To Reviewer 1: Thanks for the review. **To Reviewer 2: (1) Relaxing Gaussianity**: Our analysis used Gaussian conditions in two places. (i) Estimation of the covariance matrix (Prop 1) requires only matrix concentration inequalities. We can replace the Gaussian assumption by moments and/or boundedness assumptions. Sometimes a log factor needs to be paid. (ii) Step 2&3 (in appendix) requires the sum of a sufficiently large subset of coordinates of $\mathbf{z}$ to concentrate around the expectation. We can replace Gaussians by sub-Gaussian random variables. Further relaxation is possible by deriving *ab initio* concentration bounds. **(2) Suggested algorithm**: The proposed formulation can be exactly solved (zero MSE) with multiple solutions of M. The solution under many circumstances (e.g., M has low rank) can have the noise term removed but may still be unsuitable for the application context. If the reviewer can elaborate on the idea, we can properly discuss potential barriers of implementing the idea.

To Reviewer 3: (1) Corollaries to Thm 1 (many parameters). Sec. 5 discussed simplified corollaries in narrative, including the rank assumption suggested (L315&317). We will give pointers to the Cor's after Thm 1. **(2) Power Law.** Thank you for pointing this out. We will fix it. **(3) Effective bound (related to X).** Lower bounds for solving the regression problem is in Sec. 4; A lower bound for the sub-procedure of covariance matrix estimation is indeed open. We suspect proving this lower bound requires some heavy anti-concentration inequalities for matrices because we need to show the gap paid by the Davis-Kahan theorem is inevitable. **(4) New data (height based on genome).** Thank you for the suggestion. Is UKBB (used in Lello et. al 18) the dataset in your mind? We will look into it (our method appears to be applicable). **(5) A. Ng's quote.** See the end of the page; we will contextualize this better.

To Reviewer 4: We comment on the reproducibility issue. **(1) Baseline implementation details (cross validation):** Cross validation was used for baselines. More specifically, in p. 36 of App, "We use three years of data for training, one year for validation, and one year for testing. The model is re-trained every test year..." See Fig. 1 for the sample code for parameter sweeping for RRR. **(2) Metrics & R^2.** R^2 is measured in basis points (10^{-4}). See Table 1's caption in the main text. The test R^2 (18bps) is within plausible range (E. Chan 17, Zhou & Jain 14). We believe using multiplicative metrics (3-fold improvement) is inappropriate because the baselines' R^2 are too close to 0 in equity datasets (suggesting they overfit). **(3) Intuition of performance gap.** The performance gap comes from the overfitting of baselines (recall that we proved near-optimality of our algorithm's prediction power). Take for example PCR, which can be considered a special case of ARRR with $k_1 = k_2$. For the equity dataset, the k_2 found through cross validation is around 1/3 of k_1. This means PCR has effectively 3x more parameters than ARRR. When the models fit against approximately 750 training points (750 trading days in training), PCR experiences a more pronounced overfitting problem. Note that the severity of overfitting depends on the signal-to-noise ratio (as predicted by Thm 1). Therefore, because the Twitter dataset has higher signal-to-noise ratio, the performance gap between PCR and ARRR is less pronounced. Other baselines can be analyzed in a similar manner. **(4) See Table 1 for a point-to-point response.** Blue text is non-opt statements (we do not disagree); red text is problematic ones (factual error or major misunderstanding); green text is clarified (mostly already in appendix).

Review	Response
The algorithm consists of applying PCA to X as in PCR and then doing SVD of cross-covariance between Y and PCs of X as in PLS. The entire method could be called PCR-PLS. Of course PLS is related to RRR but they are not equivalent. For example, the proposed algorithm does not seem to converge to the standard RRR for infinite data.	In high-dim, we consider n and d_1 grow together and $n < d_1$ (L88). n grows with d_1 and d_2 fixed is a **non-scope**. **L109: "We examine the setting where existing algorithms fail to deliver non-trivial MSE'.**
, or when the Step-1-PCA is modified to keep all PC components of X. Doing SVD of ZY where Z is standardized PCs of X is not equivalent to RRR. This is confusing.	This statement is incorrect. When all PCs of X is kept, the problem reduces to RRR. See Reinsel & Velu 98.
I suspect that the authors already had to deal with similar criticim (that their method is not very novel) because the Appendix contains a dedicated section about why this approach is novel. One part of claims that PCA is traditionally *not* seen as a regularization method. . . .it is standard textbook knowlegde that PCA in principal component regression (PCR) can very effectively prevent overfitting and in fact is closely related to ridge regression. See Hastie et al., The Elements of Statistical Learning. It is strange to claim that PCA regularization is a novel idea.	See below for clarification. We remark that **Hastie et al's view** (Sec. 18.6 PCR in high dim) belongs to "view 1" (analysis is possible only under factor models). **References cited therein (e.g., Bair et al 06) in fact re-confirm the accuracy of our discussion**.
What exactly is R^2_{out}? If it actually is R^2 then it cannot be above 1. Unclear what R^2=18 means.	Explained above (see also Table 1's caption in the main text.)
Many comparison methods have hyper-parameters . . . How were they set? It only makes sense to use cross-validation to set the optimal parameters. Was CV used here? It is not mentioned. I cannot believe that the method suggested here outperforms PCR, RR, RRR, etc. by more than 3-fold (as in Table 1). This is an enormous difference that makes one suspect that other methods were not applied correctly.	See implementation details and intuition of the performance gap.
I am not convinced that this paper provides a contribution of NeurIPS level. The suggested estimator is extremely simple; one could even say "naive".	That an estimator is naive does not mean the analysis is trivial. Proving that simple estimators are theoretically sound and work in practice is an important research direction in CS and Statistics.

Table 1: Specific questions and responses.

A. Ng's view. One view (view 2; Ng) claims PCA should not be used for preventing overfitting whereas the other view (view 1) claims PCA can tackle overfitting problems but the analysis is possible only under factor models. We will change the wording on view 1 to avoid confusion. Our view is new: it can tackle overfitting problems (different from v2) and can do so even without factor models (strengthened v1) but this needs to be analyzed in a specific high-dim setting (not mentioned in v1 or v2).

```
function  rrr_rank = find_optima_rrr(Xtrain, ytrain, Xtest, ytest)
    rgbeta = 1:1:50;
    errorMatrix = zeros(1,length(rgbeta));
    for j = 1:length(rgbeta)
            keep_beta = rgbeta(j);
            [~, predY_test, ~] = f_rrr(Xtrain, ytrain, Xtest, keep_beta);
            err = norm(ytest(:)- predY_test(:));
            errorMatrix(j) = err;
    end
    min_ie = find(errorMatrix == min(errorMatrix(:)));
    rrr_rank = rgbeta(min_ie(1));
end
```

Figure 1: Sample code for parameter sweeping for RRR.