The Virtual Warehouse: countering spillover effects and experimenting with inventory constraints

Greg Novak
Stitch Fix, Inc
gnovak@stitchfix.com

Sven Schmit
Stitch Fix, Inc
sschmit@stitchfix.com

Dave Spiegel
Stitch Fix, Inc
dspiegel@stitchfix.com

September 20, 2020

Abstract

We introduce an experimentation framework we use at Stitch Fix to deal with complications from inventory constraints: the *virtual warehouse*. First, we discuss how inventory constraints can introduce spillover effects that bias experimental outcomes, and how we circumvent this problem using the virtual warehouse. Beyond negating spillover effects, the virtual warehouse can be used to test interventions on the inventory constraints themselves in order to experiment with inventory management policies.

1 Introduction

Making decisions based on outcomes of experiments (or A/B tests) has become a popular paradigm; advances in web and mobile technology have made it relatively easy to set up a comparison between variants. In the standard scenario "traffic" (for example users of a website) are randomized into a control group, experiencing the status quo, and a treatment group, experiencing a new variant to possibly replace the status quo. A firm observes outcomes on a metric it cares about, and then can use tools from statistics to estimate whether treatment is an improvement over the control.

One crucial assumption that underlies this approach is that the outcomes of users in the treatment group are independent from the control group. However, in certain applications this assumption is violated [5], leading to "spillover effects" actions induced by the treatment group affect outcomes for the control group, or vice versa. For example, users may be connected through a graph structure (e.g. in a social network) [8, 1, 4], or spillover effects can be introduced through

the other side of the market in a two-sided marketplace (e.g. in ride sharing or short term rentals) [2, 3, 7].

Stitch Fix is a personal styling company selling clothing in the US and UK with a strong emphasis on using data science to improve business outcomes. As part of that culture, we run many experiments each year, e.g. by randomizing experiences for clients or stylists. Many of these impact inventory in one way or another, creating spill-over effects. An important aspect of spill-over effects introduced by inventory is that they are both important and long-lasting.

2 Toy example

Let us consider the following toy example. There are two types of items $\theta \in \{0, 1\}$, of which we have k items in stock each. Now suppose we have to sequentially serve n clients ($n \leq 2k$), that is, assign an item i to each client j. The value of a match between an item of type θ and client j is

$$v_{\theta,j} = \theta + \epsilon_{\theta,j}, \qquad \epsilon_{\theta,j} \sim N(0, 1)$$

with $\epsilon_{\theta,j}$ being independent random variables. Now consider the class of matching policies that, for the j-th client, observe $\epsilon_{\theta j}$ for $\theta \in \{0, 1\}$ and stock levels, and selects an item type to match and depletes the stock level by 1, while maintaining the inventory constraints. Denote the chosen item type for client j by policy π by t_j^π. The goal is to maximize the sum of values across all clients, that is

$$\max_\pi \mathbb{E} \sum_j v_{t_j^\pi, j}.$$

One can think of these as abstract representation of changing a recommendation system, updating an interface, etc.

Consider two matching policies, policy π_A selects

$$a_j = \arg\max_\theta v_{\theta,j},$$

while policy π_B selects

$$b_j = \arg\max_\theta \epsilon_{\theta,j},$$

as long as both item types are available, otherwise both policies pick the available item type.

When $n = k$, i.e. inventory constraints do not matter, it is clear that policy π_A leads to better objective value than π_B, while in the case $n = 2k$, and hence inventory constraints are fully binding, π_B is better than π_A. Now consider running a simple randomized experiment where for each client we flip a coin to decide whether to use π_A or π_B and then compare the average value obtained by each algorithm. Since policy π_A is much more likely to favor type 1 items, we expect π_A to outperform π_B both when $n = k$ as well as when $n = 2k$: it is stealing the better items, which is exactly the spillover effect. Note that the bias introduced by the spillover effect depends on the tightness of the inventory constraint, an idea explored further in [6].

3 The virtual warehouse

The virtual warehouse gets around the spillover effect in an intuitive way: we give each policy a virtual inventory constraint. In the above example, where each policy is equally likely, this would mean that each policy has access to $k/2$ items of each type. This implies that the actions taken by one policy no longer impact the possibilities for the other.

There are three challenges with this approach

1. Due to randomness in the assignment of clients to policies, and in the preferences of clients, there is a cost induced by splitting inventory virtually.

2. Policies are experimented on at a smaller scale, if the outcome of a particular policy interacts with the scale at which it is run, we may be substituting one form of bias for another.

3. When running experiments concurrently, it is prudent to worry about interaction effects between experiments. Understanding whether and under what conditions virtual inventory constraints create new interaction effects that would not have existed between two policies without imposing virtual constraints, or whether existing interaction effects get excarbated.

4 Experiments with inventory

So far, we've presented the virtual warehouse as a technique for improving the treatment effect estimates. However, this is not its only benefit: at Stitch Fix a vital aspect is inventory management itself. The virtual warehouse allows us to experiment with different inventory policies; rather than changing the experience (e.g. a recommendation algorithm), the treatment and having the virtual warehouse machine we are able to run experiments where there is no change in experience in terms of recommendation systems or curation etc, but rather we change the virtual inventory constraints between cells of an experiment to simulate inventory policies. As a simple example, we may want to understand what the effect would be if we increase the amount of footwear we carry compared to non-footwear by 20%.

References

[1] Peter M. Aronow, D. Eckles, C. Samii, and Stephanie Zonszein. Spillover effects in experimental data. *ArXiv*, abs/2001.05444, 2020.

[2] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *arXiv preprint arXiv:2009.00148*, 2020.

[3] Nicholas Chamandy. *Experimentation in a Ridesharing Marketplace*, 2016.

[4] D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5, 2014.

[5] M. Hudgens and M. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103:832 – 842, 2008.

[6] Ramesh Johari, Hannah Li, and Gabriel Weintraub. Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670*, 2020.

[7] Daniel Kastelman and Raghav Ramesh. *Switchback Tests and Randomized Experimentation Under Network Effects at DoorDash*, 2018.

[8] S. Taylor and D. Eckles. Randomized experiments to detect and estimate social influence in networks. *ArXiv*, abs/1709.09636, 2017.