Discussion on: "A scale-free approach for false discovery rate control in generalized linear models" by Dai, Lin, Zing, Liu

Gerda Claeskens[1], **Maarten Jansen**[2] and **Jing Zhou**[3]
[1] ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
gerda.claeskens@kuleuven.be
[2] Departments of Mathematics and Computer Science, Université libre de Bruxelles, Boulevard du Triomphe Campus Plaine, CP213, B-1050 Brussels, Belgium
maarten.jansen@ulb.be
[3] School of Mathematics, University of East Anglia, UEA, Norwich Research Park, Norwich, NR4 7TJ, UK.
J.Zhou6@uea.ac.uk

We wish to congratulate the authors for their contribution to the selection of variables in generalized linear models with controlled false discovery rate, in both moderate and high-dimensional settings.

1 Maximum likelihood estimators in moderately high dimensions

One of the noteworthy results in this paper is the asymptotic characterization of maximum likelihood parameter estimators in generalized linear models where the number of parameters p grows with the sample size n in such a way that the ratio p/n converges to a limit in the open interval $(0, 1)$. The mean and variance adjustment, as compared to the classical low-dimensional setting with a fixed p extends the results that were earlier obtained only for logistic regression models. While this was not the main topic of this paper, estimation of the mean and variance adjustment parameters would be quite interesting. That would pave the way to develop other ways of variable selection in such moderate high dimensional models.

In the numerical studies for the moderate dimensional settings, Figures 2 and 3 in Dai et al. (2023) considered logistic regression models, and Figure 4 is about a negative binomial model with dispersion parameter 2. The Benjamini-Hochberg procedure (BH) and the adjusted BH procedure based on the bias and variance adjustment scaling factors (Sur and Candès, 2019) attracted our attention. The BH procedure has surprisingly high power among the competitors, i.e., the proposed Gaussian mirror (GM), data splitting (DS), multiple data splitting (MDS), and the knockoff filter (KN). The classic BH in Figure 2 ($n = 500, p = 60$) also properly achieves the FDR control when the asymptotic normality has not been severely affected. In Figure 3 ($n = 3000, p = 500$), the BH procedure lost FDR control but gained it back after the bias and variance adjustments for the logistic regression model. Figure 4 is similar to Figure 3, but the adjusted BH procedure is yet to be derived. From the numerical results in Section 5.1, it appears that the p-value-based procedure should be favored if applicable. The p-values are based on the asymptotic distribution, in this case, the asymptotic normality of the regression parameter vector. The asymptotic normality is the golden rule for frequentist statistical inference. Once the asymptotic normality is derived, a simple yet elegant and powerful FDR control can be achieved with low computational complexity.

Dai et al. (2023) have established the asymptotic normality of the MLE for distributions in the exponential family in Proposition 3.1, which agrees with Sur et al. (2019); Sur and Candès (2019); Zhao et al. (2022). This result indicates that the asymptotic distribution of the MLE should be adjusted by bias and variance scaling factors as compared to the values in the low dimensional case. An open question here is how to estimate the scaling factors, which is challenging but worth looking into.

2 Highly sparse models in high dimensions

In the case of $p \gg n$, the authors considered the desparsified Lasso (van de Geer et al., 2014) for generalized linear models in Section 4 to construct mirror statistics. The asymptotic normality of the desparsified Lasso requires a sparsity condition $s = o(\sqrt{n}/\log(p))$, where s is the number of non-null components of the regression coefficient vector of length p. Further, the sparsity condition is always mandatory for all regularization-based estimation approaches to guarantee certain consistency properties of the regularized estimators. This could explain why the knockoff filter performs well only in sparse settings since it relies on a good estimator. Surprisingly, the proposed GM, DS, MDS have robust and fair performance in the medium sparse settings, which is a great contribution where variable selection by regularization faces challenges. One reason could be that the proposed methods rely on the symmetry of the mirror statistics for the null components instead of directly relying on the consistency of the regularized estimators.

One question we have is: how do the methods behave in high-sparsity settings where estimator-based FDR control procedures should have reasonable performance? A concern is that when only a handful of variables are relevant to the response variable, the FDR could possibly be arbitrarily large. For example, when only five variables are relevant, making one mistake out of 5 in the selection set could already make the false discovery proportion equal to 0.2. In the paper, the authors focused on large n, p cases ($p = 2000, n = 800$), which is helpful in validating FDR control asymptotically. We were more curious about the small n, p cases since it obviously favors the BH procedure in the moderate dimensional settings with $p/n \to \kappa \in (0, 1)$, see Dai et al. (2023, Figure 2).

Following our observation in the moderate dimensional settings, we proposed to simply apply a BH procedure to the p-values based on the desparsified Lasso (van de Geer et al., 2014). The simulation setting we took for logistic regression models is similar to the left panel of Figure 5 in Dai et al. (2023), and we focus on even sparser settings where $s = 5, 10, 15, 20$. For comparison, we also consider $s = 50$. We proportionally reduce n, p and set $n = 200, p = 500$ such that $p/n = 2.5$ agrees with the setup in Dai et al. (2023, Figure 5). We consider the nominal FDR level $q = 0.1, 0.2$. As was done in Dai et al. (2023), the averaged FDR and power are calculated by averaging over 50 randomly generated datasets for all five methods, i.e., BH procedure (Dai et al., 2023) (BHq), data splitting (DS), multiple data splitting (MDS), the knockoff filter (KN), and a naive BH simply apply the BH procedure to the p-values from the desparsified Lasso in van de Geer et al. (2014) (BHdesparsify). The simulation results are presented in Figure 1. As expected, the BHq and BHdesparsify have reasonable FDR control and promising power in high sparse settings where $s = 5, 10, 15$.

Surprisingly, the biggest difference between BH and DS is only in the very sparse setting $s = 5$ and $q = 0.1$. The DS has good performance with power and FDR control for the other settings. We wonder what could be an explanation for the observation that for smaller n, p settings, DS and MDS no longer outperform BHq. In which ways could DS and MDS be further improved for the moderate n, p settings with high sparsity?

3 A three-step approach in high dimensions

For the case of regularized estimation for high dimensional data the procedure of Dai et al. (2023) proceeds in three steps. *Step 1*: In high dimensional settings the ℓ_1 regularization (lasso estimation) takes care of a selection of the variables. This means that some of the coefficients will actually be set to zero, depending on the value of the regularization constant λ. *Step 2*: Since the point mass at zero for the asymptotic distribution of the variables that are not selected may cause difficulties for inference, the desparsification or debiasing methods add a certain random quantity to the vector of estimated coefficients such that all coefficients become non-zero. While this results in a multivariate normal random vector (all the zeros became nonzero), it has thus also undone the selection aspect of the regularized estimation method since there are no longer

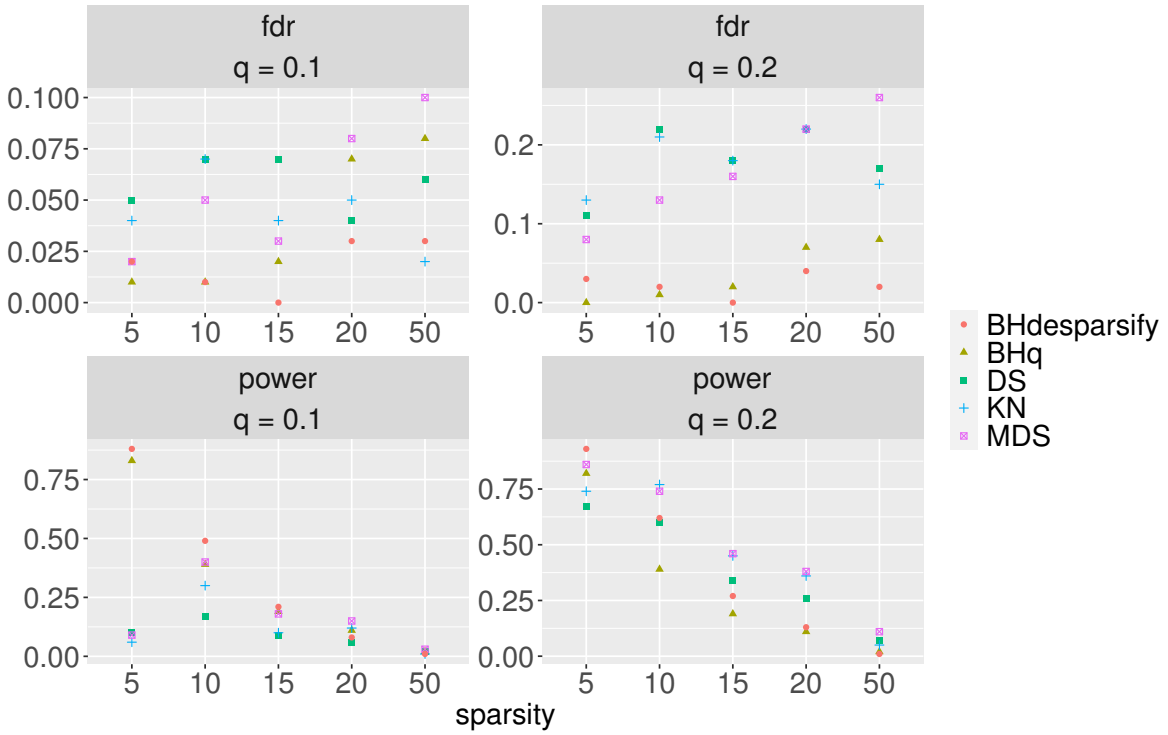

Figure 1: The averaged FDR and power are calculated by averaging over 50 randomly generated datasets for all five methods, i.e., BH procedure (Dai et al., 2023) (BHq), data splitting (DS), multiple data splitting (MDS), the knockoff filter (KN), and a naive BH simply apply the BH procedure to the p-values from the desparsified Lasso in van de Geer et al. (2014) (BHdesparsify). The first and second columns present results for the nominal level $q = 0.1$ and $q = 0.2$, respectively.

components set to zero. *Step 3*: This is another stage of selection, this time by means of the (multiple) data splitting and mirror statistics. One is left wondering whether this back-and-forth approach of selection - unselection and again selection might be open for improvement.

We have the following questions regarding this three-step approach.

3.1 How does the choice of λ in step 1 of the procedure influence the final result?

In the current implementation of Dai et al. (2023) 10-fold cross-validation is used for all of the nodewise regressions used in the debiasing step, as well as for the lasso-estimation problems in each of the data splitting sets. To be precise, the authors' code computes the cross-validation value for the full dataset and divides this value by $\sqrt{2}$ for each of the split datasets. In order to investigate the effect of the initial choice of the λ in Step 1 on the final FDR, we ran a small simulation study. The number of replicates is set to 500. The considered setting takes: $n = 200$, $p = 500$, $q = 0.2$ or $q = 0.1$ and the number of nonzero regression coefficients $s = 50$ in a logistic regression model with normal covariates, as in our previous setting in Section 2, based on the simulation by Dai et al. (2023). While we kept the cross-validation value for the nodewise regression problems in the debiasing step 2, we investigated the choice of λ in step 1 to be ten times smaller and 10 times bigger than the authors' choice by cross-validation.

Table 1 gives a numerical summary of (i) the average number of FDR = number of false positives/max(1,number of selected variables) over these simulation replicates, (ii) the average number of selected variables/number of true nonzero variables (power) and (iii) the percentage of times that the FDR was found to be at least 50%. Histograms of the simulated FDR and

		$\lambda_{\text{cv}}/10$	λ_{cv}	$10\lambda_{\text{cv}}$
$q = 0.2$	average FDR	0.15	0.16	0.27
	$\#(\text{FDR}\geq 50\%)/N_{\text{rep}}$	0.13	0.11	0.31
	average power	0.05	0.06	0.10
$q = 0.1$	average FDR	0.03	0.03	0.08
	$\#(\text{FDR}\geq 50\%)/N_{\text{rep}}$	0.02	0.02	0.09
	average power	0.01	0.01	0.03
Step 1 Lasso	average FDR	0.74	0.62	0.03
	$\#(\text{FDR}\geq 50\%)/N_{\text{rep}}$	1.00	0.94	0.03
	average power	0.60	0.42	0.01

Table 1: Results from $N_{\text{rep}} = 500$ simulation replicates using three different values for the regularization parameter λ in step 1. Average number of false positives (FDR), percentage of simulation runs where FDR$\geq$ 50% and average number of selected positives/true positives (power) when $q = 0.2$ and $q = 01$ for the single data split method (DS). The last part of the table presents these summary values for the regularized estimator from Step 1.

power values are shown in Figure 2, which depict a clear bimodal shape for the FDR values from the 500 simulation replicates and indeed shows a dependence on the chosen value of the regularization parameter. Table 1 further indicates that some of the FDR values might be quite large, e.g., for a nominal $q = 0.2$ and using cross-validation in step 1 to set λ, 11% of the FDR values was found to be larger or equal to 0.5. The observed values for the power, see Figure 2, might indicate that a strong emphasis on the FDR might result in a low power. It would be interesting to better understand this behavior of FDR and its relation to power.

3.2 Could the three-step procedure be simplified in a single-step procedure?

We wonder whether it would be possible to achieve the same goal by an adequate choice of λ that keeps the FDR under control, as opposed to the cross-validation selection of λ. As a numerical experiment, Table 1 also shows the simulated FDR and power values for the full dataset with cross-validation as a method to set λ, and choices that are 10 times smaller or 10 times larger. It appears that a larger value of λ as compared to the cross-validated one would be needed in order to control FDR.

The lasso regularization defines a convex optimization problem leading to an estimation in which many covariates end up being zero. The implicit covariate selection is interesting in the sense that it mimics the computationally infeasible search for the subset of covariates of given cardinality that minimizes the squared residual sum after orthogonal projections, see, for instance, results in Donoho (2006). Next to a covariate selection, however, Lasso also includes a shrinkage on the selected covariates, thus tempering the effect of false positive selections. With two sources of bias, namely false negatives and shrinkage bias, and reduced variance in false positives due to the shrinkage, the trade-off between bias and shrinkage in fine-tuning the smoothing parameter λ in the lasso regularization shifts towards larger models that are less sensitive to bias. In other words, being tolerant towards false positives, fine-tuning lasso for a bias-variance trade-off tends to largely overestimate the size of the least false model. This holds, in particular, if the criterion for the assessment of the selected model is based on the prediction error or the expected log-likelihood (i.e., Kullback-Leibler divergence). This is the case when using information criteria such as AIC and Mallows's C_p. Cross-validation and generalized cross-validation are well known to be linked to Mallows's C_p, see, for instance, Jansen (2015). The information criteria can be adjusted to assess debiased lasso (Jansen, 2014), thus balancing between bias and variance, or closeness and complexity *after* undoing the shrinkage.

A three-steps procedure in which the prediction error or Kullback-Leibler divergence is alternated with a more conservative approach, such as FDR, may also lead to a balanced approach

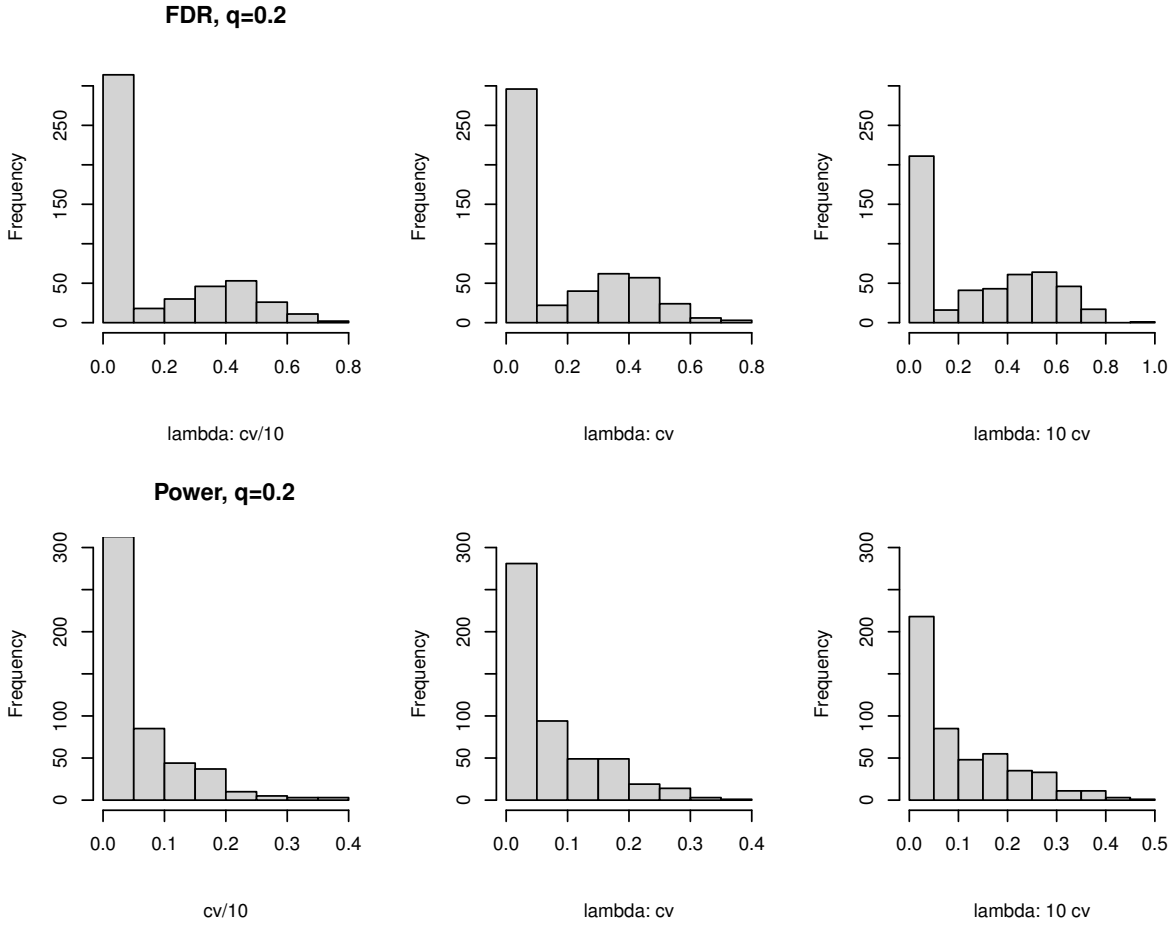

Figure 2: Results from $N_{\text{rep}} = 500$ simulation replicates using three different values for the regularization parameter λ in step 1. Average number of false positives (FDR), percentage of simulation runs where FDR$\geq$ 50% and average number of selected positives/true positives (power) when $q = 0.2$ and $q = 01$ for the single data split method (DS).

serving different objectives in an application dependent order of importance. However, if the final focus lies on controlling the false discovery rate, it would be useful to examine if the lasso fine-tuned for FDR control could be beneficial in terms of computation speed, the reported false discoveries and the interpretability of the obtained results. An FDR-controlled lasso could be integrated into the subsequent debiasing, data splitting and mirror statistics steps, leading to a coherent approach.

4 Post-selection inference

Like any variable selection method, also the selected models resulting from FDR control with the method from Dai et al. (2023) come with uncertainties. While in the theoretical work of Dai et al. (2023) for the high-dimensional case the regularization parameters λ that appear at several instances were kept fixed, in their numerical work the values are obtained in a data-driven way. Also this aspect introduces additional variability.

When inference is to be performed on the parameters in the model obtained after selection, one should use proper inference methods that account for the selection of the variables and of the regularization parameters when present. This aspect is not included in the asymptotic distribution result of Proposition 3.1 (Dai et al., 2023), where the model is assumed to be correctly specified. It would be interesting to develop post-selection inference results, for example, by conditioning on the selection event in the style of Lee et al. (2016), for inference in the models obtained by this FDR controlled selection, both in moderate and high-dimensions.

Acknowledgements

G.Claeskens acknowledges the support of the Research Foundation Flanders and KU Leuven Research Fund C1-project C16/20/002.

References

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023). A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, (just-accepted):1–31.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Mathematics*, 59:797–829.

Jansen, M. (2014). Information criteria for variable selection under sparsity. *Biometrika*, 101(1):37–55.

Jansen, M. (2015). Generalized cross validation in variable selection with and without shrinkage. *Journal of Statistical Planning and Inference*, 159:90–104.

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1):487–558.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Zhao, Q., Sur, P., and Candes, E. J. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861.