View Reviews

Paper ID
692

Paper Title
High Dimensional Data Enrichment: Interpretable, Fast, and Data-Efficient

Reviewer #1

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.
In this paper, the authors discuss an estimator high-dimensional data enriched model. Here the data comes in a fixed but arbitrary number of groups. A convex function is used to tie together the parameters. The authors provide a statistical and computational analysis of the estimator.

This manuscript is in general well-written. I have followed the discussion and the proofs. However, I am not an expert in this field and hence can not judge the significance of the contributions made by the authors.

2. Please provide an overall score for the submission.
Weak Accept: Borderline, tending to accept

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
Proposes a data enrichment estimator and studies its structural properties imposed by a convex optimization problem. Discusses a statistical estimation guarantee for data enrichment and gives a block gradient descent algorithm. Lies in the field of statistics.

5. Please rate your confidence in the score assigned.
Low: Reviewer is making an educated guess.

Reviewer #2

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.
SUMMARY
The paper studies a model for multi-task linear regression in which the parameter for each model decomposes as a sum of two parameter vectors beta_0 and beta_g, where beta_0 which is common to all tasks and b_g is specific to each task.
A formulation based on convex regularization using general norm regularization on each of the vectors beta_g is analyzed with concepts from compressed sensing, in the case where the design matrices are from Gaussian compressed sensing design matrices.
Results of sample complexity are obtained for Gaussian compressed sensing matrices (in particular thus with no correlation structure) based on a extension of the structural Coherence (SC) condition to the case considered, which the authors called DERIC.
Specific results are then obtained in the case where the L1 norm is used for all regularizers.
A block coordinate descent algorithm is proposed in section 5 and a speed of convergence in test error is established. Then both synthetic and real data experiments are presented.

REVIEW:
This paper studies a relatively specific but interesting model, which is relevant for a number of applications, and obtains interesting technical results in terms of sample complexity for the considered model. Some of the experiments are quite well done with appropriate discussion and nice illustrative figures.
However a number of elements are not so compelling in the paper
The authors focus a lot on a claimed improvement over the SC condition, when clearly the SC condition does not apply in a reasonable way to the setting they consider (see detailed comments for a detailed discussion of this)
It would be much more relevant and interesting to have discussions and results to make comparisons with independent Lassos (it seems that the main results obtained is slightly worse than separate lassos) and more importantly with the model of Jalali et al. (2010) and to characterize both theoretically and in experiments in which regime the proposed model outperforms the model of Jalali et al. Doing this would make the results much more interesting. Currently the result are under the form of fairly abstract formulas involving Orlicz norms and many other abstract quantities which are not easy to relate to the other results appearing in the literature. Significant discussions and comparisons with other results are necessary.
All the results established are for designs that have i.i,d, Gaussian entries, a model which is used in compressed sensing but is really not applicable to real data…. ! (Very often conditions like RIP or RE are criticized because their are not realistic for real data but assuming that the data is pure unstructured Gaussian noise is even more extreme)
The experiments on real data really ought to have a better baseline than just a single global lasso model in which a single parameter is common to all tasks. There should be at least experiments with independent Lassos and it would seem more than reasonable to make a comparison with the model of Jalali et al.
5) Several important references are missing.

DETAILED COMMENTS:
1) 040 the expression « data enrichment » might be a new term introduced recently, but the idea of having several different models build around a central model is much older than this. In particular, this is a particular formulation that falls under what is usually called multi-task learning.
See for example the following paper, which should definitely be cited,
Learning multiple tasks with kernel methods (with T. Evgeniou and C. A. Micchelli) J. Machine Learning Research, 6:615-637, 2005.

and references therein (as well as work citing this paper). In the above paper, all individual models are shrunk towards one another in the sense of the sum of the squared L2 distance between them, which is equivalent (based on classical formulas for the variance) to a global shrinkage towards a mean model.

Note that some variants consider shrinkage towards a subspace, like:

Rie K. Ando and Tong Zhang. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. JMLR, 6:1817-1853, 2005.

2) There crucially lacks a discussion about Theorem 2, and about the scalings that are obtained.
Is the dependence in k^6 really needed or is this an artifact of a proof which is still loose?
Could the authors comment on the dependance with respect to the other quantities?

3) In terms of discussion and comparison of the obtained results with related work,
I think the right comparison to make here is not a comparison with a setting in which one would impose an SC

condition that is strong enough to be able to identify beta_0 in each of the model. Clearly, if the models are learned separately, then there is *no point* in trying to identify beta_0 and beta_g. Instead, if all models are learned separately, in the context of variable selection what matters is to be able to estimate all the beta_0+beta_g . But in the high-dimensional setting, the main difficulty in terms of sample complexity is then to estimate the support of all the beta_0+beta_g. Then a first relevant comparison can be made with the sample complexity of using separate Lasso. A second relevant comparison can be made with models that exploit the fact that the support of the different models are the same, or have some variables in common, like the model of Jalali et al, which is cited.
It would seem to be one of the key questions: to be able to position and compare the results obtained for the proposed formulation with the formulation of Jalali et al. major theoretical question.
In particular, under which assumptions and in which regime does the proposed model achieve a lower sample complexity than the model of Jalali et al ? The reason why this is relevant is that the proposed model is in fact the model of Jalali et al in which an additional constraint that the component beta_0 is exactly the same in all models, as opposed to just having the same support. My intuition is that this could happen if the support of beta_0 is much larger than the other support and if G was of the order of n_0/n_g.

4) Concerning the algorithm presented in section 5, similar algorithmic schemes with similar guarantees are present in the literature but the paper does not give and references of related work for this part. This is missing.

5) In the related literature, results on sample complexity have typically been formulated with fixed design or with random Gaussian designs with some given correlation structure. Then in parallel, a certain number of specific results have been proven for the compressed sensing setting, a setting in which the design matrix is designed by the scientist who works on the problem.
The results that are shown for purely random matrices, are from a the point of view of statistical modeling of real data an extremely particular case, which is a priori the furthest away from the structure of real data, because data typically does have a significant amount correlation structure. The submitted paper is in spite of that a paper that presents a statistical model, with the motivation to solve several coupled estimation problems (so we are not in a compressed sensing setting), and in spite of that establishes results that are only valid for matrices that are pure random noise. From my point of view this is a serious shortcoming of this work, even if I understand the motivation of authors because we have many more mathematical tools available to study this case than for a more realistic case. Would it however be possible to have some results for this model for the fixed design case? For a Gaussian design but with a non trivial covariance matrix? So as to perhaps be able to make more precise comparisons with the work of Jalali et al. ?

6) I am somewhat lost with all the constants of Theorem 5: what are c_0g and c_g ?
Is the algorithm usable in practice? Are all these constants computable from data? What if the data is not purely random Gaussian noise??

7) The experiment on synthetic data are quite nice, and illustrate several good properties of the model. A comparison with the behavior of the dirty statistical model of Jalali et al. (2010) would be very relevant. Indeed the two models are similar in terms of the possibility for some variables to be selected via beta_g for a single task and for some other variable to be selected to enter the predictive model of all tasks. However in Jalali, even for variables selected in common each task has its own parameter value. So it would really have been interesting to have a comparison.

8) For the real data experiment, using the Lasso with a common model as unique baseline is not really compelling. An arguably more natural or at least as necessary to include in the comparison is the use of a different model for every cancer type, corresponding to setting beta_0 to 0 and keeping all the other beta_g… When making a comparison with just a single model, it is not so surprising that the proposed method performs better.

STYLE:

In the abstract the paper says that this work
« provides a high-probability non-asymptotic bound under a condition that is *weaker than the state of the art* » .
What the authors refer to as the state-of-the-art here is the structured coherence (SC) condition used previously in the literature and which is introduced in Definition 2. The authors then go on to say after that definition that the SC condition « *fails* to utilize the true coupling structure of the data in the data enrichment model ». The discussion goes on to blame the poor SC condition for yielding *radically pessimistic* estimates for the sample complexity, for being a *stringent condition* etc. Later the authors say that they introduce a « * considerably weaker* geometric condition ". However, the model that the authors consider is so specific that it is arguably ABSOLUTELY OBVIOUS that the SC condition will not be appropriate here, almost as much as a hammer is not the right tool to screw a screw. So this is perhaps subjective, but to me this discussion is at best useless. It would be much more appropriate to say that a counterpart of the SC condition is needed in this case and that this paper proposes such a definition. The choice of the words « weaker than the state of the art », « fails to », « radically pessimistic », « considerably weaker " all appear to me as contributing to an annoying rhetorical style that aims at emphasizing how nice the proposed results are by arguing that they are superior to the existing literature. Clearly there is no point in doing this here, and bashing an extremely poorly chosen baseline is not very useful scientifically…

NOTATIONS, FORMULATIONS etc

Defined at the bottom of page 2

I assume that omega(.) denotes the Gaussian width of an argument. But it seems that omega is not defined or only implicitly via its use from page 5 onwards. Note also that omega is also used to refer to the noise in the linear model, which is perhaps not a very good choice for that reason.

Begining of Section 3.

The text refers to an RE condition without saying that this means restricted eigenvalue (which might require a reference as well) and without being clear as to which inequality this expression refers to.

The text says "Figure 3d considers a very high-dimensional setting where p=1000" -> p=1000 cannot be considered a *very* high-dimensional setting either in absolute or relative terms: the ratio of the total number of parameters on the total number of observations is less than 0 and with scaling in s log p this is more than comfortable. No we are not at all in a *very* high-dimensional regime.

X_0 does not seem to be defined.

I am personally really not convinced at all by the expression « data enrichment ». There are regularly new buzzwords that are introduced in the ML community for concepts and ideas that already have a name and this makes for a serious lack of traceability of the ideas. As a result your paper can fail to be cited or other parts of the relevant literature might be lost at some point. As a community, I think we should take more into account the principle of Ockham, which was to say that one should not multiply the concepts beyond necessity… For me, this work is about multi-task learning, a very interesting field in which we still have a lot to understand…
To me data enrichment is an expression which is very very vague and can mean anything.
These question of terminology are somewhat subjective, but in mathematics it is important to have definitions that are common to the group of researchers that work on a problem. In the recent years too many research scientist in machine learning feel compelled to do invent new names and new acronyms to do branding of their work. Our field risks dramatically loosing in scientific quality.

034 through -> thorough

045: the model is not « based on n samples ». Maybe the learning is based on n samples
047: "the unique aspect of such high dimensional" -> there was nothing that was high-dimensional in the previous

sentence. If you write « such setup" then « such » should refer to the setup in the previous sentence.
Col 2 011: makes the estimation possible -> "makes the estimation possible in spite of the fact that n<p"
When introducing \beta and \beta_g it would be useful to say that both are vectors in RR^p. This is not obvious in the first place: \beta_g looks like it could be subvector of \beta.

2. Please provide an overall score for the submission.
Weak Reject: Borderline, tending to reject

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
This paper studies a relatively specific but interesting model, which is relevant for a number of applications, and obtains interesting technical results in terms of sample complexity for the considered model. Some of the experiments are quite well done with appropriate discussion and nice illustrative figures.
However a number of elements are not so compelling in the paper
The authors focus a lot on a claimed improvement over the SC condition, when clearly the SC condition does not apply in a reasonable way to the setting they consider (see detailed comments for a detailed discussion of this)
It would be much more relevant and interesting to have discussions and results to make comparisons with independent Lassos (it seems that the main results obtained is slightly worse than separate lassos) and more importantly with the model of Jalali et al. (2010) and to characterize both theoretically and in experiments in which regime the proposed model outperforms the model of Jalali et al. Doing this would make the results much more interesting. Currently the result are under the form of fairly abstract formulas involving Orlicz norms and many other abstract quantities which are not easy to relate to the other results appearing in the literature. Significant discussions and comparisons with other results are necessary.
All the results established are for designs that have i.i,d, Gaussian entries, a model which is used in compressed sensing but is really not applicable to real data…. ! (Very often conditions like RIP or RE are criticized because their are not realistic for real data but assuming that the data is pure unstructured Gaussian noise is even more extreme)
The experiments on real data really ought to have a better baseline than just a single global lasso model in which a single parameter is common to all tasks. There should be at least experiments with independent Lassos and it would seem more than reasonable to make a comparison with the model of Jalali et al.
5) Several important references are missing.

5. Please rate your confidence in the score assigned.
High: Reviewer has understood the main arguments in the paper, and has made high level checks of the proofs.

Reviewer #3

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.
The paper presents evaluates linear models for selecting/learning shared and mutually exclusive information content. It introduces an information sharing/coherence condition DERIC and a corresponding estimator. The resulting predictive model is shown to be more interpretable and efficient (robustness).

The conjecture that structural coherence is stronger than DERIC for superposition models is sensible. The simplicity of the eventual estimation procedure is attractive for scalability to high-dimensional datasets, possibly to output structures of nonlinear models like deep networks etc. There are several issues though with the current structure of the paper.
1. Overall, the paper looks like tightly squeezed version of a long/detailed manuscript, jamming several technical results into a 8-page paper. More specifically, decoding the technical results has been rather difficult. Remarks for each technical result is required. And moving some of them to supplement. The visualization in Figure 2 is not helpful. Empirically plotting what the two conditions are implying would be useful. What is an intuitive interpretation of the error cones (from the perspective of an for example, the cat example from Fig 1?).

2. Why is classical Lasso good baseline here? I would assume some state of the art multi-task learning model makes better baseline no? Is DICER better by statistical margin in Figure 5? Synthetic experiments can be moved to supplement.

2. Please provide an overall score for the submission.
Weak Reject: Borderline, tending to reject

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
As stated, my main concerns are with (a) the presentation of model/technical results etc. and (b) comparison to a baseline.

5. Please rate your confidence in the score assigned.
Medium: Reviewer has understood the main points in the paper, but skipped the proofs and technical details.

Reviewer #4

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.
This paper considers data fusion problem with a common parameter and different individual parameters in each individual model, generating a group of data from the total data. However, I think that most of the current results may be parallel to those of existing no-enrichment results in high dimensional problems, and additional technique challenges may be lacked. Moreover, the definition of the common parameter and all individual parameters is vague and cannot be identified well.

2. Please provide an overall score for the submission.
Weak Reject: Borderline, tending to reject

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
The current version of this paper is written well, and the theoretical part and numerical experiments are stated explicitly.
However, the model to be investigated may be not interesting enough, as their parameters in the model are not unique.
The proof procedures seem similar to those existing results in the classical high dimensional setting. The authors did not clearly explain mutual effects between
the estimation of \beta_0 and the estimation of \beta_g.

5. Please rate your confidence in the score assigned.
High: Reviewer has understood the main arguments in the paper, and has made high level checks of the proofs.