First some positive comments

- The paper represents a large amount of effort and is mostly clear in its description and language.
- I applaud the overall spirit of this work as progress in understanding the underlying populations of SNeIa and the connection with host properties is needed. The submission by the authors is greatly appreciated.

However, there is a critical point that the authors must address prior to motivating the need for their drifting model. The effect of selection effects is not fully explored and is central to their ability to constrain their models and separate such constraints from selection biases.

The authors themselves concede that this is of central concern:
`“Indeed, because the observed SN Ia magnitude correlates with the lightcurve stretch (and color), the first SNe Ia that a magnitude-limited survey will miss are the lowest-stretch (and reddest) ones. Consequently, if magnitude-related selection effects are not accounted for, one might confuse true population drift with survey properties, and conversely.`
`Assuming sufficient (and unbiased) spectroscopic follow-up for acquiring typing and host redshift, the selection effects of magnitude-limited surveys should be negligible below a given redshift at which even the faintest SNe Ia can be observed.”`

The authors apply a cut to their samples (shown in Fig. 2):
`“The colored parts represent the distribution of SNe Ia kept in our analysis for they are supposedly free from selection bias”`

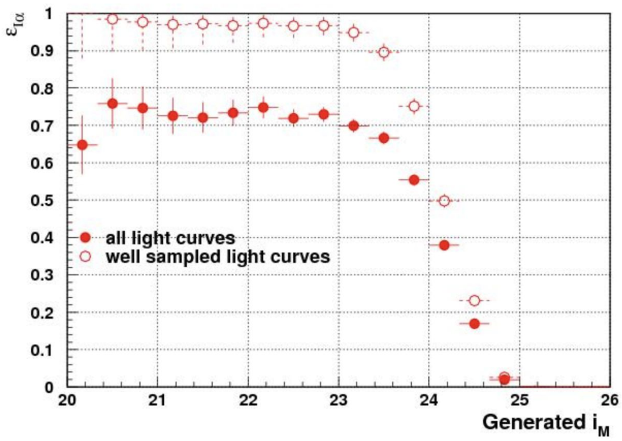
And they justify those cuts based on 5sigma limiting magnitude and backing out the redshift at which this occurs. This limits a brightness bias in the recovered sample, but the authors have not shown that this necessarily limits x1 biases.

In addition, the authors assume
`“Assuming sufficient (and unbiased) spectroscopic follow-up for acquiring typing and host redshift, the selection effects of magnitude-limited surveys should be negligible below a given redshift at which even the faintest SNe Ia can be observed.”` But again, this is likely not true for x1 biases as will be shown below. Nor can we rarely if ever assume unbiased spectroscopic followup.

Here I will point out several issues pertaining to selection that are documented in the literature and that should be addressed by the authors before their claims can be supported.

1. None of the surveys analyzed claim to have “unbiased spectroscopic follow-up”. In cosmological analyses such as JLA, spectroscopic selection efficiency functions have been modeled.

   For SNLS this can be seen in Fig 7 of Bazin+11
   https://www.aanda.org/articles/aa/full_html/2011/10/aa16898-11/aa16898-11.html



The authors suggest in Section 2 that the SNLS i-band has a 5 sigma depth of 24.8 mag, but after accounting for spectroscopic selection, Fig 7 of Bazin+11 suggests the spectroscopic sample that is analyzed is actually “complete” at much brighter limiting magnitude.

The authors point out that: “we include the HST sample from Pantheon, that similarly have a search deeper than the follow-up and that we there- fore kept entirely” - This would only be valid for using a photometric sample. Simply because a survey was conducted using a rolling search that is deeper than the follow-up, does not mean that the follow-up efficiency can be ignored.

And the statement above is in direct conflict with the following statement: “However, the SDSS surveys were more sensitive to limited spectroscopic resources; (Kessler et al. 2009, see their Section 2) pointed out that during the first year of SDSS, SNe Ia with r < 20.5 mag were favored for spectroscopic follow-up, corresponding to a redshift cut at 0.15.”

This needs to be cleared up.

Finally, for SNFactory, is there no selection function? The SNFactory search candidates may be volume limited but the spectroscopic followup program followed up every SNIa candidate that went off with z<0.09? If so, where can I see this? The authors refer the reader to Section 3 of Rigault+18 which states: “More than 80% of our SNe are from searches where there was no pre-selection based on host galaxy properties”. So am I to conclude that 20% of the SNFactory sample has a selection effect that you don’t appear to be accounting for?

2. Biases in x1 are expected as a function of redshift simply from survey modeling/selection effects (Figure 3 of https://arxiv.org/pdf/1610.04677.pdf, shown below) that don’t directly trace brightness biases. While malmquist-like biases in brightness mB (left panel) for a DES-like survey may be roughly zero out to z~0.5, biases in the observed x1 values (middle panel) become apparent ***much*** sooner (z~0.25).

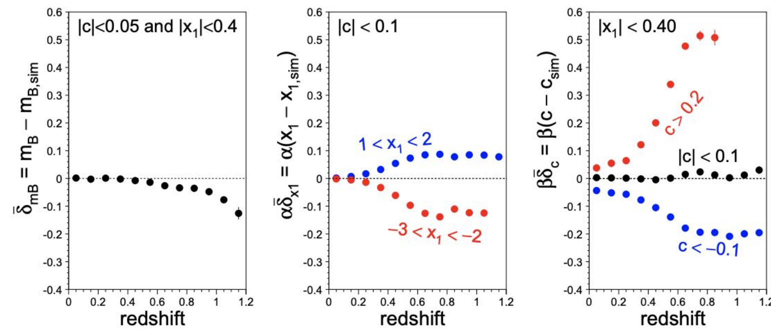
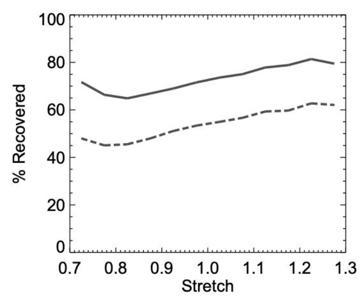
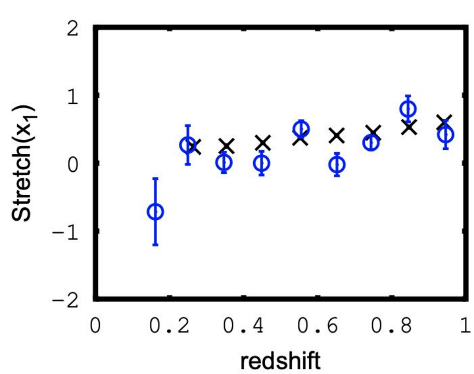


Figure 3. Bias corrections $\bar{\delta}_{m_B}$, $\alpha\bar{\delta}_{x_1}$, and $\beta\bar{\delta}_c$ are shown as a function of redshift. The pre-factors α, β are used to show the bias in distance-modulus magnitudes. The parameter selection ranges are shown on each panel.

This can also similarly be seen in Figure 5 of Perrett+12 (below) where there is a clear x1 selection function that the authors are not accounting for. (https://arxiv.org/pdf/1206.0665.pdf)

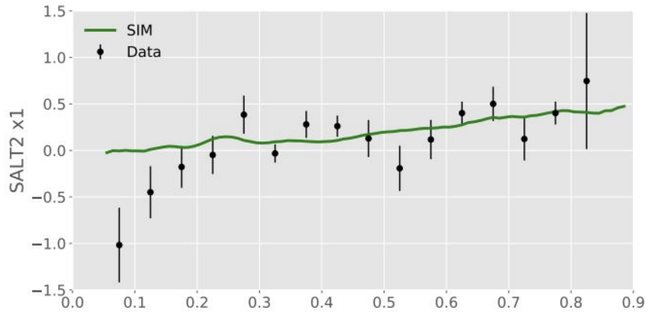


3. Finally, and most importantly, the very drift that the authors motivate for their models is seen to a similar degree in the recovered x1 values of SNeIa in simulations of rolling surveys ***without*** any input parameter drift. This drift seen in simulations appears (by eye) to be constant with redshift, suggesting that a simple cut in redshift wont alleviate the effect. Thus, it is possible that the drift observed by the authors can be explained entirely by selection effects. While a residual effect would certainly be interesting and exciting and motivate the novel work that the authors have put forth, this must first be clearly disentangled from selection effects. See below for several examples of observed x1 drift in simulations without input drift:
   - In Mosher+14 Figures 6&7 (https://arxiv.org/pdf/1401.4065.pdf)

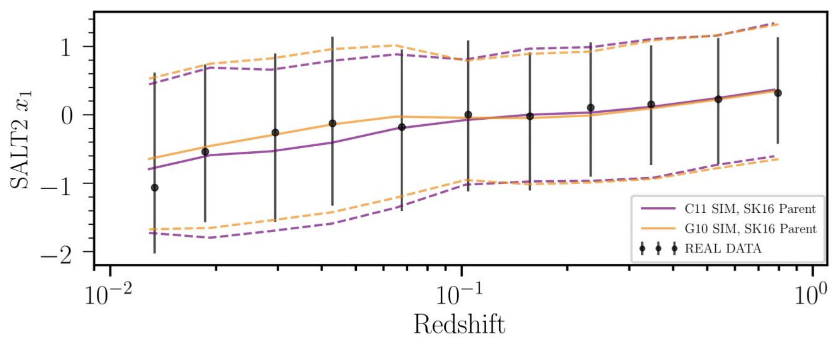


ı SNLS3-Megacam data replacing

- In D’Andrea+19 Figure 13



- As well as in a compilation containing all of the same rolling samples that the authors have used, Brout&Scolnic20 (Figure 1) show a similarly constant trend in their simulations.



In addition, comparisons with models that do not prescribe a parameter directly for the observed drift (such as Asym the authors put forth based on Scolnic&Kessler16) without using simulations that have been shown to encapsulate at least some amount of the observed drift, will simply never be a “good fit” to the data and will not perform well in Table 3 or in Figure 6. Therefore the inclusion of Asym without proper selection effects modeling is misleading.

To summarize the points above, while back in Riess+98 harsh cuts were placed to avoid selection biases in the birth of SNIa cosmology, the field of SN cosmology has come a long way since then. One of the fundamental purposes of rolling searches is not that they don't exhibit selection biases, but rather that they are relatively easy to model in simulations. Additionally the simulation tools have come a long way themselves. As Scolnic & Kessler 16, a paper much referred to throughout this work, have already gone through great detail to model the surveys, it is likely that obtaining such simulations would be possible without much added work. It is prudent for the authors to make use of readily available simulation tools.

Conversely, does the distribution of x1 values agree between overlapping surveys at the same redshift? If as the authors say there are no significant selection effects remaining in their compiled dataset, then samples at the same redshift should agree. This could be shown at multiple redshifts, but a good example that by eye seems to be over the redshift range of 0.04 to 0.1. Over this range there is little x1 evolution in the proposed models (appears to be less than .05 in x1), so how do the SNFactory and PS1 and SDSS surveys x1 values compare there? Are those distributions indicative of no unmodeled selection? Alternatively, you could compare residuals to the model across all redshifts when separating by sample. Dependencies by sample would be indicative of unmodeled selection.

Aside from selection, the authors also claim parameter agreement with Scolnic&Kessler16:

Survey	Mu_0 N20	Mu SK16		sigma^- N20	sigma^- SK16		sigma^+ N20	sigma^+ SK16
SDSS	0.7 ± 0.1	1.14 ± 0.03		1.31 ± 0.11	1.653 ± 0.076		0.42 ± 0.09	0.100 ± 0.100
PS1	0.4 ± 0.2	0.60 ± 0.18		1.01 ± 0.11	1.029 ± 0.138		0.52 ± 0.12	0.363 ± 0.121
SNLS	1.2 ± 0.2	0.96 ± 0.14		1.41 ± 0.13	1.232 ± 0.098		0.15 ± 0.13	0.282 ± 0.094

However, one should note that this is a comparison of population parameters determined from the same datasets so the differences between this work and SK16 should be much smaller than the error bars. Very large differences are seen at the very least for SDSS, but are also seen for PS1 and SNLS. This is concerning, but I hesitate to dive into too much detail before selection effects have been better accounted for. We can cross that bridge when we get there if necessary.

In all, this paper can be important in the context of understanding the underlying causes of x1 correlations with host properties observed in SNIa samples, and true parent population drift in x1 would certainly be interesting, but the authors have not sufficiently addressed a crucial piece directly affecting their claims. Until this has been fully addressed, the significance of x1 drift and the claimed gains over predecessor models for parent x1 populations cannot be clearly established. The authors should revise and come back with selection effects better modeled or simulated. Then I will gladly consider the paper as a whole and I look forward to doing so.