Data Scientist Technical Test

David Torres Sanchez

November 6, 2018

1 Stochastic Optimisation

Given the stochastic nature of the problem I decided to try and implement some stochastic optimisation techniques. After some research, I decided that the literature stream that best fitted the problem is the stochastic machine scheduling problem. Particularly, I borrowed some of the concepts in Li and Hu (2017) (included in `/files/`) to formulate and solve the problem using scenario generation.

2 Modelling

2.1 Notation

Sets

$\mathcal{G}$**:** Set of glovers $\mathcal{G} = \{1, 2, \ldots, G\}$;

$\mathcal{T}$**:** Set of slots $\mathcal{T} = \{1, 2, \ldots, T\}$;

$\mathcal{S}$**:** Set of scenarios $\mathcal{S} = \{1, 2, \ldots, S\}$.

Parameters

$\mathbf{st}_t/\mathbf{et}_t$**:** Start/end time of slot t;

p_s**:** Probability associated with scenario s;

$\mathbf{PL}_{ts}$**:** Sample $\{0, 1\}$ from leaving distribution for slot t under scenario s;

D_{ts}**:** Order demand for slot t under scenario s;

τ_{ts}**:** Delivery duration for a single delivery in slot t under scenario s;

Variables

x_{gt}**:** Binary variable with value 1 if glover g is assigned to a delivery in slot t, 0 otherwise;

z_{ts}**:** Binary variable (auxiliary) with value 1 if glover g begins their shift in slot t, 0 otherwise;

Z_{gts}**:** Binary variable with value 1 if glover g leaves (finishes their shift) after slot t under scenario s, 0 otherwise;

Q^F_{ts}**:** Continuous variable representing the number of orders **fulfilled** in slot t under scenario s;

Q^C_{ts}**:** Continuous variable representing the number of orders **cancelled** in slot t under scenario s;

$$\begin{aligned}
&\textbf{min} \quad \sum_s p_s \,(10\mathrm{CR} - \mathrm{UR}) && && (1)\\
&\textbf{subject to} && &&\\
&\textbf{Orders} && &&\\
&Q^C_{ts} \leq Q^F_{ts} && \forall t, s && (2)\\
&\sum_g x_{gt} \leq Q^F_{ts} && \forall t, s && (3)\\
&\textbf{Demand} && &&\\
&\sum_g x_{gt} \geq D_{ts} && \forall t, s && (4)\\
&\textbf{Delivery Duration} && &&\\
&x_{gt} \leq x_{gt'} && \forall g, t, \{t' : \mathrm{st}_t < \mathrm{st}_{t'} < \mathrm{et}_t\} && (5)\\
&\textbf{Glovers Leaving} && &&\\
&Z_{gts} \geq \mathrm{PL}_{ts} && \forall g, t, s && (6)\\
&x_{gt'} \leq 1 - Z_{gts} && \forall g, t, s, t' = \inf\{t' : \ \mathrm{et}_t \leq \mathrm{st}_{t'}\} && (7)\\
&\textbf{Transitivity (consecutive values)} && &&\\
&z_{gt_0} \geq x_{gt_0} && \forall g && (8)\\
&z_{gt} \geq x_{gt} - x_{g,t-1} && \forall g, t && (9)\\
&\sum_t z_{gt} \leq 1 && \forall g && (10)\\
&\textbf{Variable Definitions} && &&\\
&x_{gt}, z_{gt} \in \{0, 1\} && \forall g, t && (11)\\
&Z_{gts} \in \{0, 1\} && \forall g, t, s && (12)\\
&Q^F_{ts}, Q^C_{ts} \in \mathbb{R}^+ && \forall t, s && (13)
\end{aligned}$$

Where the objective function components are,

$$\mathrm{CR} = \frac{\sum_t Q^C_{ts}}{\sum_t Q^F_{ts}}$$

and,

$$\mathrm{UR} = \frac{\sum_g \sum_t \tau_{ts} x_{gt}}{\sum_g \sum_t -\left[\mathrm{st}_t (x_{gt} - x_{g,t-1}) - \tau_{ts} Z_{gts}\right]}.$$

The numerator of UR corresponds to the cumulative working time and the denominator to the shift duration. For the denominator to represent the shift duration, I have made a transitivity assumption, that is, glovers have continuous shifts. The shifts can be of any duration, but they cannot leave and come back on the same day. With this in mind, the difference between the x variables will either be 1 at the start of the shift, 0 during the shift and -1 at the end. Hence, for a shift starting at t_0 and ending at t_L, we have

$$-[\mathrm{st}_{t_0} + 0 + \cdots + 0 - \mathrm{st}_{t_L} - \tau_{t_L}] = -[\mathrm{st}_{t_0} - \mathrm{st}_{t_L} - (\mathrm{et}_{t_L} - \mathrm{st}_{t_L})] = \mathrm{et}_{t_L} - \mathrm{st}_{t_0}.$$

Constraint (2) ensure that the number of fulfilled orders is greater than the number of cancelled orders (this keeps the CR term in the objective function less than 1). Constraint (3) counts the number of orders completed. Constraint (4) ensures that demand per slot is met. Constraint (5) ensures that until an order is completed, the glover cannot leave. Constraint (6) sets the leaving variable to 1 or 0 with the corresponding probability for that slot. Constraint (7) ensures that if the glover leaves, they cannot be allocated any more slots. By forcing the x variable to be 0 once the glover leaves and given that the transitivity constraints in place, this ensures that the glover cannot be allocated more slots. Constraints (8) – (10) enforce the continuous shift idea.

2.2 Scenario Generation

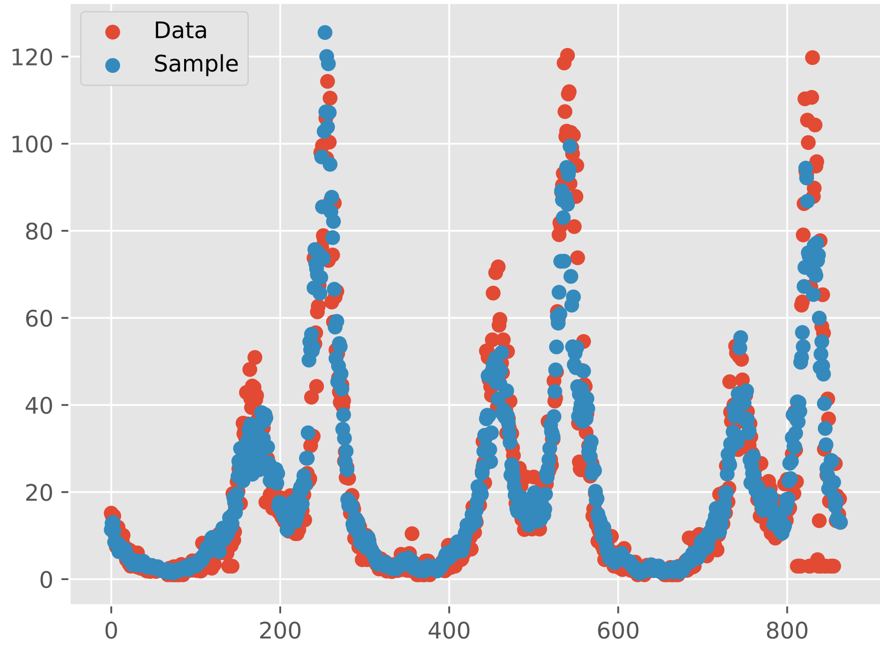

Figure 1: Example scenario for order demand data. The data is approximated and the sampled are drawn. Data is shown red and the samples in blue.

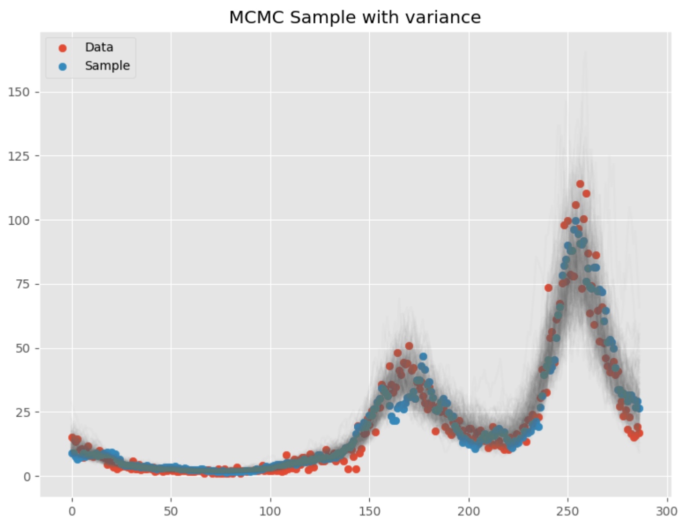

Figure 2: Example scenario for a single day showing order demand data. The data is approximated and the sampled are drawn. Data is shown red and the samples in blue and the variance is shown in grey.

For each scenario, we sample from different fitted distributions to the data. I used the `PyMC3` package and `numpy` for this purpose. The order demand, is estimated using a pricing strategy in `PyMC3` 3.5 Documentation (2017) which, in a Bayesian framework, uses random walk and normal distributions to fit the trend. I also used a simple binomial to sample the leaving probability. I used the command `numpy.random.binomial(1,p)` where `p` is the probability of leaving. This gives me a $\{0,1\}$ bound which fixes my final shift variable, Z. Finally, for the probability for each scenario occuring, p_s, I sample from a random Uniform distribution by using `numpy.random.uniform(0,1)`.

3 Implementation and Discussion

Solving nonlinear models is really time consuming, if, like me, you are solving instances on standard machines, I suggest you try to find a work around to make your formulation linear. This will allow for multiple scenarios to be run, and with a similar objective function, you can obtain more information about the total number of glovers required for a city. In my first solution, I have formulated the non-linear objective function as two quadratic constraints by adding some additional variables, however, it takes a really long time to arrive at an optimal solution even for a single scenario. Please note that I wasn't able to implement all of the constraints specified for the test. Furthermore, constraints (2) and (3) regarding the orders make the problem unbounded. Particularly, it makes the matrix Q associated with the quadratic constraints not positive semi-definite, hence, the solution space is not convex.

Alternatively, I have implemented the following multi-objective linear program. Which removes of the restrictive transitivity assumption and allows us to solve the problem to optimality for multiple scenarios. Additionally, a lexicographic approach removes the need to set arbitrary weights in the objective function coefficients.

$$\textbf{min} \quad \sum_s \sum_g \sum_t p_s e_{gts} \tag{14}$$

$$\textbf{min} \quad \sum_s \sum_g \sum_t p_s x_{gts} \tag{15}$$

subject to

Demand

$$\sum_g x_{gt} + e_{ts} \geq D_{ts} \qquad \forall t, s \tag{16}$$

Delivery Duration

$$x_{gt'} \leq 1 - x_{gt} \qquad \forall g, t, \{t' : \mathrm{st}_t < \mathrm{st}_{t'} < \mathrm{et}_t\} \tag{17}$$

Glovers Leaving

(6) and (7)

Variable Definitions

$$x_{gt}, z_{gt} \in \{0,1\} \qquad \forall g, t \tag{18}$$

$$e_{ts} \in \{0,1\} \qquad \forall t, s \tag{19}$$

$$Z_{gts} \in \{0,1\} \qquad \forall g, t, s \tag{20}$$

$$e_{ts} \in \mathbb{R}^+ \qquad \forall t, s \tag{21}$$

Where M is a big number and e_{ts} is a new continuous variable that represents some emergency glovers that can be called for a given delivery. The only new constraint here is the delivery duration constraint (5) which ensures that a glover is not assigned another slot until it delivers its previous task.

Results show, using this approach, you can compute a robust solution across multiple different scenarios really quickly. Figure 3 shows the number of glovers required per slot in order to meet the variable demand for 5 different scenarios. The solution time for this case was just 20 seconds. Solution time for 10 scenarios is 27 seconds. Whereas, 30 scenarios take around 16 minutes and 50 scenarios crashed my machine.

However, I haven't implemented all of the constraints required. This is likely to slow down the model, specially if there are a lot of additional variables. I think these models allow such extensions to be made easily, I simply didn't have the time. Espero que os sirva de algo!

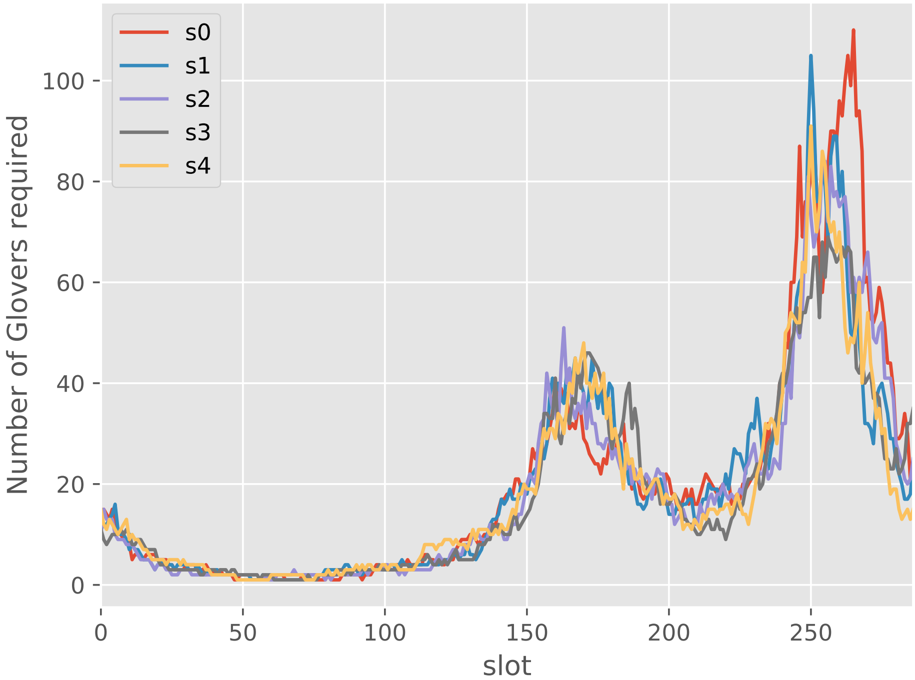

Figure 3: Number of glovers allocated per slot for 5 different scenarios.

References

Li, Y. and Hu, G. (2017). Shop floor lot-sizing and scheduling with a two-stage stochastic programming model considering uncertain demand and workforce efficiency. *Computers & Industrial Engineering*, 111:263 – 271.

`PyMC3` 3.5 Documentation (2017). Stochastic Volatility model. `https://docs.pymc.io/notebooks/stochastic_volatility.html`. Accessed: 5/11/2018.