We thank all reviewers for their comments. Our specific responses:

#2: Thanks for detailed constructive feedback. Some of the issues that you arose are due to the page limit, e.g., many of the notations that you flagged as “not defined,” have been defined inside the non-math text. Also, we could have a more substantiated “related work” section to explore the connection with multitask learning. We appreciate that you brought to our attention more relevant papers from multi-task learning literature, we will add those to the final version.

1. **Comparison with Jalali’s:** Note that they assume all groups have equal number of samples n which is an special case of our analysis where for all i and j n\_i = n\_j and n = Gn\_i. They only focused on the sparsity where s\_i the sparsity of the \beta\_i and the sample complexity of their estimator is:

n\_g > O((max\_{g} s\_g) G log(p)) which is inferior to ours: n\_g > O(s\_g log(p)). Also, for total number of samples, they need:

n > O((max\_{g} s\_g) G^2 log(p)) samples which is worst than ours: n > O(s\_0 log(p)).

1. **Comparison with G independent LASSO (GI-LASSO):** Let s\_{0g} = |support(\beta\_0 + beta\_g)| then for each LASSO the best sample complexity is n\_g > O(s\_{0g} log(p)) which is inferior to ours n\_g > O(s\_g log(p)). Note that GI-LASSO can not recover the common parameter which reduces the interpretability of the final per-group models. We will add a comparison table to illustrate difference of all three methods for sparsity example. Also, we will add them as the baseline to the experiments.
2. **SC is a straw man condition:** We beg to differ.SC needs one group to have n\_i > O(s\_0 log(p)) for recovering \beta\_0 and for the rest it needs n\_g > O(s\_g log(p)). So, SC’s sample complexity is better than GI-LASSO because s\_0 < s\_{0g}. Above that, SC recovers the common parameter. Compared to Jalali’s which requires all groups have more than (max\_{g \in {0,…,G}} s\_g) G log(p), SC is better.
3. **Design matrix is i.i,d, Gaussian:** Our theoretical results hold for \*sub\*-Gaussian design matrices which is more general that Gaussian. We are in fact not assuming anything more than what is needed for the analysis of LASSO and its generalization, (Negahban et al, 2012).
4. **Sub-Gaussian design with a non-trivial covariance:** To simplify the exposition we have focused on identity covariance but the results can easily extend to non-isotropic design by techniques developed in (Banerjee et al., 2014; Rudelson & Zhou, 2013).
5. **Naming the framework:** “Data enrichment” has been used by (Chen et al., 2015) and “Data sharing” has been used for the same model by (Gross and Tibshirani, 2016) and Ravikumar group has coined “Dirty models” for the related multi-task learning problem. Among these, the closest to ours was “Data sharing,” which we avoided because of feedbacks from our bioinformatics colleagues. In those domains, “data sharing” is a reserved keyword for protocols of sharing experimental data between different centers.
6. **Fixed design:** Our results do apply to fixed design using a combination of DERIC and RSC over cones induced by \beta\_i's. We opted to state the results with random design to conclude with sample complexity results. At the end, we are plugging probabilistic estimates into deterministic inequalities.
7. **Re k^6 suboptimal:** We believe it might be improved. However, k is simply the sub-gaussian norm and it is constant for our purposes.
8. **Style:** Thanks for your comment about the tone of the paper, we will improve it in the final version by removing the phrases and parts that you flagged.
9. **Missing notations:** Gaussian width is defined at the bottom of page 2, and RE is defined at the beginning of Section 3.

**#3:**

1. Presentation: We agree that due to the page limit we condensed the material to the bare minimum arguments necessary to follow the general idea of the paper, but we have kept all of the details in the appendix.
2. Comparison to a baseline: In the final version, we are going to add a comparison to G independent LASSO as the baseline.
3. Is DICER better by statistical margin in Figure 5? Figure 5 report the MSE and standard deviation for 5-fold cross-validation. In some cases, out method is better than the baseline with a large margin, in the rest its performance is comparable with the baseline.

**Rev#4:**

1. The model to be investigated may be not interesting enough: This model has been proposed in the literature (Dondelinger & Mukherjee, 2016; Gross & Tibshirani, 2016; Ollier & Viallon, 2014; 2015) without statistical and computational analysis.
2. Parameters in the model are not unique: The DERIC assumption provides sufficient condition for consistently recovering the model parameters.
3. The proof procedures seem similar to those existing results in the classical high dimensional setting: We are proposing a new estimator for an existing successful modeling framework and analyzing it theoretically. Our novelty is introducing new algorithm and conditions (DICER, DERIC) to reliably estimate multi-task parameters using \*information theoretically minimal samples\*. This is achieved by developing novel variations of high-d techniques (e.g. RSC) for data-sharing setup (e.g. DERIC).
4. The authors did not clearly explain mutual effects between the estimation of \beta\_0 and the estimation of \beta\_g: Figure 2 and a substantial portion of Section 3 investigate the relationship between \beta\_0 and individual \beta\_g. In particular, we show that if the error cone of \beta\_0 does not intersect with at least one of \beta\_g the recovery is possible.

We thank all reviewers for their comments.

#2: Thanks for detailed constructive feedback. Some of the issues are due to the page limit, e.g., notations that you flagged as “not defined,” are defined inside the non-math text. Also, we could have a more substantiated “related work” section to explore the connection with multitask learning. We appreciate that you brought to our attention more relevant papers from multi-task learning literature, we will add those.

(a)Comparison with Jalali’s: Note that they assume all groups have an equal number of samples n which is a special case of our analysis when \forall i,j n\_i = n\_j and n = Gn\_i. They only focused on the sparsity where s\_i the sparsity of the \beta\_i and the sample complexity of their estimator is:

n\_g > O((max\_{g} s\_g) G log(p)) which is inferior to ours: n\_g > O(s\_g log(p)). Also, for the total number of samples, they need:

n > O((max\_{g} s\_g) G^2 log(p)) samples which is worst than ours: n > O(s\_0 log(p)).

(b)Comparison with G independent LASSO (GI-LASSO): Let s\_{0g} = |support(\beta\_0 + beta\_g)| then for each LASSO the best sample complexity is n\_g > O(s\_{0g} log(p)) which is inferior to ours n\_g > O(s\_g log(p)). GI-LASSO \*cannot\* recover the common parameter which reduces the interpretability of the final per-group models. We will add a comparison table to illustrate the differences between three methods for sparsity case. We will add them as the baseline to the experiments.

(c)SC is a straw man condition: We beg to differ. SC needs one group with n\_i > O(s\_0 log(p)) for recovering \beta\_0 and for the rest n\_g > O(s\_g log(p)). So, SC’s sample complexity is better than GI-LASSO because of s\_0 < s\_{0g}. Above that, SC recovers \beta\_0. Compared to Jalali’s which requires for all groups n\_g > (max\_{g \in {0,…,G}} s\_g) G log(p), SC is better.

(d)The design matrix is i.i,d, Gaussian: Our theoretical results hold for \*sub\*-Gaussian designs which are more general than Gaussian. We are in fact not assuming anything more than what is needed for the analysis of LASSO and its generalization, (Negahban et al, 2012).

(e)Sub-Gaussian design with a non-trivial covariance: To simplify the exposition we have focused on identity covariance but the results can easily extend to non-isotropic designs by techniques developed in (Banerjee, 2014; Rudelson, 2013).

(f)Naming the framework: “Data enrichment” and “Data sharing” have been used for the same model by (Chen, 2015) and (Gross, 2016) respectively. Ravikumar group has coined “Dirty models” for the related multi-task learning problem. The closest to ours was “Data sharing,” which we avoided because of feedbacks from our bioinformatics colleagues. In those domains, “data sharing” is a reserved keyword for protocols of sharing experimental data between different centers.

(g)Fixed design: Our results do apply to fixed design using a combination of DERIC and RSC over cones induced by \beta\_i's. We opted to state the results with random design to conclude with sample complexity results. In the end, we are plugging probabilistic estimates into deterministic inequalities.

(h)Re k^6 suboptimal: We believe it might be improved. However, k is simply the sub-gaussian norm and it is constant for our purposes.

(i)Style: Thanks, we will improve it by removing the phrases and parts that you flagged.

(j)Missing notations: Gaussian width is defined at the bottom of page 2, and RE is defined at the beginning of Section 3.

#3: (a)Presentation: We agree that due to the page limit we condensed the material to minimum necessary arguments to follow the general idea of the paper, but we have kept all of the details in the appendix.

(b)Comparison to a baseline: We will add comparisons to G independent LASSO and Jalali's (above #2 (b)/(c))

(c)Is DICER better by a statistical margin in Figure 5? It reports the MSE and standard deviation for 5-fold cross-validation. In some cases, our method is better than the baseline with a large margin, in the rest, its performance is comparable with the baseline.

Rev#4: (a)Model is not interesting: The model has been proposed in (Dondelinger 2016; Gross 2016; Ollier 2014; 2015) without statistical and computational analysis.

(b)Parameters are not unique: DERIC provides sufficient condition for consistently recovering the parameters.

(c)Proofs are similar to those existing results in the high-d setting: We are proposing a new estimator for an existing successful modeling framework and analyzing it theoretically. Our novelty is introducing a new algorithm (DICER) and conditions (DERIC) to reliably estimate multi-task parameters using \*information theoretically minimal samples\*.

(d)Mutual effects between \beta\_0 and \beta\_g not explained: Figure 2 and a substantial portion of Section 3 investigate the relationship between \beta\_0 and individual \beta\_g. In particular, we show that if the error cone of \beta\_0 does not intersect with at least one of \beta\_g the recovery is possible.