, 0.5, 0.63 and 0.9. It is important to recall that the optimal service level is a critical solution of the optimisation problem. In most case, the most profitable service level is not a "high" stocking level. Other levels can also be considered, but in additional experiments that we have conducted, we have not found any significant changes in performance of the approaches.

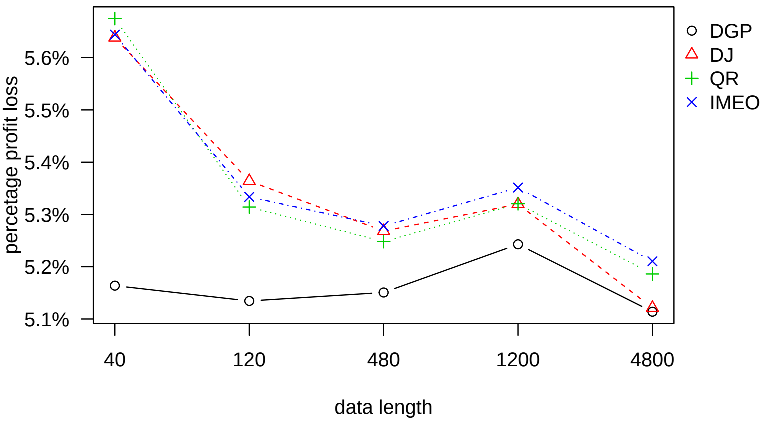
Initially, we choose the following parameters for our (linear) profit function:

- $p = 20$, $v = 10$, $c_h = -3$, $c_s = -7$.

One can check that the corresponding pair $\left(c_u, c_o\right)$ is $(3,7)$, and the target service level evaluates to 0.3. Applying the methods to this scenario, we get values of PPL, SL and FR.

Figure 1 shows the mean percentage profit loss obtained when the target service level was set to 0.3. As one would hope, the losses for all methods converge as the sample size increases. It is also apparent that IMEO and QR have very similar performance. This is to be expected, given that IMEO can be considered equivalent to QR in the linear NVP case. (The small gap between those two methods may due to different optimisation algorithms

Figure 1: Mean percentage profit loss vs. data size at 0.3 target service level



that applied, as mentioned in Section 3.) Another thing to note is that the integrated approaches perform very similar to the disjoint one, which is also expected, given the knowledge of the true model and the simplicity of the NVP.

Figure 2: Service level vs. sample size at 0.3 target service level

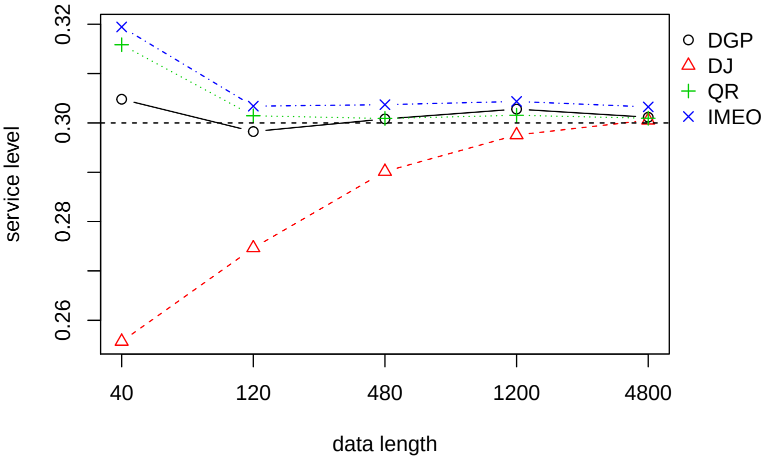


Figure 2 represents the service level, i.e., the proportion of iterations in which the demand was satisfied in the simulation. We see that all four methods converge to the desired target of 0.3 as the sample size grows. Interestingly, DJ approaches the target from below, while IMEO and QR slightly over estimate the level on small samples. A possible explanation of this phenomenon is that in the first phase of DJ, the estimated parameters tend to be closer to zero when the sample size is small. This causes the estimated order quantities to be further away from the mean than they should be, and leads to lower than needed SL. On the other hand, the integrated methods take the underage and overage costs into account. For the given cost parameters, over-stocking is less costly than under-stocking, which leads to a higher SL than needed.

Figure 3: Mean fill rate vs. sample size at 0.3 target service level

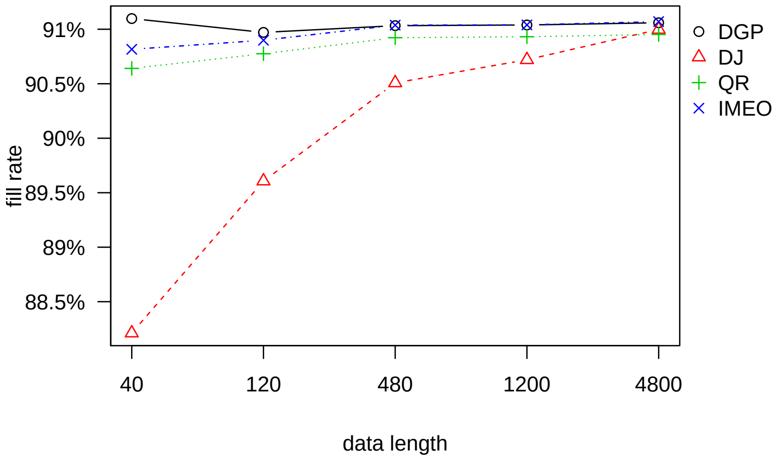


Figure 3 shows the mean fill rate. As one might expect, the performance of all methods improves with the growth of the sample size. The thing to note is that the DJ requires a much larger sample size than the other methods in order to achieve the same MFR, although the differences between the DJ and other methods are not very big (only a couple of percent points). The possible explanation of this behaviour of DJ is probably similar to the situation with the service levels.

Next, we explore whether the target service level has a significant effect on the performance of the methods. Specifically, we consider the following three alternative parameter settings:

- $p = 20$, $v = 8$, $c_h = -3$, $c_s = -7$
- $p = 20$, $v = 8$, $c_h = 3$, $c_s = 7$
- $p = 20$, $v = 8$, $c_h = -7$, $c_s = -3$.

These correspond to pairs (c_u, c_o) are $(5, 5)$, $(19, 11)$ and $(9, 1)$, respectively, leading to the target service levels of 0.5, 0.63 and 0.9, respectively. It is important to recall that the target service level refers to the most profitable service level for given parameters.

The relevant data is given in Tables 1 and 2, for the cases of sample sizes $s = 40$ and $s = 4800$, respectively. The best cases (excluding DGP, which is expected to be the best by construct) are marked in boldface for each of the error measures and each of the service levels.

We can see from Table 1 that IMEO performs consistently better than the other approaches on the small sample, regardless of the target service

Table 1: Target service level effect with $s = 40$

	Mean percentage profit loss				**Service level**				**Mean fill rate**			
Target service level	DGP	DJ	QR	IMEO	DGP	DJ	QR	IMEO	DGP	DJ	QR	IMEO
0.3	5.2%	**5.6%**	5.7%	**5.6%**	0.30	0.26	**0.32**	**0.32**	91.1%	88.2%	90.6%	**90.8%**
0.5	4.9%	5.3%	5.3%	**5.2%**	0.50	0.44	**0.50**	**0.50**	95.2%	93.4%	94.7%	**94.8%**
0.63	13.8%	15.1%	15.0%	**14.8%**	0.63	0.58	**0.62**	**0.62**	97.1%	95.8%	**96.6%**	**96.6%**
0.9	2.1%	2.4%	2.4%	**2.3%**	0.90	0.91	0.91	**0.90**	99.6%	99.1%	99.2%	**99.4%**

level, although the difference in performance between the methods is not substantial. QR is the second best approach in this scenario.

Table 2: Target service level effect with $s = 4800$

	Mean percentage profit loss				**Service level**				**Mean fill rate**			
Target service level	DGP	DJ	QR	IMEO	DGP	DJ	QR	IMEO	DGP	DJ	QR	IMEO
0.3	5.1%	**5.1%**	5.2%	5.2%	0.30	0.30	0.30	0.30	91.1%	91.0%	91.0%	**91.1%**
0.5	4.9%	**4.9%**	5.0%	5.0%	0.50	0.50	0.50	0.50	95.3%	**95.2%**	**95.2%**	**95.2%**
0.63	13.8%	**13.8%**	14.0%	14.0%	0.63	0.63	0.63	0.63	97.0%	**97.0%**	**97.0%**	96.9%
0.9	2.1%	**2.1%**	2.2%	**2.1%**	0.90	0.90	0.90	0.90	99.5%	99.2%	99.1%	**99.4%**

When it comes to large samples (Table 2), IMEO and QR perform slightly worse than DJ, although the difference between the methods is not substantial. This holds for all target service levels, irrespective of the error measures. A thing to note is that the MPPL is particularly high for all methods when the target service level is 0.63. This is probably due to the relatively large overage and underage costs in that case. In addition, note that, asymptotically, all methods approach the target service level. This makes IMEO and QR more desirable than DJ, because they do at least as well as DJ both on small and large samples.

Summarising, we can see that IMEO does at least as well as the classical disjoint method and quantile regression in different scenarios of the linear NVP, when the model is correctly specified.

4.2.2 Nonlinear profit function

In this subsection, we examine the relative performance of the methods when applied to the NNVP. As before, we assume that the true model for the demand is known. We use the following nonlinear profit function (Kyparisis and Koulamas, 2018):

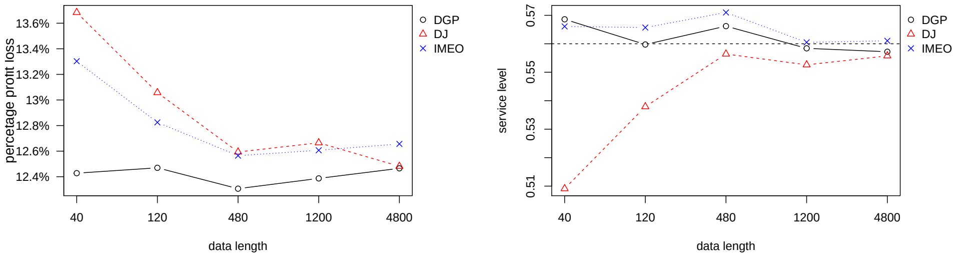
$$\pi(Q, y) = \begin{cases} 20y - 8Q - 4(Q - y) + 5\,\mathbb{E}[\min\{(Q - y), u\}], & \text{if } Q \geq y \\ 20Q - 8Q - 0.01(y - Q)^2, & \text{if } Q < y, \end{cases} \tag{10}$$

where $u \sim \mathcal{N}(30, 5^2)$. Given that the target service level does not have a closed form in the NNVP, one can use the technique proposed by Kyparisis and Koulamas (2018), or other numerical approaches, to verify that it is approximately equal to 0.56.

Since the nonlinearity makes it impossible to apply QR (as discussed in Subsection 2.3), we present results only for DJ and IMEO, together with the "idealised" method based on the DGP. Moreover, we found that the MFR plots were similar in all scenarios and did not give any additional important information. Therefore, we do not present them in what follows.

Figure 4 shows the MPPL and SL of each method, for each sample size. It is apparent that both DJ and IMEO perform well in terms of MPPL, with very similar values for all sample sizes. IMEO is slightly better on small samples (similar to what we have seen in Subsection 4.2.1), but the differences in performance between the methods are not substantial. When it comes to SL, the picture is similar to what we have observed in Subsection 4.2.1: the SL of IMEO is very close to the the target service level even when the sample size is small, while the disjoint method needs more data to reach the target service level, and it approaches it from below. The reason for such performance is similar to the one discussed in Subsection 4.2.1. As expected, both methods converge to the DGP values in both MPPL and SL with the increase of the sample size.

Figure 4: Performance vs. sample size with nonlinear profit function



(a) percentage profit loss vs. data size

(b) service level vs. data size

We would like to stress that, unlike disjoint method, IMEO does not need any complicated numerical optimisation or simulation methods to estimate the optimal order quantity – it does that directly. In addition, we have found in our experiments that it typically required less computational time than the disjoint method: approximately only one-tenth of the time spent on DJ was needed for IMEO, when $s = 40$.

Overall, we see that IMEO performs at least as good as DJ when the true model is known. However, the true model is typically not known in practice, so next we investigate situations, when the model is misspecified.

4.3 When the true model is not known

We examine the effect of *model misspecification* on the relative performance of the various methods. We consider three scenarios:

1. Model omits important variable, which typically leads to biased estimates of parameters;
2. Model has redundant variables, which usually leads to ineficient estimates of parameters;
3. The assumed distribution of error term is wrong.

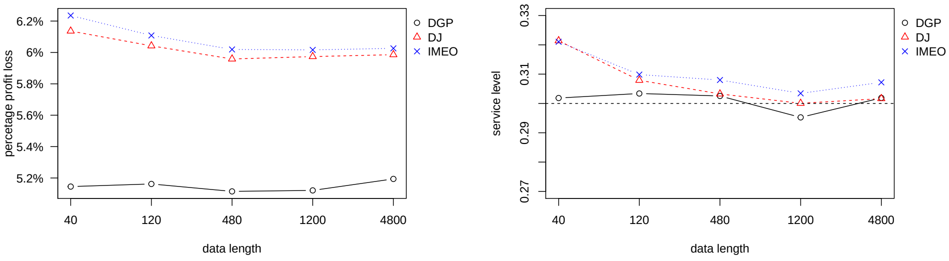
While in reality there can be other scenarios, the proposed three scenarios cover the main possible issues with the model misspecification.

We don't include the QR method in further discussions, because it performs similarly to IMEO in the linear case, and cannot be used in the non-linear case.

4.3.1 Linear case

We first consider the situation in which **the model omits important variables** (i.e., is under-parameterised): the seasonal order of ARIMA is dropped in the estimation and the solution of the NVP. In other words, when estimating the optimal order quantity, we apply the AR(1) model with constant.

Figure 5: Performance vs. sample size with under-parameterised linear model



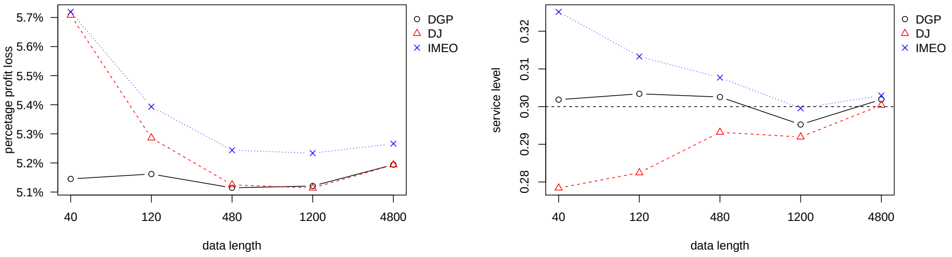
(a) percentage profit loss vs. data size

(b) service level vs. data size

The MPPL and SL for this case are shown in Figure 5. By comparing with Figure 1, we see that the use of an under-parameterised model causes both DJ and IMEO to incur small additional losses in MPPL, irrespective of the sample size. This is expected, as the seasonal pattern in the data is ignored by the models. These losses do not significantly improve as the number of data points increases, because seasonality plays a key role when making one-step ahead forecasts. When it comes to the SL, both approaches perform similarly, reaching the target on larger samples. But when it comes to small samples, they both reach higher than needed levels. Interestingly, the performance of DJ and IMEO is similar, even though they are very different approaches.

Next, we consider the situation in which **the model contains redundant variables** (i.e., is over-parameterised): an unnecessary lag term is included. Specifically, we apply an ARIMA$(2, 0, 0)(1, 0, 0)_4$ model to the data.

Figure 6: Performance vs. sample size with over-paramaterised linear model



(a) percentage profit loss vs. data size

(b) service level vs. data size

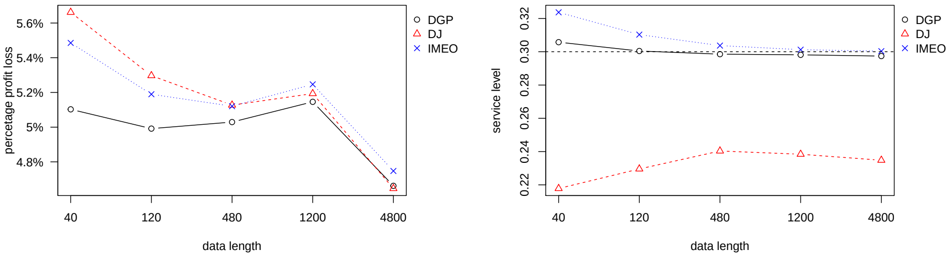
The MPPL and SL for this case can be seen in Figure 6. Both DJ and IMEO incur a loss in profit, as before. Now, however, the profit loss decreases with the increase of the sample size. This can be explained by a well-known statistical phenomenon when estimating models with redundant variables (Farrar and Glauber, 1967): they tend to lead to less efficient estimates of parameters on small samples but do not lead to systematic bias (as omitted variables typically do). As for the service level, it can be seen that both methods asymptotically converge to the target, but tend to be less precise on small samples, with IMEO reaching higher levels than needed and DJ leading to lower values. The performance is similar to what we saw in Subsection 4.2.1.

We do not present results for other target service levels in this part, since they were very similar to what we observed above.

Finally, we explore the effect on performance of different methods, when **the assumed distribution is incorrect**. In particular, we generate data using a modified version of our seasonal ARIMA$(1, 0, 0)(1, 0, 0)_4$ model, in which the *error term* follows the Laplace distribution with mean 0 and scale 141 (which will have standard deviation of 200). We remark that the Laplace distribution has “fatter tails”, or, more formally, a higher kurtosis than the normal distribution. When estimating the optimal order quantity, however, we use the incorrect assumption that the error term follows the normal distribution.

The results for this scenario are presented in Figure 7. We can see that the incorrect distributional assumption leads only to a very small loss in profit for both DJ and IMEO. They both converge to DGP as the sample size increases. However, the analysis of SL shows that while IMEO rapidly converges to the target service level from above, the DJ achieves a lower than needed service level, producing a biased value – and this drawback is not remedied by an increase of the sample size. A possible explanation is that the integrated approach works directly with the data, and does not rely on the assumption of normality, being in this sense "non-parametric", while the DJ relies on normality and consistently underestimates the uncertainty in the data. This example suggests that IMEO my be more robust.

Figure 7: Performance vs. sample size with Laplace distributed error term



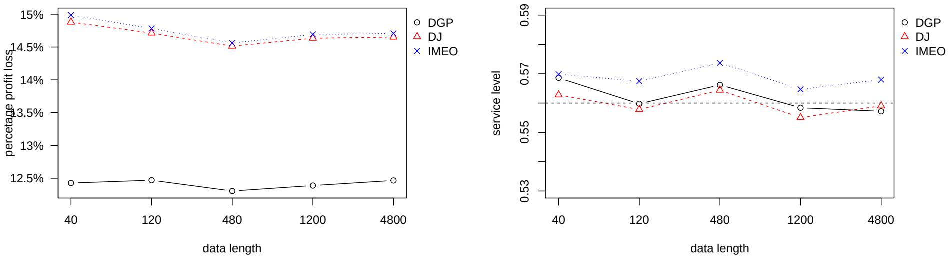
(a) percentage profit loss vs. data size

(b) service level vs. data size

4.3.2 Nonlinear case

Finally, we explore the case in which there is a nonlinear profit function and misspecification simultaneously. The same three scenarios of misspecification are considered.

Figure 8: Performance vs. sample size with under-parameterised nonlinear model

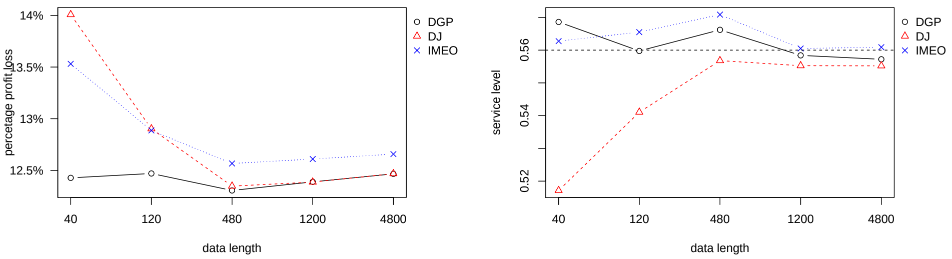


(a) percentage profit loss vs. data size

(b) service level vs. data size

Figure 8 represents the scenario in which the applied models are under-parameterised. We can see that the performance of the methods in terms of MPPL and SL are similar to the case of linear NVP discussed in Subsection 4.3.1. The percentage profit loss for both DJ and IMEO stabilises around 14.75% and never converges to the level of DGP, due to the absence of the important variable in the model. This is compensated by SL, for which DJ performs similar to DGP, with IMEO reaching a slightly higher than needed level.

Figure 9: Performance vs. sample size with over-parameterised nonlinear model



(a) percentage profit loss vs. data size

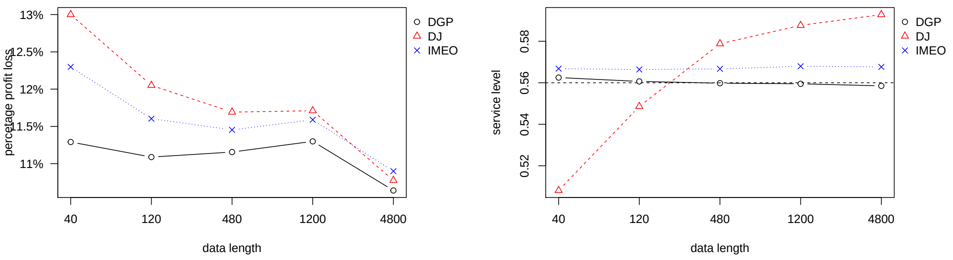
(b) service level vs. data size

Figure 9 demonstrates the over-parameterised scenario. We can see that IMEO has lower MPPL than DJ on small samples. With an increase in sample size, the latter method converges to the DGP, while IMEO still has a slight bias, producing around 0.1% higher loss than the DGP. The good performance of IMEO on small samples could be because it does not need to estimate the variance of the error term. As for the SL, IMEO reaches the target level much faster than DJ, performing especially well on small samples, where the latter method reaches much lower service level than needed. As the sample size increases, both methods converge to the target level.

Finally, we consider the scenario in which an incorrect distribution of error term is assumed. The results are presented in Figure 10. It becomes apparent that IMEO outperform DJ in terms of MPPL on small samples and converges to DGP together with DJ. IMEO also seems more stable than DJ in terms of SL across all sample sizes. While IMEO does not converge to the target level on larger samples, it is consistent and less biased than DJ, which converges to a value much higher than needed. One interesting thing we can see from both Figure 10 and Figure 7 is that DJ always gives a very low SL when the sample size is small, and converges to a higher than needed target as the sample size grows. This “wrong” target is obtained in both situations, not surprisingly, because DJ cannot overcome the incorrectness in the distributional assumption.

Summarising this subsection, we see that IMEO is robust and does not fail as badly as the classical disjoint method does in severe cases of misspecification. At worst, IMEO performs similarly to DJ. In addition, it looks like IMEO does consistently better than the DJ in terms of service level, so if this is more important for a company than profit loss, then we would recommended using IMEO.

Figure 10: Performance vs. sample size with Laplace distributed error term, nonlinear model



(a) percentage profit loss vs. data size

(b) service level vs. data size

5 Real-life Case

In this section, we examine the performance of IMEO on a real-life nurse staffing problem in a hospital, which must determine the next-day staffing level Q. The hospital incurs an underage cost (unexpected high death rate, reputation damage, etc.) if there are not enough nurses, and an overage cost (unnecessary exposure risk, big salary payment, etc.) if there are too many nurses. Both types of costs are considered to be nonlinear in this case (Fernandez et al., 2020; Al Thobaity and Alshammari, 2020). The objective is to minimise the expected daily cost.

The data we use comes from the NHS open data-set (NHS England, 2020). It includes the total bed occupancies for a large UK general hospital from April to October 2020 on a daily basis.

We assume a fixed 1 to 3 nurse-to-bed ratio (similar to how it was done in Ban and Rudin, 2019), hence the demand y is the total number of beds occupied divided by 3. In addition, we do not require the number of nurses to be integer, due to possibility of them working part-time. Based on the study of Chen et al. (2020) and Liu et al. (2020), we can approximate this problem as:

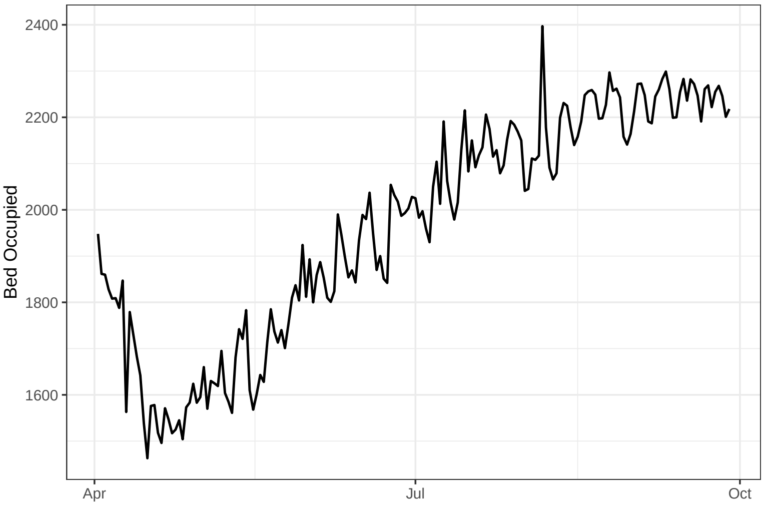
$$\pi(Q, y) = \alpha[Q - y]^+ - \beta\,\mathbb{E}\left[\min\left\{[Q - y]^+, u\right\}\right] + \zeta\left([y - Q]^+\right)^2. \qquad (11)$$

Here, we assume that each over-scheduled nurse incurs a fixed cost $\alpha > 0$, but can be reassigned to help other departments and reduce the cost by the amount β; whereas demand of other under-staffed departments can be seen as a random variable u. Moreover, we assume that the shortage penalty of nurse staffing is proportional to the shortage quantity with rate ζ, since that shortage may be covered by reassignment or overcharging, but significant shortage could be lethal to patients. The parameter values are chosen based on the studies of Chen et al. (2020), Liu et al. (2020) and Coronini-Cronberg et al. (2020):

- $\alpha = 10$, $\beta = 4$, $\zeta = 1$, and $u \sim U(0, 15)$.

To get some sense of the data, we provide a time-series plot in Figure 11. It can be seen that from mid-April to October, the number of beds occupied exhibits seasonality, and that from late-August, the number becomes stable. According to the public information in the NHS data set, the hospital nearly reaches the maximum capacity.

Figure 11: Time-series plot of bed occupancy

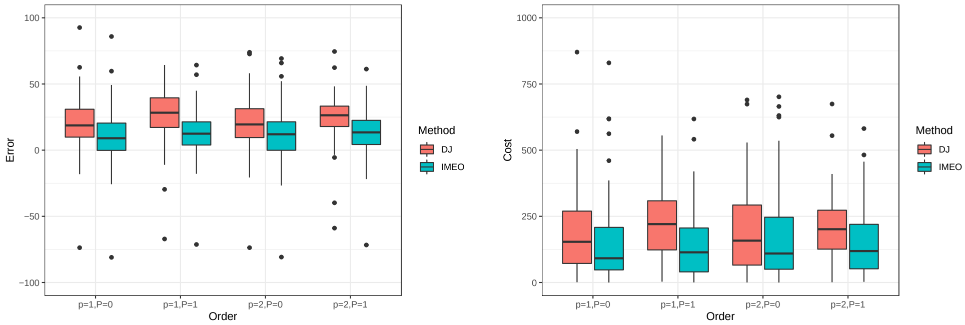


To perform a fair comparison, we apply both the disjoint method ("DJ") and the proposed method ("IMEO") to this NNVP, with several ARIMA models with orders $p = 1, 2$ and $P = 0, 1$. We use the ARIMA model here to maintain consistency with previous sections. It is, of course, possible that other models would fit better than the chosen ARIMA. However, our intention here is not to find the best-fitting model, but to compare the performance of IMEO and the disjoint method fairly. After all, it is more interesting to see how the two approaches compare, when the applied model is wrong.

For the purpose of generality, we include four pairs of orders in ARIMA ($p = 1, P = 0$; $p = 2, P = 0$; $p = 1, P = 1$ and $p = 2, P = 1$). To compare the performance of the methods, we obtain their 1-step ahead forecasts with rolling horizon (Tashman, 2000), where origin length is $s = 100$ and the origin is shifted $n = 80$ times. For each forecasted value, we compute the over-scheduled/under-scheduled nurse number, the service level, and the daily cost.

The boxplots are shown in Figures 12a and 12b. The black lines in the boxes represent mean values rather than medians. From the plots, we can see that IMEO outperforms DJ with all ARIMA models, in terms of both the mean cost and the mean scheduling error. Moreover, DJ produces larger variance of both errors and costs than IMEO. Thus, IMEO not only has lower mean costs, but also works more efficiently overall. This finding is in line with the MPPL result at data length 120 in Subsection 4.2.2.

Figure 12: Boxplot of the out-of-sample performance. The black lines in the boxes represent mean values



(a) Boxplot of the out-of-sample error

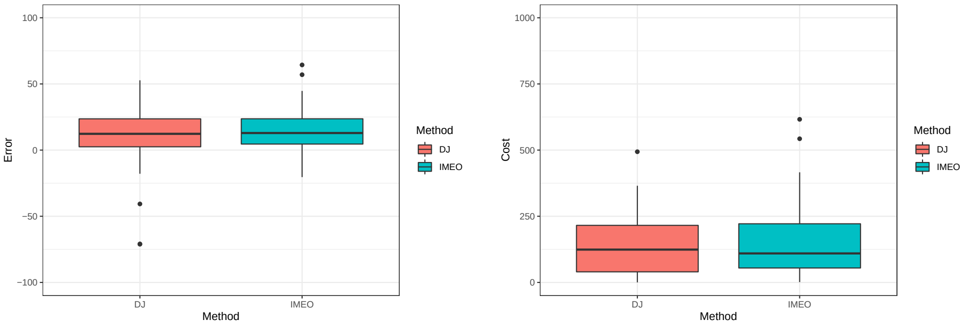
(b) Boxplot of the out-of-sample cost

We also used the `auto.arima()` function from forecast package (Hyndman et al., 2020) for R in order to select the most appropriate ARIMA model for the DJ method and then we used the same model for the IMEO (Figures 13a and 13b). In this case we are favouring DJ approach, producing the model closer to the true one based on Akaike's information criterion. At the same time, the appropriate order selection mechanism for IMEO is not yet developed, so the same model will not be optimal for it. Still, analysing the plots on Figure 13, we can see that the difference between the two methods is not substantial, with boxplots being very close to each other. This additional experiment shows that even when the DJ approach is done properly, using the state of the art forecasting techniques, IMEO does not fail substantially and can be considered as a decent alternative to DJ. Together with the results from Figures 12, this experiment shows that IMEO works well in a wide variety of cases.

The service level achieved by each method is summarised in Table 3. One can check by simulation or numerical methods that the "target service level" that minimises cost for the given parameters is around 0.8. In that case, IMEO achieves a much closer service level to the target than the disjoint method, no matter what ARIMA model is used, while the disjoint method provides a higher than needed service level in all four cases.

This example shows that IMEO is a robust approach that results in lower costs and a service level closer to the target, even when the model is specified incorrectly. The disjoint method is much more sensitive to the specification of the model, performing poorly when the model is misspecified.

Figure 13: Boxplot of the out-of-sample performance with auto-fitting. The black lines in the boxes represent mean values



(a) Boxplot of the out-of-sample error with auto-fitting

(b) Boxplot of the out-of-sample cost with auto-fitting

Table 3: Service level of each methods for real-life case

Order	$p = 1, P = 0$	$p = 2, P = 0$	$p = 1, P = 1$	$p = 2, P = 1$
DJ	0.950	0.950	0.975	0.975
IMEO	0.750	0.750	0.825	0.850

Note: The service level achieved by auto-fitting algorithm is 0.950.

6 Concluding Remarks

In this paper, we extended the method of Ban and Rudin (2019) to a broader framework, called "Integrated Method for Estimation and Optimisation" or IMEO. IMEO attempts to maximise the expected profit instead of minimising the expected opportunity cost, which turns out to be an important distinction in the nonlinear case. We showed that IMEO reduces to the method of Ban and Rudin in the linear case, when both methods turn out to be equivalent to quantile regression. Our experiments indicate that IMEO performs at least as well as the benchmark methods, in terms of both mean percentage profit loss and service level. It also appears that IMEO is more robust with regards to model misspecification. It also does well in terms of service level, which could be attractive in real-life applications.

While the focus of the experiments in this paper was on ARIMA models, the proposed approach could be applied to other models as well, such as linear regression with explanatory variables, nonlinear regression and ETS.

There are several interesting topics for further research. First, it would be interesting to study the performance of IMEO with other demand models, such as ETS. Second, it would be desirable to develop a variable selection mechanism in IMEO. The conventional disjoint method allows one to do this in the first phase, for example by using cross-validation or a stepwise technique based on information criteria, while Ban and Rudin (2019) use regularisation for the selection and estimation. While these are good approaches, they require large samples and are computationally expensive. Our hope is that a more efficient feature selection method can be developed. Third, we focused our research on NVP, but IMEO could be potentially extended to multi-period inventory problems. Finally, it would be interesting to extend IMEO to multi-item problems, either with or without substitution effects between products.

A Proof of Equivalence to Quantile Regression

Proof. In quantile regression, the objective function is defined as (Koenker and Hallock, 2001):

$$\min \sum_{t=1}^{s} \rho_\tau(y_t - q(\mathbf{x}_t)),$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}_{(u<0)})$, and $\mathbb{I}$ is an indicator function. Thus, simply by setting $\tau = {}^{c_u}/_{(c_o+c_u)}$, we have:

$$\begin{aligned}
&\min \sum_{t=1}^{s} \big(c_u[y_t - q(\mathbf{x}_t)]^+ + c_o[q(\mathbf{x}_t) - y_t]^+\big) \\
&\quad = \min(c_o + c_u) \sum_{t=1}^{s} \big(\frac{c_u}{c_o + c_u}[y_t - q(\mathbf{x}_t)]^+ + \frac{c_o}{c_o + c_u}[q(\mathbf{x}_t) - y_t]^+\big) \\
&\quad = \min \sum_{t=1}^{s} \big(\tau[y_t - q(\mathbf{x}_t)]^+ + (1-\tau)[q(\mathbf{x}_t) - y_t]^+\big) \\
&\quad = \min \sum_{t=1}^{s} \rho_\tau(y_t - q(\mathbf{x}_t)).
\end{aligned}$$

□

B Proof of Maximising-Minimising Transformation

Proof. We have:

$$\min[a, b] = a - [a - b]^+,$$

and

$$a - b = [a - b]^+ - [b - a]^+.$$

We can transform:

$$\begin{aligned}
&\pi(q(\mathbf{x}_t), y_t) \\
&\quad = p \min[q(\mathbf{x}_t), y_t] - vq(\mathbf{x}_t) - c_h[q(\mathbf{x}_t) - y_t]^+ - c_s[y_t - q(\mathbf{x}_t)]^+ \\
&\quad = p\{q(\mathbf{x}_t) - [q(\mathbf{x}_t) - y_t]^+\} - vq(\mathbf{x}_t) - c_h[q(\mathbf{x}_t) - y_t]^+ - c_s[y_t - q(\mathbf{x}_t)]^+ \\
&\quad = (p - v)q(\mathbf{x}_t) - (c_h + p)[q(\mathbf{x}_t) - y_t]^+ - c_s[y_t - q(\mathbf{x}_t)]^+.
\end{aligned}$$

Therefore, we have (since y_t is fixed):

$$
\begin{aligned}
&\max \sum_{t=1}^{s} \pi(q(\mathbf{x}_t), y_t) \\
&\quad = \max \sum_{t=1}^{s} \{(p-v)q(\mathbf{x}_t) - (c_h+p)[q(\mathbf{x}_t)-y_t]^+ - c_s[y_t - q(\mathbf{x}_t)]^+\} \\
&\quad = \max \sum_{t=1}^{s} \{(p-v)[q(\mathbf{x}_t)-y_t] - (c_h+p)[q(\mathbf{x}_t)-y_t]^+ - c_s[y_t - q(\mathbf{x}_t)]^+\} \\
&\quad = \max \sum_{t=1}^{s} \{(p-v)[q(\mathbf{x}_t)-y_t]^+ - (p-v)[y_t - q(\mathbf{x}_t)]^+ \\
&\qquad - (c_h+p)[q(\mathbf{x}_t)-y_t]^+ - c_s[y_t - q(\mathbf{x}_t)]^+\} \\
&\quad = \min \sum_{t=1}^{s} \{(v+c_h)[q(\mathbf{x}_t)-y_t]^+ + (p-v+c_s)[y_t - q(\mathbf{x}_t)]^+\} \\
&\quad = \min \sum_{t=1}^{s} \{c_o[q(\mathbf{x}_t)-y_t]^+ + c_u[y_t - q(\mathbf{x}_t)]^+\}.
\end{aligned}
$$

□

References

Al Thobaity, A. and Alshammari, F. (2020). Nurses on the frontline against the COVID-19 pandemic: an integrative review. *Dubai Medical Journal*, 3:87–92.

Arrow, K., Harris, T., and Marschak, J. (1951). Optimal inventory policy. *Econometrica*, 19:250–272.

Ban, G.-Y. and Rudin, C. (2019). The big data newsvendor: practical insights from machine learning. *Operations Research*, 67:90–108.

Bassok, Y., Anupindi, R., and Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research*, 47:632–642.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66:1025–1044.

Bertsimas, D. and Thiele, A. (2005). A data-driven approach to newsvendor problems. Technical report, Operations Research Center, MIT.

Beutel, A.-L. and Minner, S. (2012). Safety stock planning under causal demand forecasting. *International Journal of Production Economics*, 140:637–645.

Bruzda, J. (2016). Quantile forecasting in operational planning and inventory management—an initial empirical verification. *Dynamic Econometric Models*, 16:5–20.

Chen, S.-C., Lai, Y.-H., and Tsay, S.-L. (2020). Nursing perspectives on the impacts of COVID-19. *Journal of Nursing Research*, 28:1682–3141.

Choi, T.-M. (2012). *Handbook of Newsvendor Problems.* Springer, New York.

Congzheng, L. (2020). Forecasting-and-newsvendor. `https://github.com/Joshua960411/Forecasting-and-Newsvendor`.

Coronini-Cronberg, S., Maile, E., and Majeed, A. (2020). Health inequalities: the hidden cost of COVID-19 in NHS hospital trusts? *Journal of the Royal Society of Medicine*, 113:179–184.

Eeckhoudt, L., Gollier, C., and Schlesinger, H. (1995). The risk-averse (and prudent) newsboy. *Management Science*, 41:786–794.

Elmachtoub, A. N. and Grigas, P. (2017). Smart "predict, then optimize". *arXiv preprint arXiv:1710.08005.*

Farrar, D. and Glauber, R. (1967). Multicollinearity in regression analysis: the problem revisited. *Review of Economic and Statistics*, pages 92–107.

Fernandez, R., Lord, H., Halcomb, E., Moxham, L., Middleton, R., Alananzeh, I., and Ellwood, L. (2020). Implications for COVID-19: a systematic review of nurses' experiences of working in acute care hospital settings during a respiratory pandemic. *International Journal of Nursing Studies.* vol. 111, article 103637.

Hadley, G. and Whitin, T. (1963). *Analysis of Inventory Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Halman, N., Orlin, J., and Simchi-Levi, D. (2012). Approximating the nonlinear newsvendor and single-item stochastic lot-sizing problems when data is given by an oracle. *Operations Research*, 60:429–446.

He, B., Dexter, F., Macario, A., and Zenios, S. (2012). The timing of staffing decisions in hospital operating rooms: Incorporating workload heterogeneity into the newsvendor problem. *Manufacturing & Service Operations Management*, 14(1):99–114.

Huber, J., Müller, S., Fleischmann, M., and Stuckenschmidt, H. (2019). A data-driven newsvendor problem: from data to decision. *European Journal of Operational Research*, 278:904–915.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2020). *forecast: Forecasting functions for time series and linear models.* R package version 8.13.

Karlin, S. and Carr, C. (1962). Prices and optimal inventory policy. In Arrow, K. and Scarf, H., editors, *Studies in Applied Probability and Management Science*, pages 159–172. Stanford University Press.

Karmarkar, U. (1994). A robust forecasting technique for inventory and leadtime management. *Journal of Operations Management*, 12:45–54.

Kashefi, M. (2016). Effect of salvage market on strategic technology choice and capacity investment decision of firm under demand uncertainty. *Facility Design eJournal*, 17:140–155.

Khouja, M. (1995). The newsboy problem under progressive multiple discounts. *European Journal of Operational Research*, 84:458–466.

Kim, G., Wu, K., and Huang, E. (2015). Optimal inventory control in a multi-period newsvendor problem with non-stationary demand. *Advanced Engineering Informatics*, 29:139–145.

Kocherginsky, M., He, X., and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14:41–55.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives*, 15:143–156.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94:1296–1310.

Korpela, J. and Tuominen, M. (1996). Inventory forecasting with a multiple criteria decision tool. *International Journal of Production Economics*, 45:159–168.

Kyparisis, G. and Koulamas, C. (2018). The price-setting newsvendor with nonlinear salvage revenue and shortage cost. *Operations Research Letters*, 46:64–68.

Lau, A. and Lau, H. (1996). The newsstand problem: a capacitated multi-product single period inventory problem. *Operations Research*, 94:29–42.

Liberopoulos, G., Tsikis, I., and Delikouras, S. (2010). Backorder penalty cost coefficient “b”: what could it be? *International Journal of Production Economics*, 123:166–178.

Liu, D. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Liu, Q., Luo, D., Haase, J., Guo, Q., Wang, X., Liu, S., Xia, L., Liu, Z., Yang, J., and Yang, B. (2020). The experiences of health-care providers during the COVID-19 crisis in China: a qualitative study. *The Lancet Global Health*, 8:e790–e798.

Liyanage, L. and Shanthikumar, J. (2005). A practical inventory control policy using operational statistics. *Operations Research Letters*, 33:341–348.

Mills, E. (1959). Uncertainty and price theory. *Quarterly Journal of Economics*, 73:116–130.

Moon, I. and Silver, E. (2000). The multi-item newsvendor problem with a budget constraint and fixed ordering costs. *Journal of the Operational Research Society*, 51:602–608.

Morse, P. and Kimball, G. (1951). *Methods of Operations Research.* MIT Press.

NHS England (2020). Statistics on COVID-19 hospital activity. Available at www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity.

Pal, B., Sana, S., and Chaudhuri, K. (2015). A distribution-free newsvendor problem with nonlinear holding cost. *International Journal of Systems Science*, 46:1269–1277.

Pantumsinchai, P. and Knowles, T. (1991). Standard container size discounts and the single-period inventory problem. *Decision Sciences*, 22:612–619.

Petruzzi, N. and Dada, M. (1999). Pricing and the newsvendor problem: a review with extensions. *Operations Research*, 47:183–194.

Porteus, E. (2002). *Foundations of Stochastic Inventory Theory.* Stanford University Press.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Scarf, H. (1958). *Studies in the Mathematical Theory of Inventory and Production.* Stanford University Press.

Silver, E., Pyke, D., and Peterson, R. (1998). *Inventory Management and Production Planning and Scheduling.* Wiley, New York.

Solis, F. and Wets, R. (1981). Minimization by random search techniques. *Mathematics of Operations Research*, 6:19–30.

Syntetos, A., Babai, Z., Boylan, J., Kolassa, S., and Nikolopoulos, K. (2016). Supply chain forecasting: theory, practice, their gap and the future. *European Journal of Operational Research*, 252:1–26.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting.*

Zipkin, P. (2000). *Foundations of Inventory Management.* McGraw-Hill.