Two-sample testing of high-dimensional linear regression coefficients

Tengyao Wang

University College London

ETH young data science researcher seminar

12 Feb 2021

Collaborators

Fengnan Gao
Fudan University

Setting up the scene

- Two samples $(X_1, Y_1) \in \mathbb{R}^{n_1 \times p} \times \mathbb{R}^{n_1}$ and $(X_2, Y_2) \in \mathbb{R}^{n_2 \times p} \times \mathbb{R}^{n_2}$, generated from the linear models:

$$\begin{cases} Y_1 = X_1\beta_1 + \epsilon_1 \\ Y_2 = X_2\beta_2 + \epsilon_2, \end{cases}$$

  where $\epsilon_1 \sim N_{n_1}(0, \sigma^2 I_{n_1})$ and $\epsilon_2 \sim N_{n_2}(0, \sigma^2 I_{n_2})$ are independent.

- Given (X_1, Y_1) and (X_2, Y_2), we want to test

$$H_0 : \beta_1 = \beta_2 \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_2.$$

Classical F-test

- Classically, the generalised likelihood ratio test can be used to distinguish H_0 and H_1.

$$\mathrm{RSS}_0 := \|Y_1 - X_1\hat{\beta}\|_2^2 + \|Y_2 - X_2\hat{\beta}\|_2^2,$$
$$\mathrm{RSS}_1 := \|Y_1 - X_1\hat{\beta}_1\|_2^2 + \|Y_2 - X_2\hat{\beta}_2\|_2^2,$$

$$F := \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/p}{\mathrm{RSS}_1/(n_1+n_2-p)} \sim F_{p,n_1+n_2-p} \quad \text{under } H_0.$$

- Works well when $n_1, n_2 \gg p$.

High-dimensional settings

- Existing works mostly assume sparsity of both β_1 and β_2 (e.g. Xia, Cai and Cai (2018))
- But parameter of interest is really

$$\theta := \frac{\beta_1 - \beta_2}{2}$$

  and $\gamma := (\beta_1 + \beta_2)/2$ is a possibly dense nuisance parameter.
- We would like to impose sparsity on θ alone.

Problem setup in high-dimensions

- More precisely, define

$$\Theta_{p,k}(\rho) := \{\theta \in \mathbb{R}^p : \|\theta\|_2 \geq \rho, \|\theta\|_0 \leq k\},$$

  we would like to test

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \in \Theta_{p,k}(\rho).$$

- **Question:** what is the smallest ρ such that we can 'test apart' H_0 and H_1.

- More precisely, define

$$\Theta_{p,k}(\rho) := \{\theta \in \mathbb{R}^p : \|\theta\|_2 \geq \rho, \|\theta\|_0 \leq k\},$$

  we would like to test

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \in \Theta_{p,k}(\rho).$$

- **Question:** what is the smallest ρ such that we can 'test apart' H_0 and H_1.

- **Remark.** If we allow arbitrary γ, it is necessary to assume that $n_1 + n_2 > p$. Otherwise, for any $\theta \in \mathbb{R}^p$, the system of equations

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta_1 = \begin{pmatrix} Y_1 \\ Y_2 - 2X_2\theta \end{pmatrix}$$

  have at least 1 solution β_1, which together with $\beta_2 := \beta_1 - 2\theta$, gives perfect fit to $Y_1 = X_1\beta_1$ and $Y_2 = X_2\beta_2$.

Related literature

- Two sample testing of high-dimensional means against a sparse alternative (Cai, Liu and Xia, 2014; Chen, Li and Zhong, 2019)
- One sample testing of global null of a high-dimensional regression coefficient against sparse alternative (Ingster, Tsybakov and Verzelen, 2010 and Arias-Castro, Candés and Plan, 2011)
- Testing equality of two sparse regression coefficients (Städler and Mukherjee, 2012; Xia, Cai and Cai, 2018; Xia, Cai and Sun, 2020)
- Testing equality of two dense high-dimensional regression coefficients against a sparse alternative (Charbonnier, Verzelen and Villers, 2015; Zhu and Bradic, 2016)

Our method: complementary sketching

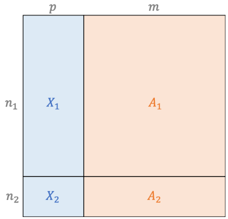
A complimentary sketching

Complementary sketching

- **Notation:** $n := n_1 + n_2, m := n - p$.
- **Procedure:** Given data X_1, X_2, Y_1, Y_2,
  1. Construct $A_1 \in \mathbb{R}^{n_1 \times m}$ and $A_2 \in \mathbb{R}^{n_2 \times m}$ such that $\binom{A_1}{A_2}$ has orthonormal columns orthogonal to the column space of $\binom{X_1}{X_2}$.
  2. Compute

$$W := \begin{pmatrix} A_1^\top & -A_2^\top \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in \mathbb{R}^{m \times p},$$

$$Z := \begin{pmatrix} A_1^\top & A_2^\top \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \in \mathbb{R}^{m}.$$



- Similar to orthogonal sketching, but sketches the covariate matrix and the response vector in opposite ways in the second block.

A key reduction

- Observe that

$$
\begin{aligned}
Z &= A_1^\top Y_1 + A_2^\top Y_2 = A_1^\top X_1\beta_1 + A_2^\top X_2\beta_2 + A_1\epsilon_1 + A_2\epsilon_2 \\
&= A_1^\top X_1\theta + \cancel{A_1^\top X_1\gamma} - A_2^\top X_2\theta + \cancel{A_2^\top X_2\gamma} + A_1\epsilon_1 + A_2\epsilon_2 \\
&= W\theta + \xi,
\end{aligned}
$$

  where $\xi \sim N_m(0, \sigma^2 I_m)$.

- We have reduced the two-sample testing problem to a one-sample problem of sample size m without the nuisance parameter.

Test construction from the sketch

- Let $\tilde{W}$ be W with columns normalised to have ℓ_2 norm 1.

$$
\begin{aligned}
&\text{— sparse case} \quad \psi^{\mathrm{sparse}}_{\lambda,\tau} := \mathbb{1}\{\|\mathbf{hard}(\tilde{W}^\top Z, \lambda)\|_2^2 \geq \tau\} \\
&\text{— dense case} \quad \psi^{\mathrm{dense}}_{\eta} := \mathbb{1}\{\|Z\|_2^2 \geq \eta\},
\end{aligned}
$$

  where $\mathbf{hard}(\cdot, \lambda)$ is the entrywise hard-thresholding function with threshold λ.

- Note that both tests $\psi^{\mathrm{sparse}}_{\lambda,\tau}$ and $\psi^{\mathrm{dense}}_{\eta}$ depends on A_1, A_2 only through the column span of $\binom{A_1}{A_2}$.

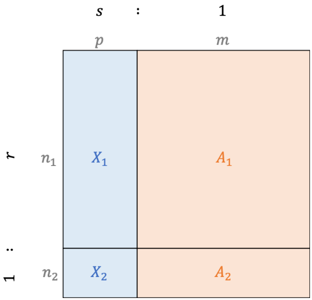
Theoretical results

Assumptions

- We assume that

  (C1) All entries of X_1, X_2 are independent $N(0,1)$ distributed.

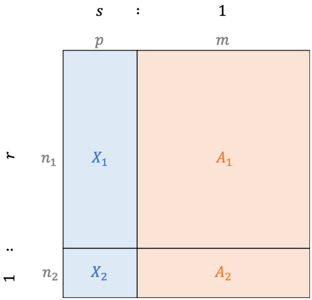
  (C2) Asymptotic regime: $n_1, n_2, p \to \infty$ such that $n_1/n_2 \to r$ and $p/m \to s$.



Assumptions

- We assume that

  (C1) All entries of X_1, X_2 are independent $N(0,1)$ distributed.
  (C2) Asymptotic regime: $n_1, n_2, p \to \infty$ such that $n_1/n_2 \to r$ and $p/m \to s$.



- Results are asymptotic in nature due to the use of limiting spectral distribution results from random matrix theory.
- The entrywise normality assumption can be replaced with a restricted isometry assumption on the transformed matrix W.

Upper and lower bounds, sparse case

- Let $P^X_{\beta_1,\beta_2}$ be the distribution of Y_1, Y_2 conditional on X_1, X_2 and for parameter β_1, β_2.
- **Minimax risk**

$$\mathcal{M}_X(k,\rho) := \inf_{\psi}\Big\{\sup_{\beta\in\mathbb{R}^p} P^X_{\beta,\beta}(\psi \neq 0) + \sup_{\substack{\beta_1,\beta_2\in\mathbb{R}^p\\(\beta_1-\beta_2)/2\in\Theta_{p,k}(\rho)}} P^X_{\beta_1,\beta_2}(\psi\neq 1)\Big\},$$

- Let $P^X_{\beta_1,\beta_2}$ be the distribution of Y_1, Y_2 conditional on X_1, X_2 and for parameter β_1, β_2.
- **Minimax risk**

$$\mathcal{M}_X(k,\rho) := \inf_{\psi}\Big\{\sup_{\beta\in\mathbb{R}^p} P^X_{\beta,\beta}(\psi \neq 0) + \sup_{\substack{\beta_1,\beta_2\in\mathbb{R}^p \\ (\beta_1-\beta_2)/2\in\Theta_{p,k}(\rho)}} P^X_{\beta_1,\beta_2}(\psi \neq 1)\Big\},$$

Theorem. Assume (C1) and (C2). If $\frac{k\log p}{n} \to 0$ and $\rho \geq \sqrt{\frac{8k\log p}{n\kappa_1}}$ for $\kappa_1 := \frac{r}{(1+r)^2(1+s)}$, then $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 0$. Moreover, this asymptotic zero risk can be achieved by $\psi^{\text{sparse}}_{\lambda,\tau}$ with $\lambda = 2\sqrt{\log p}$ and $\tau \in (0, 2k\log p]$.

Upper and lower bounds, sparse case

- Let $P^X_{\beta_1,\beta_2}$ be the distribution of Y_1, Y_2 conditional on X_1, X_2 and for parameter β_1, β_2.
- **Minimax risk**

$$\mathcal{M}_X(k,\rho) := \inf_{\psi}\Big\{\sup_{\beta\in\mathbb{R}^p} P^X_{\beta,\beta}(\psi \neq 0) + \sup_{\substack{\beta_1,\beta_2\in\mathbb{R}^p\\ (\beta_1-\beta_2)/2\in\Theta_{p,k}(\rho)}} P^X_{\beta_1,\beta_2}(\psi\neq 1)\Big\},$$

Theorem. Assume (C1) and (C2). If $\frac{k\log p}{n} \to 0$ and $\rho \geq \sqrt{\frac{8k\log p}{n\kappa_1}}$ for $\kappa_1 := \frac{r}{(1+r)^2(1+s)}$, then $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 0$. Moreover, this asymptotic zero risk can be achieved by $\psi^{\text{sparse}}_{\lambda,\tau}$ with $\lambda = 2\sqrt{\log p}$ and $\tau \in (0, 2k\log p]$.

Theorem. Assume (C1) and (C2). If $k \leq p^\alpha$ for some $\alpha < 1/2$ and $\rho \leq \sqrt{\frac{(1-2\alpha-\epsilon)k\log p}{n\kappa_1}}$, then $\mathcal{M}_X(k,\rho) \xrightarrow{\text{a.s.}} 1$.

- For any $\alpha < 1/2$, the ℓ_2 radius of separation condition on ρ matches that in the upper bound (up to constants depending only on α).
- Thus $\psi_{\lambda,\tau}^{\mathrm{sparse}}$ is essentially minimax optimal — the complementary sketching eliminates the nuisance parameter without throwing away too much information.
- The quantity $n\kappa_1$ is in some sense the effective sample size in this problem:

$$n\kappa_1 = \frac{nr}{(1+r)^2(1+s)} \sim \frac{m}{n_1/n_2 + n_2/n_1 + 2}.$$

Proof sketch

1. Columns of W has ℓ_2 norms $\sim \sqrt{4n\kappa_1}$ and $\tilde{W}$ satisfies k-RIP with constant $C_{s,r}\sqrt{\frac{k\log p}{n}}$. To show this, we use the following decomposition

$$W^\top W = 4(X_1^\top X_1)(X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2) = 4T^\top B(I-B)T,$$

where $T \perp\!\!\!\perp B$ and B has a matrix-variate Beta distribution.

Proof sketch

1. Columns of W has ℓ_2 norms $\sim \sqrt{4n\kappa_1}$ and $\tilde{W}$ satisfies k-RIP with constant $C_{s,r}\sqrt{\frac{k\log p}{n}}$. To show this, we use the following decomposition

$$W^\top W = 4(X_1^\top X_1)(X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2) = 4T^\top B(I-B)T,$$

where $T \perp\!\!\!\perp B$ and B has a matrix-variate Beta distribution.

2. Size control under H_0: **hard**$(\tilde{W}^\top Z, \lambda)$ has most entries 0.

Proof sketch

1. Columns of W has ℓ_2 norms $\sim \sqrt{4n\kappa_1}$ and $\tilde{W}$ satisfies k-RIP with constant $C_{s,r}\sqrt{\frac{k\log p}{n}}$. To show this, we use the following decomposition

$$W^\top W = 4(X_1^\top X_1)(X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2) = 4T^\top B(I-B)T,$$

   where $T \perp\!\!\!\perp B$ and B has a matrix-variate Beta distribution.
2. Size control under H_0: $\textbf{hard}(\tilde{W}^\top Z, \lambda)$ has most entries 0.
3. Power control under H_1: let $\tilde{W}\tilde{\theta} = W\theta$, we have

$$\begin{aligned}\|\tilde{W}^\top Z\|_2 &= \|\tilde{W}^\top \tilde{W}\tilde{\theta} + \tilde{W}\xi\|_2 \\ &\approx \|\tilde{W}^\top \tilde{W}\tilde{\theta}\|_2 \approx \|\tilde{\theta}\|_2 \approx \sqrt{4n\kappa_1}\|\theta\|_2 \geq \sqrt{4n\kappa_1\rho^2}.\end{aligned}$$

Proof sketch

1. Columns of W has ℓ_2 norms $\sim \sqrt{4n\kappa_1}$ and $\tilde{W}$ satisfies k-RIP with constant $C_{s,r}\sqrt{\frac{k\log p}{n}}$. To show this, we use the following decomposition
$$W^\top W = 4(X_1^\top X_1)(X_1^\top X_1 + X_2^\top X_2)^{-1}(X_2^\top X_2) = 4T^\top B(I-B)T,$$
where $T \perp\!\!\!\perp B$ and B has a matrix-variate Beta distribution.
2. Size control under H_0: $\mathbf{hard}(\tilde{W}^\top Z, \lambda)$ has most entries 0.
3. Power control under H_1: let $\tilde{W}\tilde{\theta} = W\theta$, we have
$$\begin{aligned}\|\tilde{W}^\top Z\|_2 &= \|\tilde{W}^\top \tilde{W}\tilde{\theta} + \tilde{W}\xi\|_2 \\ &\approx \|\tilde{W}^\top \tilde{W}\tilde{\theta}\|_2 \approx \|\tilde{\theta}\|_2 \approx \sqrt{4n\kappa_1}\|\theta\|_2 \geq \sqrt{4n\kappa_1\rho^2}.\end{aligned}$$
4. Lower bound: constructing a mixture of the alternative that is close in chi-squared divergence to a null distribution.

Upper and lower bounds, dense case

Theorem. Assume (C1) and (C2). If $\frac{k \log p}{n} \to 0$ and $\rho \geq \sqrt{\frac{2m^{1/2} \log^{1/2} p}{n\kappa_1}}$, then $\psi_\eta^{\mathrm{dense}}$ with $\eta := m + \sqrt{8m \log p} + 4 \log p$ has asymptotic size 0 and power 1 almost surely.

Theorem. Assume (C1) and (C2). If $p^{1/2} \leq k \leq p^\alpha$ for some $\alpha \in [1/2, 1]$ and $\rho = o(p^{-1/4} \log^{-3/4} p)$, then $\mathcal{M}_X(k, \rho) \xrightarrow{\text{a.s.}} 1$.

- The upper and lower bounds are only matched up to logarithmic factors and constants depending on s and t.
- $k \asymp p^{1/2}$ is the boundary between the sparse and the dense cases.

Numerical studies

Simulation settings

- Generate X_1, X_2 with independent $N(0,1)$ entries and $\beta_1 \sim N_p(0, I_p)$.
- Given k and ρ, draw $\beta_2 - \beta_1$ uniformly at random from the set of k-sparse vectors with ℓ_2 norm ρ.
- Noise variance σ^2 estimated by a method-of-moment estimator (Dicker, 2014).
- Tuning parameters $\lambda = \sqrt{4 \log p}$, $\tau = 3 \log p$ and $\eta = m + \sqrt{8m \log p} + 4 \log p$.

- Define

$$\nu := \frac{m\rho^2}{(n_1/n_2 + n_2/n_1 + 2)k \log p}$$

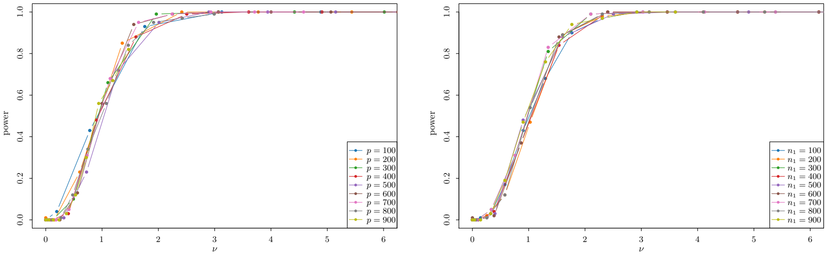
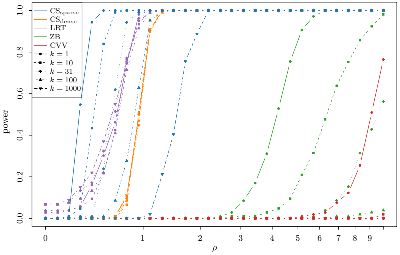
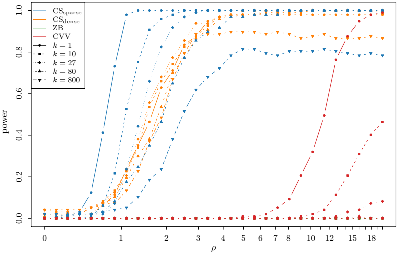


Figure: Power function of $\psi^{\text{sparse}}_{\lambda,\tau}$, estimated over 100 Monte Carlo repetitions, plotted against ν in various parameter settings. Left panel: $n_1 = n_2 = 500$, $p \in \{100, 200, \ldots, 900\}$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$. Right panel: $n_1 \in \{100, 200, \ldots, 900\}$, $n_2 = 1000 - n_1$, $p = 400$, $k = 10$, $\rho \in \{0, 0.2, \ldots, 2\}$.

Comparison with existing procedures

- Classical likelihood ratio test (requires $\min\{n_1, n_2\} > p$)
- Test proposed by Charbonnier, Verzelen and Villers (2015)
- Test proposed by Zhu and Bradic (2016) (requires $n_1 = n_2$)





Comparison with existing procedures

- Classical likelihood ratio test (requires $\min\{n_1, n_2\} > p$)
- Test proposed by Charbonnier, Verzelen and Villers (2015)
- Test proposed by Zhu and Bradic (2016) (requires $n_1 = n_2$)

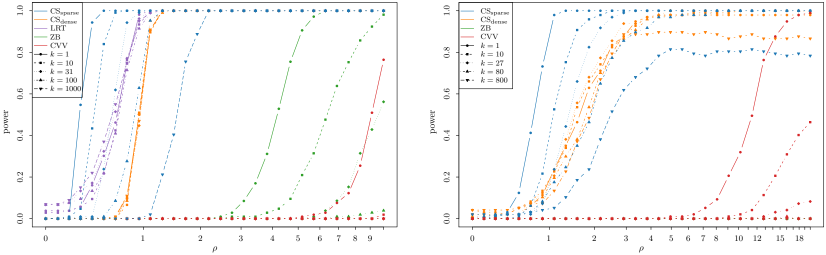


Figure: Power comparison of different methods at different sparsity levels $k \in \{1, 10, \lfloor p^{1/2} \rfloor, 0.1p, p\}$ and different signal ℓ_2 norm ρ on a logarithmic grid. Left panel: $n_1 = n_2 = 1200$, $p = 1000$; right panel: $n_1 = n_2 = 500$, $p = 800$.

Model misspecification

- What if design or noise assumptions are violated:
  (a) Correlated design: assume rows of X_1 and X_2 are independently drawn from $N(0, \Sigma)$ with $\Sigma = (2^{-|j_1 - j_2|})_{j_1, j_2 \in [p]}$.
  (b) Rademacher design: assume entries of X_1 and X_2 are independent Rademacher random variables.
  (c) One way balanced ANOVA design: assume $d_1 := n_1/p$ and $d_2 := n_2/p$ are integers and X_1 and X_2 are block diagonal matrices

$$X_1 = \begin{pmatrix} \mathbf{1}_{d_1} & & \\ & \ddots & \\ & & \mathbf{1}_{d_2} \end{pmatrix} \qquad X_2 = \begin{pmatrix} \mathbf{1}_{d_2} & & \\ & \ddots & \\ & & \mathbf{1}_{d_2} \end{pmatrix},$$

  (d) Heavy tailed noise: we generate both ϵ_1 and ϵ_2 with independent $t_4/\sqrt{2}$ entries.

Model misspecification

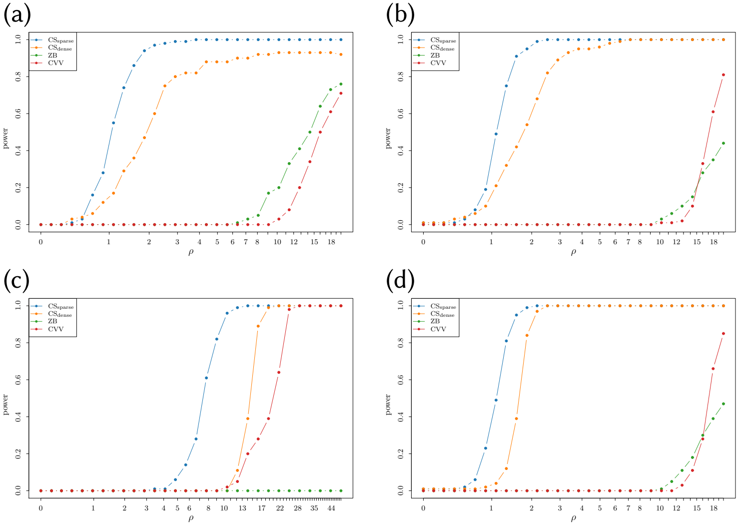


Figure: Power functions against signal ℓ_2 norm ρ on a logarithmic grid.

Summary

- It is possible to test for a sparse difference between two high-dimensional regression coefficients, even if both coefficients are dense.
- Complementary sketching eliminates the nuisance parameter and tests built upon it has an essentially optimal testing radius.

- Main reference:

  Gao, F. and Wang, T. (2020+) Two-sample testing of high-dimensional linear regression coefficients via complementary sketching. *arXiv preprint*, arxiv:2011.13624.

References

- Arias-Castro, E., Candés, E. J. and Plan, Y. (2011) Global testing under sparse alternatives. *Ann. Statist.*, **39**, 2533–2556.
- Cai, T. T., Liu, W. and Xia, Y. (2014) Two-sample test of high dimensional means under dependence. *J. Roy. Statist. Soc., Ser. B*, **76**, 349–372.
- Charbonnier, C., Verzelen, N. and Villers, F. (2015) A global homogeneity test for high-dimensional linear regression. *Electron. J. Statist.*, **9**, 318–382.
- Chen, S. X., Li, J. and Zhong, P.-S. (2019) Two-sample and ANOVA tests for high dimensional means. *Ann. Statist.*, **47**, 1443–1474.
- Dicker, L. H. (2014) Variance estimation in high-dimensional linear models. *Biometrika*, **101**, 269–284.
- Ingster, Y. I., Tsybakov, A. and Verzelen, N. (2010) Detection boundary in sparse regression. *Electron. J. Statist.*, **4**, 1476–1526.
- Städler, N. and Mukherjee, S. (2012) Two-sample testing in high-dimensional models. *arXiv preprint*, arxiv:1210.4584.
- Xia, Y., Cai, T. and Cai, T. T. (2018) Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statist. Sinica*, **28**, 63–92.
- Xia, Y., Cai, T. T. and Sun, W. (2020) Gap: A general framework for information pooling in two-sample sparse inference. *J. Amer. Statist. Assoc.*, **115**, 1236–1250.
- Zhu, Y. and Bradic, J. (2016) Two-sample testing in non-sparse high-dimensional linear models. *arXiv preprint*, arxiv:1610.04580.