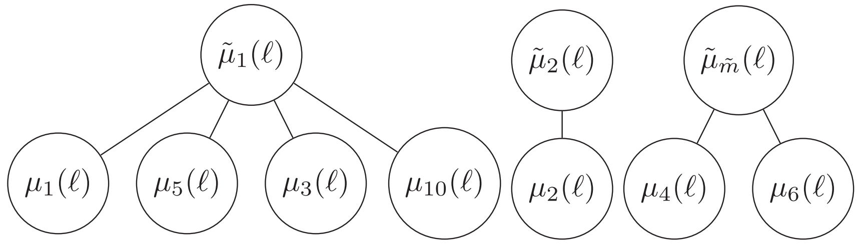

Fig. 11. Schematic diagram of constraint (6) with $\sigma(3) = 1$

Moreover, $T_{\mathrm{OR}}^{(j)}(t_j, r_j, s_j)$ enjoys the properties in theorem 3 of the paper. Detailed proofs are available from `https://hugogogo.github.io/paper/cars_discussion_supplement.pdf`. If there is not a one-to-one mapping between $\sigma(j)$ and j, then $T_{\mathrm{OR}}^{(j)}(t_j, r_j, s_j)$ must be estimated carefully.

The **authors** replied later, in writing, as follows.

We thank the discussants for their insightful comments and excellent contributions. It is our great delight to meet some discussants in London, and we are pleased to participate in further discussions in writing. The discussions are wide ranging. For brevity, we focus only on some key topics.

Key message: data reduction, information loss and optimality
Data reduction via constructing linear contrasts has long been used as an essential tool for statistical analyses. Examining the process at a high level, the conventional practice involves first dividing raw data into 'relevant' and 'irrelevant' parts (or data carving, per Professor Ramdas), and then developing inference procedures based solely on the summary of the relevant part. This practice is widespread in statistical analysis. A major surprise is that such standard practices in data processing could lead to significant information loss in large-scale inference. Our work marks a clear departure from the existing work where auxiliary information is gleaned from external data. We propose new strategies to extract structural information *within the same sample* by using auxiliary covariates. We thank Professor Fan for the comments on our contributions to the optimality theory in false discovery rate FDR control, which has been lacking in the literature. This is an important direction in large-scale inference, considering that optimality has been the goal in the development of many fundamental results in statistics including Fisher's theory on the asymptotic efficiency of maximum likelihood estimation and the Neyman–Pearson lemma on the optimality of the likelihood ratio test.

*Structural information in high dimensional inferenc*e
We appreciate the illuminating comments from Roquain and Nichols on the role of auxiliary data in amplifying the signals. From a decision theoretic view, classical ideas such as Robbins's compound decision theory and the James–Stein shrinkage estimator show that the joint structure of primary statistics can be exploited to construct more efficient estimation or testing rules. A key message conveyed through this work is that *extra* valuable structural information can be extracted from the seemingly irrelevant part of the data. This point is particularly crucial in high dimensional settings. When the number of parameters is small, the information loss is inconsequential (since the joint structure cannot be estimated well). However, in the case with thousands of parameters structural information can be recovered with good precision from auxiliary statistics, which can play a key role in improving the power.

The sufficiency principle and broad applicability of covariate-assisted ranking and screening
We concur with Fan and Habiger in their insightful comments on *sufficient statistics*: a fundamental principle that seems to have been largely ignored in the FDR-literature. The comments also shed new lights on the applicability of covariate-assisted ranking and screening (CARS) beyond the case that requires doubly sparse means. The general idea in CARS works for a broad class of bivariate models (see Section 4.1 and our remark 7) and the doubly sparse assumption should be viewed as a special setting to explain intuitively why CARS works. Moreover, violations of the sufficiency principle are common in data processing and CARS can benefit from a non-sparse auxiliary sequence as long as the covariate encodes useful structural information. These points have been nicely corroborated by Professor Habiger's several interesting papers on heterogeneous discrete data. Professor Habiger's inspiring discussion also points the way forward for developing effective data combination strategies across qualitative and quantitative variables.

Using covariate-assisted ranking and screening in other high dimensional problems

CARS provides a generic tool for inferring sparsity structure by integrating evidence from multiple sources. The interesting discussions by Bogomolov, Heller and Yekutieli, Yu, Bien and Witten, Yang and Cheng, Li and Wong, and Banerjee and Mukherjee, among others, show that CARS has considerable potential for providing better solutions to a wide range of high dimensional problems including large-scale analysis-of-variance tests, high dimensional replicability analysis, sparse linear discriminant analysis, multiview analysis, hierarchical inference and sparse compound estimation. It is encouraging to see some preliminary successes reported by the discussants. We feel that revisiting the fundamental sufficiency principle in large-scale inference and carefully investigating possible information loss in data reduction would be an important and fruitful direction for future research. We appreciate the creative ideas and stimulating comments from the discussants on applying CARS to various high dimensional inference problems. We very much look forward to further explorations along these lines.

The dependent case and the 'grouping, adjusting and pooling' procedure

Fan, Goeman and Solari expressed legitimate concerns on the independence assumption. Although the robustness of CARS under dependence has been investigated numerically in the on-line appendix B.5, we take this opportunity to describe briefly our recent work aiming to address the important dependence issue. Xia *et al*. (2018) developed a general information pooling framework that involves grouping, adjusting and pooling (GAP) to leverage the structural information from an auxiliary sequence. GAP is built on the Benjamini–Hochberg (BH) procedure and utilizes weighted p-values to capture the heterogeneity among hypotheses. We generalize the weighted multiple-testing theory in Genovese *et al*. (2006) to show that GAP controls FDR under a range of dependence structures, including weakly dependent tests arising from high dimensional linear regression and Gaussian graphical models. However, the optimal choice of weights is still an open issue that deserves more research; inspiring discussions can be found in the comments by Dobriban, Ramdas and Habiger on use of weighted p-values and interactive use of masked p-values.

Asymptotic false discovery rate control and variations of the Benjamini–Hochberg procedure

CARS and Lfdr-methods offer asymptotic FDR-control and work better for large-scale testing problems where the density functions can be well estimated. By contrast, the BH procedure offers guaranteed FDR-control under a range of dependence structures. For smaller-scale problems with a few dozen or several hundred tests as considered by Goeman and Solari, we recommend GAP and other variations of the BH procedure (see the discussions by Dobriban, Ramdas and Habiger) to incorporate useful side information. It would be of great interest to investigate the performance of Bayesian CARS (see Professor Roquain's comments) to increase the stability in small sample settings.

The sparse case and updated covariate-assisted ranking and screening package

We thank Roquain and Nichols for noticing the issues of our CARS package under the very sparse case. We have uploaded to the Comprehensive R Archive Network the updated package that includes the 'sparse option' described in Section 5.1 and a new section on the vignette illustrating that CARS, using the sparse option, controls FDR when $m = 10000$ and $k = 10$: a setting considered by Professor Nichols. The key idea for the sparsity adjustment is to use the known densities to stabilize the bivariate density estimate in regions with few observations. Through communication with Professor Mark van de Wiel, we recognize that, for methods based on CARS and Lfdr, the instability of the non-parametric density estimator (in the denominator) seems to be a common issue. In the sparse regime, Professor Roquain's proposal of employing Cauchy slab priors is a promising direction with the potential of having the best of both worlds: the method avoids non-parametric modelling of a bivariate density, while the choice of priors has great promise of leading to good frequentist properties.

On choosing the auxiliary sequence

We briefly address the interesting question from Goeman and Solari whether the total variance could be a good competitor as an auxiliary variable. First, it can be shown that, with known and homoscedastic variances, the pair (T_1, T_2) is a sufficient statistic (per Professor Fan); hence T_2 is optimal in the sense that it has no information loss. Although the sufficiency principle may be satisfied by other pairs, our choice of T_2 not only is intuitively appealing but also simplifies the development of both methodology and theory; see the discussion in Section 2.1. Second, an important consideration in choosing the auxiliary variable is to avoid selection bias. As noted in a post by Professor Ryan Tibshirani on the 'Normal deviate' blog, screening based on between-group variance leads to severe selection bias. The total variance is not promising either, at least under the CARS framework, because under heteroscedasticity it is correlated with the primary statistic and cannot capture the sparsity structure effectively. Moreover, the total variances are not suitable as useful structures to inform BH algorithms. The p-value null distribution is likely to be distorted when screening, grouping or weighting is carried out via total variance.

Open issues and concluding remarks
Large-scale multiple testing is a fundamental building block in contemporary statistics and developing efficient procedures that control the FDR, a celebrated innovation in the past two decades, has been a prominent and impactful research area. Although the hypothesis testing framework is not omnipotent as pointed out by Professor Longford, we believe that some concerns may be possibly addressed by tailoring the general FDR-concept to the needs of specific applications; notable ideas include weighted FDR (Benjamini and Hochberg, 1997), directional FDR (Guo *et al.*, 2010) and the false important discovery rate (Sun and McLain, 2012). As pointed out by Medina and Stehlík, the null hypothesis should be carefully formulated, and existing methods should be properly modified for specific applications. Much more research is still needed in this area.

References in the discussion

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., Sklar, P., Roddey, J. C., Chen, C.-H., McEvoy, L., Desikan, R. S., Djurovic, S., Dale, A. M., Psychiatric Genomics Consortium and Bipolar Disorder and Schizophrenia Working Groups (2013) Improved detection of common variants associated with schizophrenia and polar disorder using pleiotropy-informed conditional false discovery rate. *PLOS Genet.*, **9**, article e1003455.

Banerjee, T., Mukherjee, G. and Sun, W. (2018) Adaptive sparse estimation with side information. *Preprint arXiv: 1811.11930*.

Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (1997) Multiple hypothesis testing with weights. *Scand. J. Statist.*, **24**, 407–418.

Bickel, P. J. and Levina, E. (2004) Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.

Bogomolov, M. and Heller, R. (2018) Assessing replicability of findings across two studies of multiple features. *Biometrika*, **105**, 505–516.

Bourgon, R., Gentleman, R. and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natn. Acad. Sci. USA*, **107**, 9546–9551.

Brown, L. D. (2008) In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.*, **2**, 113–152.

Cai, T. and Sun, W. (2017) Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *J. R. Statist. Soc.* B, **79**, 197–223.

Castillo, I. and Mismer, R. (2018) Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Statist.*, **12**, 3953–4001.

Castillo, I. and Roquain, E. (2018) On spike and slab empirical Bayes multiple testing. *Preprint arXiv:1808.09748*.

Dobriban, E. (2017) Weighted mining of massive collections of p-values by convex optimization. *Informn Inf.*, **7**, 251–275.

Dobriban, E., Fortney, K., Kim, S. K. and Owen, A. B. (2015) Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika*, **102**, 753–766.

Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

Efron, B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Statist. Ass.*, **99**, 96–104.

Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2018) FarmTest: factor-adjusted robust multiple testing with false discovery control. *J. Am. Statist. Ass.*, to be published.

Fithian, W., Sun, D. and Taylor, J. (2014) Optimal inference after model selection. *Preprint arXiv:1410.2597*. University of California at Berkeley, Berkeley.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B. and Kim, S. K. (2015) Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLOS Genet.*, **11**, no. 12, article e1005728.

Friedman, J. (2004) On multivariate goodness-of-fit and two-sample testing. *Report SLAC-PUB-10325*. Stanford Linear Accelerator Center, Menlo Park.

Genovese, C. R., Roeder, K. and Wasserman, L. (2006) False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.
Guo, W., Sarkar, S. K. and Peddada, S. D. (2010) Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, **66**, 485–492.
Habiger, J. D. (2017) Adaptive false discovery rate control for heterogeneous data. *Statist. Sin.*, **27**, 1731–1756.
Habiger, J., Watts, D. and Anderson, M. (2017) Multiple testing with heterogeneous multinomial distributions. *Biometrics*, **73**, 562–570.
Heller, R. and Yekutieli, D. (2014) Replicability analysis for genome-wide association studies. *Ann. Appl. Statist.*, **8**, 481–498.
Ignatiadis, N., Klaus, B., Zaugg, J. B. and Huber, W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Meth.*, **13**, 577.
Johnstone, I. M. and Silverman, B. W. (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, **32**, 1594–1649.
Johnstone, I. M. and Silverman, B. W. (2005) Ebayes Thresh: R programs for empirical Bayes thresholding. *J. Statist. Softwr.*, **12**, no. 8.
Katsevich, E. and Ramdas, A. (2018) Towards 'simultaneous selective inference': post-hoc bounds on the false discovery proportion. *Preprint arXiv:1803.06790*. Stanford University, Stanford.
Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. and Baker, C. I. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neursci.*, **12**, 535–540.
Kropf, S. and Läuter, J. (2002) Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometr. J.*, **44**, 789–800.
Lei, L. and Fithian, W. (2018) AdaPT: an interactive procedure for multiple testing with side information. *J. R. Statist. Soc.* B, **80**, 649–679.
Lei, L., Ramdas, A. and Fithian, W. (2017) STAR: a general interactive framework for FDR control under structural constraints. *Preprint arXiv:1710.02776*. University of California at Berkeley, Berkeley.
Li, A. and Barber, R. F. (2019) Multiple testing with the structure adaptive Benjamini–Hochberg algorithm. *J. R. Statist. Soc.* B, **81**, 45–74.
Longford, N. T. (2014) A decision-theoretical alternative to testing many hypotheses. *Biostatistics*, **15**, 154–169.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Michaelson, J. J., Loguercio, S. and Beyer, A. (2009) Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, **48**, 265–276.
Pecanka, J., Jonker, M. A., IPDGC, Bochdanovits, Z. and Van Der Vaart, A. W. (2017) A powerful and efficient two-stage method for detecting gene-to-gene interactions in GWAS. *Biostatistics*, **18**, 477–494.
Peña, E., Habiger, J. and Wu, W. (2011) Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.*, **39**, 556–583.
Ramdas, A., Singh, A. and Wasserman, L. (2016) Classification accuracy as a proxy for two sample testing. *Preprint arXiv:1602.02210*.
Roeder, K. and Wasserman, L. (2009) Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.*, **24**, 398–413.
Roquain, E. and van de Wiel, M. A. (2009) Optimal weighting for false discovery rate control. *Electron. J. Statist.*, **3**, 678–711.
Rosenblatt, J. D., Benjamini, Y., Gilron, R., Mukamel, R. and Goeman, J. J. (2016) Better-than-chance classification for signal detection. *Preprint arXiv:1608.08873*.
Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.
Sun, W. and Cai, T. T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Ass.*, **102**, 901–912.
Sun, W. and McLain, A. C. (2012) Multiple testing of composite null hypotheses in heteroscedastic models. *J. Am. Statist. Ass.*, **107**, 673–687.
Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E. and Ugurbil, K. (2013) The WU-Minn Human Connectome Project: an overview. *NeuroImage*, **80**, 62–79.
Westfall, P. H., Kropf, S. and Finos, L. (2004) *Weighted FWE-controlling Methods in High-dimensional Situations*, pp. 143–154. Beachwood: Institute of Mathematical Statistics.
Xia, Y., Cai, T. T. and Sun, W. (2018) GAP: a general framework for information pooling in two-sample sparse inference. *Technical Report*. Fudan University, Shanghai. (Available from `http://www.bcf.usc.edu/~wenguans/Papers/GAP.pdf`.)
Xie, X., Kou, S. and Brown, L. D. (2012) Sure estimates for a heteroscedastic hierarchical model. *J. Am. Statist. Ass.*, **107**, 1465–1479.
Yang, Q. and Cheng, G. (2018) Quadratic discriminant analysis under moderate dimension. *Preprint arXiv:1808.10065*. Purdue University, West Lafayette.
Zhou, W.-X., Bose, K., Fan, J. and Liu, H. (2018) A new perspective on robust M-estimation: finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.*, **46**, 1904–1931.