論文 / 著書情報
Article / Book Information

Title	A statistical mechanics approach to de-biasing and uncertainty estimation in LASSO for random measurements
Authors	Takashi Takahashi, Yoshiyuki Kabashima
Citation	Journal of Statistical Mechanics: Theory and Experiment, Vol. 2018, , pp. 073405(1-25)
Pub. date	2018, 7
Creative Commons	See the 3rd page.
Note	This article is published under a CC BY licence. The Version of Record is available online at https://doi.org/10.1088/1742-5468/aace2e

PAPER: INTERDISCIPLINARY STATISTICAL MECHANICS • OPEN ACCESS

A statistical mechanics approach to de-biasing and uncertainty estimation in LASSO for random measurements

To cite this article: Takashi Takahashi and Yoshiyuki Kabashima *J. Stat. Mech.* (2018) 073405

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

A statistical mechanics approach to de-biasing and uncertainty estimation in LASSO for random measurements

Takashi Takahashi and Yoshiyuki Kabashima

Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan
E-mail: takahashi.t.cc@m.titech.ac.jp

Received 27 March 2018
Accepted for publication 19 June 2018
Published 16 July 2018

Online at stacks.iop.org/JSTAT/2018/073405
https://doi.org/10.1088/1742-5468/aace2e

Abstract. In high-dimensional statistical inference in which the number of parameters to be estimated is larger than that of the holding data, regularized linear estimation techniques are widely used. These techniques have, however, some drawbacks. First, estimators are biased in the sense that their absolute values are shrunk toward zero because of the regularization effect. Second, their statistical properties are difficult to characterize as they are given as numerical solutions to certain optimization problems. In this manuscript, we tackle such problems concerning LASSO, which is a widely used method for sparse linear estimation, when the measurement matrix is regarded as a sample from a rotationally invariant ensemble. We develop a new computationally feasible scheme to construct a de-biased estimator with a confidence interval and conduct hypothesis testing for the null hypothesis that a certain parameter vanishes. It is numerically confirmed that the proposed method successfully de-biases the LASSO estimator and constructs confidence intervals and p-values by experiments for noisy linear measurements.

Keywords: cavity and replica method, random matrix theory and extensions, statistical inference

1742-5468/18/073405+25$33.00

Contents

1. Introduction

Estimating high-dimensional unknown variables from a limited number of data precisely and reliably is an important task in statistics, machine learning, signal processing, and so on. For instance, such demands arise in compressed sensing [1, 2] and genomics [3]. Since, in these problems, the number of parameters often far surpasses that of observed data, it is clear that some sparsity assumptions on the parameters are necessary to reasonably estimate them. Therefore, one needs to simultaneously solve two problems: variable selection, which seeks relevant (or non-zero) parameters for the data generation process, and parameter estimation. In the past few decades, a number of methods have been developed to tackle such problems. One of the most successful approaches is the least absolute shrinkage and selection operator (LASSO) [4] method

J. Stat. Mech. (2018) 073405

for high-dimensional linear regression problems in which the estimator is obtained by minimizing the L_1 norm regularized likelihood function. As LASSO estimators can be easily obtained by versatile algorithms for convex optimization [2, 5] and have appealing consistency properties [6–8], they have received considerable attention.

Specifically, let us consider the linear measurement model:

$$y_i = \boldsymbol{a}_i^\top \boldsymbol{x}_0 + \xi_i,\ \xi_i \sim_{\text{i.i.d}} \mathcal{N}\left(0, \sigma^2\right),\ i = 1, 2, ..., M, \tag{1}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^N$ and $\boldsymbol{a}_i \in \mathbb{R}^N$ are the parameter (signal) and measurement vectors, respectively, $\sigma^2 \in \mathbb{R}$ is a parameter that describes the strength of the measurement noise, and $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. Notation $\top$ means the operation of matrix/vector transpose. In matrix notation, this model is expressed as

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{\xi}, \tag{2}$$

where $\boldsymbol{a}_i^\top$ corresponds to the ith row of the matrix $A \in \mathbb{R}^{M\times N}$. A is called the observation or measurement matrix by cases. The objective of high-dimensional linear regression is to find the parameter vector $\boldsymbol{x}_0$, where the number of measurements M is smaller than that of the parameter N. Note that in this high-dimensional setting, one cannot obtain a true solution with simple linear algebra because $A^\top A$ is not invertible; by contrast, in the classical setting where $M > N$, the unique unbiased estimator is easily obtained as $\widehat{\boldsymbol{x}}_{\text{classical}} = (A^\top A)^{-1} A^\top \boldsymbol{y}$ by using the least squares method. To achieve this aim, LASSO seeks an estimator by solving an optimization problem that imposes sparsity via an L_1 penalty:

$$\widehat{\boldsymbol{x}}^{\text{LASSO}}(\boldsymbol{y}, A; \lambda) \equiv \arg\min_{\boldsymbol{x}} \left[\frac{1}{2}\|\boldsymbol{y} - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1\right], \tag{3}$$

where λ is a hyperparameter that controls the strength of the regularization. This convex optimization problem can be solved efficiently by using versatile algorithms. Although LASSO might be seen as simple heuristics, it has an appealing consistency property: in a certain sparsity condition on the true parameter $\boldsymbol{x}_0$ and an appropriate control of the regularization strength λ, the LASSO solution and $\boldsymbol{x}_0$ are consistent in the sense that $\|\widehat{\boldsymbol{x}}^{\text{LASSO}} - \boldsymbol{x}_0\|_2^2/N$ vanishes as the measurement ratio $\gamma \equiv M/N$ tends to infinity. For a more comprehensive review of LASSO in the context of high-dimensional settings, see [9].

Unfortunately, LASSO also has some drawbacks. First, the LASSO solution is biased as long as $\lambda > 0$ is finite. The amplitude of the LASSO estimator $\widehat{\boldsymbol{x}}^{\text{LASSO}}$ is shrunk toward zero by the regularization term and its absolute value is typically smaller than that of the true parameter $\boldsymbol{x}_0$ even in an ideal sparsity assumption. Second, no explicit form of the distribution is available for the estimator, as it is just expressed as a numerical solution of (3). Consequently, one can neither construct confidence intervals nor perform hypothesis testing for the null hypothesis that a certain element of the parameter vanishes. These bottlenecks are considered to be problematic in real applications in which the statistical reliability of the estimation result should be assessed. This situation is different from the one of classical statistics in which one can analytically obtain an unbiased estimator and its distribution.

J. Stat. Mech. (2018) 073405

To resolve the problems stated above, in this study, we develop a new scheme for de-biasing and uncertainty estimation in the LASSO estimation in the case that the observation matrix A is generated from rotationally invariant random matrix ensembles, which are concretely defined in the next section. The uncertainty addressed in this study concerns the randomness that arises from the random observation matrix A and measurement noise $\boldsymbol{\xi}$. Our approach is based on a careful observation of the replica analysis of LASSO and an advanced mean-field method known as expectation consistent approximation or the adaptive Thouless–Anderson–Palmer (TAP) approach [10–12] developed in machine learning [13] and statistical mechanics. We numerically show that the proposed algorithm effectively de-biases the LASSO estimator and estimates its uncertainty.

The rest of this manuscript is organized as follows. In section 2, we explain the problem setting. In section 3, we describe the result of the replica analysis of LASSO and its physical implications. Then, the design of our scheme is introduced. The derivation of the free energy density is in appendix. In section 4, the proposed scheme is numerically tested by experiments for noisy linear measurements using various matrix ensembles. The last section provides a summary.

2. Problem setting

2.1. Model specification

In this study, we focus on random design models of (2), in which A is a random matrix and the true parameter vector $\boldsymbol{x}_0$ is sparse in the sense that the number of its non-zero components is limited to ϱN ($0 \leqslant \varrho < 1$). More precisely for A, we assume that for eigenvalue decomposition $A^\top A = ODO^\top$, O can be regarded as a random sample from the uniform distribution of the $N \times N$ orthogonal matrices and the empirical eigenvalue distribution $\sum_{i=1}^{N} \delta(\lambda - \lambda_i)/N$, where $\{\lambda_i\}_i$ are the eigenvalues of $A^\top A$, converges to a certain distribution $\rho(\lambda)$ in the limit $N \to \infty$ with probability one.

2.2. De-biasing and uncertainty estimation in LASSO

Let $\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}, A; \lambda)$ be the LASSO estimator for the given $\boldsymbol{y}, A$, and λ. We are interested in the two problems associated with $\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}, A; \lambda)$. The first problem is that the LASSO estimator is biased. In other words, $\left|\mathbb{E}\left[\widehat{x}_i^{\mathrm{LASSO}}\right]_{A,\boldsymbol{\xi}} - x_{0,i}\right|$, $(i = 1, 2, ..., M)$ remains finite for $\lambda > 0$ because of the shrinkage effect caused by the regularization term $\lambda\|\boldsymbol{x}\|_1$. The second is that the LASSO estimator does not have an explicit form of the distribution. As a consequence, one can neither construct a confidence interval nor compute a p-value to conduct hypothesis testing for the null hypothesis that a certain parameter vanishes.

In response to the aforementioned problems, we construct the following quantities. The first quantity is the de-biased estimators $\{\widehat{x}_i^{\mathrm{debiased}}\}_i$ that have confidence intervals $\{\mathcal{I}_i(\alpha_{\mathrm{sig}}) \equiv [\widehat{x}_i^{\mathrm{debiased}} - L_i(\alpha_{\mathrm{sig}}), \widehat{x}_i^{\mathrm{debiased}} + U_i(\alpha_{\mathrm{sig}})]\}_i$ with significance α_{sig}. The term *de-biased* means that this estimator coincides with the true parameter on average:

J. Stat. Mech. (2018) 073405

$\mathbb{E}[\widehat{x}_i^{\rm debiased}]_{A,\boldsymbol{\xi}} = x_{0,i}$. The second quantity is the p-values to test whether the LASSO estimator is zero or not. We are interested in hypothesis testing with the null hypothesis $H_{0,i} : x_{0,i} = 0$. The confidence intervals concerning the de-biased estimators and hypothesis testing via p-values assess the uncertainty in LASSO.

In the past few years, several researchers have been working on the issue closely related to that stated here [14–16]. These studies discuss de-biasing and hypothesis testing in high-dimensional statistics for a fixed observation matrix where the randomness comes from the measurement noise, under tight sparsity assumptions on a true sparse signal, which corresponds to the $\varrho \to 0$ limit in the current setting. In contrast to these studies, we concentrate on the case that the randomness comes from both the random observation matrix and the measurement noise without an explicit sparsity assumption on the true parameter keeping $\varrho \sim O(1)$.

3. A statistical mechanics approach

3.1. Replica analysis for general rotationally invariant random design matrices and its physical implications

To investigate how the LASSO solution depends on the true solution, observation matrix, and measurement noise, we first evaluate the free energy density corresponding to the LASSO Hamiltonian $H(\boldsymbol{x}) \equiv \|\boldsymbol{y} - A\boldsymbol{x}\|_2^2/2 + \lambda\|\boldsymbol{x}\|_1$ at a zero-temperature limit:

$$f(\lambda) \equiv -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{N\beta}\mathbb{E}\left[\ln Z(\boldsymbol{y}, A; \lambda)\right]_{A,\boldsymbol{\xi}}, \tag{4}$$

where β is the inverse temperature and Z is the partition function:

$$Z(A, \boldsymbol{y}; \lambda) = \int \exp\left(-\frac{\beta}{2}\|\boldsymbol{y} - A\boldsymbol{x}\|_2^2 - \beta\lambda\|\boldsymbol{x}\|_1\right) \mathrm{d}\boldsymbol{x}. \tag{5}$$

We take the limit $N \to \infty$ with $\gamma = M/N \sim O(1)$ fixed. In the zero-temperature limit $\beta \to \infty$, the Boltzmann distribution $\mathrm{e}^{-\beta H(\boldsymbol{x})}/Z$ is dominated by the configurations of the LASSO solution (3). Hence, one can evaluate how the LASSO estimator depends on $\boldsymbol{x}_0, A, \boldsymbol{\xi}$ by analyzing the macroscopic behavior of the typical free energy density (4) using statistical mechanics.

Since the Hamiltonian defined above has a mean-field nature in the sense that all the variables are weakly connected, the free energy density (4) can be evaluated by using the replica method:

$$\begin{aligned} f = \underset{\chi,\widehat{\chi},Q,\widehat{Q},m,\widehat{m}}{\mathrm{extr}} \Bigg[& G'(-\chi; J)(Q - 2m + \varrho - \chi\sigma^2) + \frac{\gamma}{2}\sigma^2 - \frac{\widehat{Q}Q}{2} + \frac{\widehat{\chi}\chi}{2} + \widehat{m}m \\ & + \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\int \min_{x_i}\left\{-\frac{\widehat{Q}}{2}x_i^2 + \left(\widehat{m}x_{0,i} + \sqrt{\widehat{\chi}}z_i\right)x_i - \lambda|x_i|\right\} Dz_i \Bigg], \end{aligned} \tag{6}$$

J. Stat. Mech. (2018) 073405

where $\mathrm{extr}_{\chi,\widehat{\chi},Q,\widehat{Q},m,\widehat{m}}\,\mathcal{F}(\chi,\widehat{\chi},Q,\widehat{Q},m,\widehat{m})$ denotes the extremization of the function $\mathcal{F}$ with respect to its arguments and $G'(x;J)$ is the derivative of $G(x;J)$ with respect to x. We have defined $\int(...)Dz, J, G(x)$ as follows:

$$\int(...)Dz \equiv \int(...)\frac{\mathrm{e}^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\mathrm{d}z, \tag{7}$$

$$J \equiv A^{\top}A, \tag{8}$$

$$G(x;J) \equiv \underset{z}{\mathrm{extr}}\left[-\int \rho_J(s)\ln|z-s|\,ds + \frac{zx}{2}\right] - \frac{1}{2}\ln x - \frac{1}{2}, \tag{9}$$

where $\rho_J(s)$ is the asymptotic eigenvalue distribution of J. The derivative of the function $G(x;J)$ has the following form:

$$G'(x;J) = \frac{1}{2}\left(z(x) - \frac{1}{x}\right), \tag{10}$$

where $z(x)$ is implicitly determined by the extremal condition of (9):

$$x = \mathcal{S}_J(z(x)) \equiv \int \frac{\rho_J(\lambda)}{z(x)-\lambda}\mathrm{d}\lambda. \tag{11}$$

The transformation $\mathcal{S}_J$ that appears in (11) is called the Stieltjes transformation of ρ_J. The introduced function G is connected to the R-transform $\mathcal{R}_J(\cdot)$ of the asymptotic eigenvalue distribution of J in studies of free probability theory [17]: $G(x;J) = \int_0^x \mathcal{R}_J(t)\mathrm{d}t$. Appendix provides a brief derivation of the free energy density (6).

The connection between the free energy density (6) and macroscopic observables is as follows. At the extremum, Q, m, and χ correspond to the macroscopic physical observables: $Q = \mathbb{E}\left[\langle|\boldsymbol{x}|^2\rangle\right]_{A,\boldsymbol{\xi}}/N$, $m = \mathbb{E}\left[\langle\boldsymbol{x}_0^{\top}\boldsymbol{x}\rangle\right]_{A,\boldsymbol{\xi}}/N$, and $\chi = \beta\mathbb{E}\left[\langle|\boldsymbol{x}|^2\rangle - |\langle\boldsymbol{x}\rangle|^2\right]_{A,\boldsymbol{\xi}}/N$. Each of these corresponds to the self-overlap, the overlap between the LASSO solutions and true solutions, and the macroscopic susceptibility. The notation $\langle...\rangle$ represents the Boltzmann average in the zero-temperature limit: $\langle...\rangle \equiv \lim_{\beta\to\infty}\int(...)\mathrm{e}^{-\beta H(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}/Z$. In addition, from direct calculation, one can show the following relationships between the free energy density, regularization term, and residual sum of squares:

$$f = \frac{\gamma}{2}\overline{\mathrm{RSS}} + \bar{r}, \tag{12}$$

$$\bar{r} \equiv \mathbb{E}\left[\left\langle\frac{1}{N}\sum_{i=1}^{N}|x_i|\right\rangle\right]_{A,\boldsymbol{\xi}} = \widehat{\chi}\chi + \widehat{m}m - \widehat{Q}Q, \tag{13}$$

$$\overline{\mathrm{RSS}} = \mathbb{E}\left[\mathrm{RSS}\right]_{A,\boldsymbol{\xi}} \equiv \mathbb{E}\left[\left\langle\frac{1}{M}\|\boldsymbol{y}-A\boldsymbol{x}\|_2^2\right\rangle\right]_{A,\boldsymbol{\xi}}, \tag{14}$$

$$= \frac{2}{\gamma}\left[G'(-\chi;J)(Q-2m+\varrho-\chi\sigma^2) + \frac{\gamma}{2}\sigma^2 - \frac{1}{2}\chi\widehat{\chi}\right], \tag{15}$$

J. Stat. Mech. (2018) 073405

where $\bar{r}$ and $\overline{\mathrm{RSS}}$ represent the per-element average of the regularization term and residual sum of squares, respectively. By using the relationships (13) and (15) and the extremal condition concerning $\widehat{Q}, \widehat{m}, \widehat{\chi}$, the conjugate fields $\widehat{Q}, \widehat{m}, \widehat{\chi}$ can be represented via the macroscopic physical variables:

$$\widehat{\chi} = \frac{\gamma G''(-\chi; J)}{G'(-\chi; J) - G''(-\chi; J)\chi}\overline{\mathrm{RSS}} + \frac{-G''(-\chi; J)\gamma + 2\left(G'(-\chi; J)\right)^2}{G'(-\chi; J) - G''(-\chi; J)\chi}\sigma^2, \tag{16}$$

$$\widehat{Q} = \widehat{m} = 2G'(-\chi; J). \tag{17}$$

Here, $\chi, G'(-\chi; J)$ and $G''(-\chi; J)$ are given as follows:

$$\chi = -\mathcal{S}_J(z(-\chi)), \tag{18}$$

$$G'(-\chi; J) = \frac{1}{2}\left(z(-\chi) + \frac{1}{\chi}\right), \tag{19}$$

$$G''(-\chi; J) = \frac{1}{2}\left(z'(-\chi) + \frac{1}{\chi^2}\right), \tag{20}$$

where $z'(-\chi)$ is obtained from the derivative of equation (11):

$$z'(-\chi) = -\left[\int \frac{\rho_J(\lambda)}{(z(-\chi) - \lambda)^2}\mathrm{d}\lambda\right]^{-1}. \tag{21}$$

The minimization problem in equation (6) corresponds to the effective single body problem, which determines the value of the local magnetization $\langle x_i \rangle$. Splitting into effective single body problems from the original multi-body estimation problem is called the *decoupling principle* in the literature on information theory [18, 19]. A comparison with the TAP/cavity analysis indicates that $h_i \equiv \widehat{m}x_{0,i} + \sqrt{\widehat{\chi}}z_i$ and $\widehat{m}$ correspond to the local field and Onsager reaction coefficient, respectively [21]. Here, $z_i \sim \mathcal{N}(0, 1)$ effectively represents the randomness that comes from the observation matrix and measurement noise. Figure 1 schematically shows the distribution of the local fields and how the local field determines the LASSO solution. Each local field is distributed according to the normal distribution $\mathcal{N}(\widehat{m}x_{0,i}, \widehat{\chi})$ and the LASSO solution is obtained by acting the soft-thresholding operator $\mathrm{ST}_{\lambda,\widehat{Q}}$ on it:

$$\widehat{x}_i^{\mathrm{LASSO}} = \mathrm{ST}_{\lambda,\widehat{Q}}(h_i) \equiv \frac{h_i - \lambda \mathrm{sgn}(h_i)}{\widehat{Q}}\Theta\left(|h_i| - \lambda\right), \tag{22}$$

where $\Theta(z)$ is Heaviside's step function.

The LASSO solution takes a non-zero value if the amplitude of the corresponding local field is larger than λ. Conversely, if and only if it is smaller than λ, the LASSO solution is exactly zero. Hereafter, we call the non-zero and zero components of the LASSO solution the *active* and *inactive* components, respectively.

The above observations indicate that once the local fields and $\widehat{m}, \widehat{\chi}$ are estimated from the LASSO solutions, one can construct an intended p-value P_i as

J. Stat. Mech. (2018) 073405

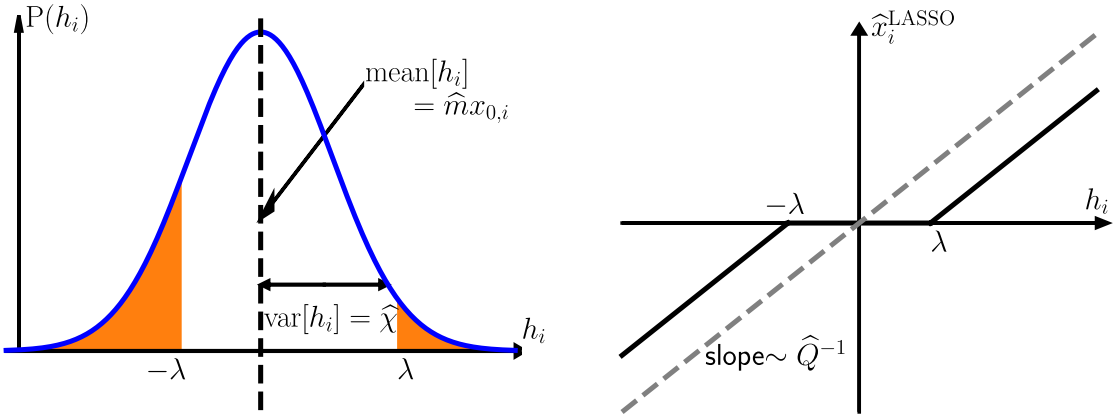

Figure 1. Left: the distribution of the local fields. The shaded region corresponds to the probability that the LASSO solution is active. Each local field is distributed according to the normal distribution $\mathcal{N}(\widehat{m}x_{0,i}, \widehat{\chi})$. In this example, $x_{0,i} < 0$. Right: local field dependence of the LASSO solution. The LASSO solution is determined by acting the soft-thresholding operator on the local field.

$$\mathrm{P}_i \equiv 2\left\{1 - \Phi\left(\frac{h_i}{\sqrt{\widehat{\chi}}}\right)\right\}, \tag{23}$$

$$\Phi(x) \equiv \int_{-\infty}^{x} \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\mathrm{d}t, \tag{24}$$

and an unbiased estimator as

$$\widehat{x}_i^{\mathrm{debiased}} \equiv \frac{h_i}{\widehat{Q}}, \tag{25}$$

with a confidence interval

$$\mathcal{I}_i(\alpha_{\mathrm{sig}}) = \left[\frac{h_i}{\widehat{Q}} - \Phi^{-1}\left(1 - \frac{\alpha_{\mathrm{sig}}}{2}\right)\frac{\sqrt{\widehat{\chi}}}{\widehat{Q}}, \frac{h_i}{\widehat{Q}} + \Phi^{-1}\left(1 - \frac{\alpha_{\mathrm{sig}}}{2}\right)\frac{\sqrt{\widehat{\chi}}}{\widehat{Q}}\right], \tag{26}$$

of significance α_{sig}. These are the key observations for the design of our scheme.

3.2. Adaptive TAP approach to constructing local fields and their variances from LASSO solutions

3.2.1. Derivation of the adaptive TAP equations. To derive the relation between the LASSO solution $\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(= \langle \boldsymbol{x} \rangle)$ and the local fields, let us consider Gibbs free energy:

$$G(\boldsymbol{m}) \equiv \operatorname*{extr}_{\boldsymbol{h}}\left[\boldsymbol{h}^{\top}\boldsymbol{m} - \frac{1}{\beta}\ln\left\{\mathrm{e}^{-\frac{\beta}{2}\|\boldsymbol{y}-A\boldsymbol{x}\|_2^2 + \beta\boldsymbol{h}^{\top}\boldsymbol{x} - \beta\lambda\|\boldsymbol{x}\|_1}\mathrm{d}\boldsymbol{x}\right\}\right]. \tag{27}$$

The average $\langle \boldsymbol{x} \rangle$ is determined as the global minimizer of $G(\boldsymbol{m})$: $\langle \boldsymbol{x} \rangle = \arg\min_{\boldsymbol{m}} G(\boldsymbol{m})$. Once the above Gibbs free energy is exactly calculated, the extremal conditions of $\boldsymbol{h}$ and $\boldsymbol{m}$ generally associate the average $\langle \boldsymbol{x} \rangle$ and local field [20]. However, the evaluation

J. Stat. Mech. (2018) 073405

of equation (27) is computationally difficult in general. To overcome this difficulty, we take the following expectation consistent with the adaptive TAP approach [10–12]. First, we define an alternative Gibbs free energy:

$$G(\boldsymbol{m},Q)\equiv\underset{\boldsymbol{h},\Lambda}{\mathrm{extr}}\left[\boldsymbol{h}^{\top}\boldsymbol{m}-\frac{N}{2}\Lambda Q-\frac{1}{\beta}\ln\left\{\int \mathrm{e}^{-\frac{\beta}{2}\|\boldsymbol{y}-A\boldsymbol{x}\|_2^2+\beta\boldsymbol{h}^{\top}\boldsymbol{x}-\frac{\beta}{2}\Lambda\|\boldsymbol{x}\|_2^2-\beta\lambda\|\boldsymbol{x}\|_1}\mathrm{d}\boldsymbol{x}\right\}\right], \tag{28}$$

which provides the constraints on the first and macroscopic second moments so that $\langle\boldsymbol{x}\rangle, \langle|\boldsymbol{x}|^2\rangle/N = \arg\min_{\boldsymbol{m},Q} G(\boldsymbol{m},Q)$.

Unfortunately, equation (28) is also difficult to evaluate directly. The adaptive TAP approach resorts this calculation to the following approximation:

$$G(\boldsymbol{m},Q)\simeq\phi_{\mathrm{ada}}(\boldsymbol{m},Q)\equiv\widetilde{\phi}(\boldsymbol{m},Q;l=0)+\phi^{\mathrm{G}}(\boldsymbol{m},Q;l=1)-\phi^{\mathrm{G}}(\boldsymbol{m},Q;l=0), \tag{29}$$

$$\widetilde{\phi}(\boldsymbol{m},Q;l)\equiv\underset{\boldsymbol{h},\Lambda}{\mathrm{extr}}\left\{\boldsymbol{h}^{\top}\boldsymbol{m}-\frac{N}{2}\Lambda Q-\frac{1}{\beta}\ln\int \mathrm{e}^{-\frac{\beta l}{2}\|\boldsymbol{y}-A\boldsymbol{x}\|_2^2+\beta\boldsymbol{h}^{\top}\boldsymbol{x}-\frac{\beta}{2}\Lambda\|\boldsymbol{x}\|_2^2-\beta\lambda\|\boldsymbol{x}\|_1}\mathrm{d}\boldsymbol{x}\right\}, \tag{30}$$

$$\phi^{\mathrm{G}}(\boldsymbol{m},Q;l)\equiv\underset{\boldsymbol{h}^{\mathrm{G}},\Lambda^{\mathrm{G}}}{\mathrm{extr}}\left\{\boldsymbol{h}_{\mathrm{G}}^{\top}\boldsymbol{m}-\frac{N}{2}\Lambda_{\mathrm{G}}Q-\frac{1}{\beta}\ln\int \mathrm{e}^{-\frac{\beta l}{2}\|\boldsymbol{y}-A\boldsymbol{x}\|_2^2+\beta\boldsymbol{h}_G^{\top}\boldsymbol{x}-\frac{\beta}{2}\Lambda_G\|\boldsymbol{x}\|_2^2}\mathrm{d}\boldsymbol{x}\right\}, \tag{31}$$

where $\widetilde{\phi}(\boldsymbol{m},Q;l=0)$, $\phi^{\mathrm{G}}(\boldsymbol{m},Q;l=1)$, and $\phi^{\mathrm{G}}(\boldsymbol{m},Q;l=0)$ are the free energies for the modified distributions: the first term is a factorized distribution but contains the original non-Gaussian prior factor $\mathrm{e}^{-\beta\lambda\|\boldsymbol{x}\|_1}$, while the second and third terms are the global and factorized multivariate Gaussian distribution that replaces the prior factor $\mathrm{e}^{-\beta\lambda\|\boldsymbol{x}\|_1}$ with a Gaussian factor $\mathrm{e}^{-\beta\Lambda_{\mathrm{G}}\|\boldsymbol{x}\|_2^2/2}$. In contrast to the original form of Gibbs free energy (28), adaptive TAP free energy (29) can be easily calculated as it is composed of only integration over the multivariate Gaussian and factorized distributions. The evaluation of the integrals and extremal conditions in the second and third terms of equation (29) provides the following expression of ϕ_{ada}:

$$\begin{aligned}\phi_{\mathrm{ada}}(\boldsymbol{m},Q)=\underset{\boldsymbol{h},\Lambda}{\mathrm{extr}}\Bigg[&\frac{1}{2}\|\boldsymbol{y}-A\boldsymbol{m}\|_2^2-\frac{N\Lambda Q}{2}-\frac{N}{\beta}G(-\chi;J)\\&+\boldsymbol{h}^{\top}\boldsymbol{m}-\frac{1}{2\Lambda}\sum_{i=1}^{N}(|h_i|-\lambda)^2\Theta(|h_i|-\lambda)\Bigg],\end{aligned} \tag{32}$$

where $\chi\equiv\beta(Q-q), q\equiv\sum_i m_i^2/N$. It has been shown [11, 22] that the above free energy $\phi_{\mathrm{ada}}(\boldsymbol{m},Q)$ is asymptotically consistent with replica theory in the sense that $\lim_{\beta\to\infty,N\to\infty}\mathbb{E}\left[\mathrm{extr}_{\boldsymbol{m},Q}\,\phi_{\mathrm{ada}}(\boldsymbol{m},Q)\right]_{A,\xi}/N=\mathbb{E}\left[f\right]_{A,\boldsymbol{\xi}}$ when A is a sample from the rotationally invariant ensemble. Thus, the extremal condition on $\boldsymbol{h},\Lambda,\boldsymbol{m},Q$ and linear response argument give the intended TAP/cavity equations, which connect the local field and LASSO estimator for the current matrix ensembles for $\beta\to\infty, N\to\infty$:

$$\boldsymbol{h}=\Lambda\boldsymbol{m}+A^{\top}(\boldsymbol{y}-A\boldsymbol{m}), \tag{33}$$

$$m_i=\frac{h_i-\lambda\,\mathrm{sgn}(h_i)}{\Lambda}\Theta(|h_i|-\lambda), \tag{34}$$

$$\Lambda=2G'(-\chi), \tag{35}$$

J. Stat. Mech. (2018) 073405

$$\chi = \frac{1}{N\Lambda}\sum_{i=1}^{N}\Theta(|h_i| - \lambda) = \frac{\varrho_{\rm active}}{\Lambda}, \tag{36}$$

where $\varrho_{\rm active} \equiv \sum_{i=1}^{N}\Theta(|h_i| - \lambda)/N = |\{i|\widehat{x}_i^{\rm LASSO} \neq 0\}|/N$ is the active component density of the LASSO solution (3).

3.2.2. General construction procedure of the de-biased estimator, confidence interval, and p-value. In summary, once the LASSO estimator $\widehat{\boldsymbol{x}}^{\rm LASSO}(\boldsymbol{y}, A; \lambda)$ is obtained for a set of $(\boldsymbol{y}, A, \lambda)$ by using versatile algorithms for the optimization problem (3) such as least-angle regression [25], coordinate descent [26], and various approximate message passing algorithms [7, 27–29], one can estimate the local fields $\boldsymbol{h}(\boldsymbol{y}, A; \lambda)$, de-biased estimator $\widehat{\boldsymbol{x}}^{\rm debiased}(\boldsymbol{y}, A; \lambda)$, confidence interval $\{\mathcal{I}_i(\alpha_{\rm sig})\}_i$, and p-value P_i as follows. We emphasize here that there is no need to use the derived TAP equation to obtain a LASSO estimator.

First, the active component density $\varrho_{\rm active}$ is calculated from the LASSO solution:

$$\varrho_{\rm active}(\boldsymbol{y}, A; \lambda) = \frac{1}{N}\left|\{i|\widehat{x}_i^{\rm LASSO}(\boldsymbol{y}, A; \lambda) \neq 0\}\right|. \tag{37}$$

Second, $z(-\chi)$ is obtained by combining equations (18), (19), (35) and (36): $z(-\chi)$ is obtained as the solution of

$$z = \frac{1 - \varrho_{\rm active}}{\mathcal{S}_J(z)}. \tag{38}$$

This equation is solved analytically or numerically depending on the cases. Note that this equation is easily solved by using a simple iteration algorithm even if an analytical expression is not obtained. Then, $z'(-\chi), \chi, G'(-\chi; J), G''(-\chi; J)$, the Onsager coefficient $\widehat{Q} = \Lambda$, the local field $h(\boldsymbol{y}, A; \lambda)$, the de-biased estimator $\widehat{\boldsymbol{x}}^{\rm debiased}(\boldsymbol{y}, A; \lambda)$, the residual sum of squares, and the variance of the local field $\widehat{\chi}$ are obtained by subsequently substituting the obtained values into equations (14), (16), (18)–(21), (25), (33) and (35):

$$z'(-\chi) = -\left[\int \frac{\rho_J(\lambda)}{(z(-\chi) - \lambda)^2}\mathrm{d}\lambda\right]^{-1}, \tag{39}$$

$$\chi = -\mathcal{S}_J(z(-\chi)), \tag{40}$$

$$G'(-\chi; J) = \frac{1}{2}\left(z(-\chi) + \frac{1}{\chi}\right), \tag{41}$$

$$G''(-\chi; J) = \frac{1}{2}\left(z'(-\chi) + \frac{1}{\chi^2}\right), \tag{42}$$

$$\widehat{Q} = \Lambda = z(-\chi) + \frac{1}{\chi}, \tag{43}$$

J. Stat. Mech. (2018) 073405

$$\boldsymbol{h}(\boldsymbol{y}, A; \lambda) = \widehat{Q}\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}, A; \lambda) + A^{\top}\left(\boldsymbol{y} - A\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}, A; \lambda)\right), \tag{44}$$

$$\widehat{\boldsymbol{x}}^{\mathrm{debiased}}(\boldsymbol{y}, A; \lambda) = \frac{\boldsymbol{h}(\boldsymbol{y}, A; \lambda)}{\widehat{Q}}, \tag{45}$$

$$\mathrm{RSS} = \frac{1}{M}\left\|\boldsymbol{y} - A\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}, A; \lambda)\right\|_2^2, \tag{46}$$

$$\widehat{\chi} = \frac{\gamma G''(-\chi; J)}{G'(-\chi; J) - G''(-\chi; J)\chi}\mathrm{RSS} + \frac{-G''(-\chi; J)\gamma + 2\left(G'(-\chi; J)\right)^2}{G'(-\chi; J) - G''(-\chi; J)\chi}\sigma^2. \tag{47}$$

Finally, the de-biased estimator's confidence interval $\{\mathcal{I}_i(\alpha_{\mathrm{sig}})\}_i$ and p-value $\{\mathrm{P}_i\}_i$ are obtained based on equations (23)–(26).

Note that a consistent estimator of the error variance σ^2 should be needed when it is unknown.

4. Numerical experiment

4.1. Settings

We perform numerical experiments to assess the usefulness of the proposed scheme[1]. For this, we artificially generate the true sparse parameter $\boldsymbol{x}_0$, observation matrix A, and measurement noise $\boldsymbol{\xi}$. The true sparse parameter $\boldsymbol{x}_0$ is generated from the Bernoulli–Gauss distribution: $x_{0,i} \sim_{\mathrm{i.i.d.}} (1-\varrho)\delta(x_{0,i}) + \varrho\mathcal{N}(0,1)$ for $i = 1, 2, ..., N$. The measurement noise is distributed according to the Gaussian distribution $\boldsymbol{\xi} \sim \mathcal{N}(0_M, \sigma^2 I_M)$. For the random observation matrix ensembles, the following ensembles are considered.

(i) The random i.i.d. Gaussian ensemble in which all entries of A are i.i.d. Gaussian variables with mean 0 and variance $1/N$. For this ensemble, the asymptotic eigenvalue distribution is given as the Marchenko–Pastur distribution [30]:

$$\rho(s) = (1-\gamma)\delta(s) + \frac{\gamma}{2\pi}\frac{\sqrt{(\lambda_+ - s)(s - \lambda_-)}}{s}\mathbb{I}_{[\lambda_-, \lambda_+]}(s), \tag{48}$$

$$\lambda_\pm = \left(1 \pm \sqrt{\gamma}\right)^2, \tag{49}$$

$$\mathbb{I}_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}. \tag{50}$$

Then, the form of $G'(-\chi; J), G''(-\chi; J), \chi$ and $\widehat{Q}$ are given as follows:

$$G'(-\chi; J) = \frac{\gamma}{2}\frac{1}{1+\chi}, \tag{51}$$

[1] Some demonstration codes are available at https://github.com/takashi-takahashi/debiasing_lasso_demo.

J. Stat. Mech. (2018) 073405

$$G''(-\chi;J)=\frac{\gamma}{2}\frac{1}{(1+\chi)^2}, \tag{52}$$

$$\chi=\frac{\varrho_{\text{active}}}{\gamma-\varrho_{\text{active}}}, \tag{53}$$

$$\widehat{Q}=\gamma-\varrho_{\text{active}}. \tag{54}$$

By substituting the above expressions of G', G'' into (16), one can show that $\widehat{\chi}$ does not depend on σ^2. This is the characteristic property of this ensemble. Generally, $\widehat{\chi}$ depends on the measurement noise σ^2.

(ii) The row-orthogonal ensemble [22, 31] constructed by randomly selecting M rows from a randomly generated $N\times N$ orthogonal matrix. For this ensemble, the asymptotic eigenvalue distribution is given as $\rho(s)=(1-\gamma)\delta(s)+\gamma\delta(s-1)$. In this case, the form of $G'(-\chi;J), G''(-\chi;J), \chi$ and $\widehat{Q}$ are given as follows:

$$G'(-\chi;J)=\frac{1}{2}\left(z(-\chi)+\frac{1}{\chi}\right), \tag{55}$$

$$G''(-\chi;J)=\frac{1}{2}\left(z'(-\chi)+\frac{1}{\chi^2}\right), \tag{56}$$

$$\chi=\frac{\rho_{\text{A}}(1-\varrho_{\text{active}})}{\gamma-\varrho_{\text{active}}}, \tag{57}$$

$$\widehat{Q}=\frac{\gamma-\varrho_{\text{active}}}{1-\varrho_{\text{active}}}, \tag{58}$$

where

$$z(-\chi)=-\frac{1-\chi+\sqrt{(\chi+1)^2-4\gamma\chi}}{2\chi}, \tag{59}$$

$$z'(-\chi)=-\frac{1-2\gamma\chi+\chi+\sqrt{(\chi+1)^2-4\gamma\chi}}{2\chi^2\sqrt{(\chi+1)^2-4\gamma\chi}}. \tag{60}$$

(iii) The random discrete cosine transform (DCT) ensemble in which A is constructed by randomly selecting M rows from $N\times N$ DCT matrix. While this ensemble shares the same eigenvalue distribution as the row-orthogonal one, it is much more relevant for practical purposes, as the computational cost for observation and inference can be significantly reduced by using the fast Fourier transform technique. In addition, although the rotationally invariant assumption on O does not hold, this ensemble is also compatible with the current adaptive TAP scheme, as pointed out by [24].

J. Stat. Mech. (2018) 073405

(iv) The geometric setup [31, 32] in which A is constructed as $A = U\Sigma V^\top$, where $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{N\times N}$ are random samples from the uniform distribution of orthogonal matrices, and $\Sigma \in \mathbb{R}^{M\times N}$ is a diagonal matrix whose (i,i)th element is given by $\nu_i \propto \tau^{i-1}$ for $i = 1, 2, ..., M$. The parameter $\tau \in (0, 1)$ is chosen so that the given value of the peak-to-average eigenvalue ratio

$$\kappa \equiv \frac{\nu_1^2}{M^{-1}\sum_{i=1}^M \nu_i^2} \tag{61}$$

is met and the singular values are scaled to satisfy the power constraint $1 = \frac{1}{N}\sum_{i=1}^M \nu_i^2$. The asymptotic eigenvalue distribution is given as

$$\rho(s) = (1-\gamma)\delta(s) + \frac{\gamma}{\eta s}\mathbb{I}_{(B\mathrm{e}^{-\eta},B)}(s), \tag{62}$$

where η and B are related to the peak-to-average ratio κ:

$$\kappa = \frac{\eta}{1-\mathrm{e}^{-\eta}}, \tag{63}$$

$$B = \frac{\kappa}{\gamma}. \tag{64}$$

In this case, the explicit form of G', G'' cannot be obtained. Thus, it should be evaluated numerically. To achieve this aim, we conduct the procedure explained in section 4.2.3, using the expression of the Stieltjes transform and $z'(-\chi)$:

$$\chi = -\mathcal{S}_J(z(-\chi)) = -\int \frac{\rho(s)}{z(-\chi)-s}\mathrm{d}\lambda = -\frac{1}{z(-\chi)}\left[1 - \frac{\alpha}{\eta}\ln\frac{z(-\chi)-B}{z(-\chi)-B\mathrm{e}^{-\eta}}\right], \tag{65}$$

$$z'(-\chi) = \frac{z(-\chi)^2}{-1+\frac{\gamma}{\eta}\ln\frac{z(-\chi)-B}{z(-\chi)-B\mathrm{e}^{-\eta}} - \frac{z(-\chi)}{(z(-\chi)-B\mathrm{e}^{-\eta})(z(-\chi)-B)}}. \tag{66}$$

We mainly use the random i.i.d. Gaussian ensemble and random DCT ensemble for the numerical experiments. The geometric setup is only used in section 4.2.3. We do not use the original row-orthogonal setup.

Once a tuple of $(\boldsymbol{x}_0, A, \boldsymbol{\xi})$ is generated, we calculate $\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}$, $\boldsymbol{h}$, $\chi, \widehat{\chi}$ and $\widehat{Q} = \Lambda$, $\widehat{\boldsymbol{x}}^{\mathrm{debiased}}$ by using the procedure explained in section 3.2.2. To estimate the error variance σ^2 needed in the random DCT case, we use the naive cross-validation-based estimator:

$$\hat{\sigma}^2(\boldsymbol{y}, A; \hat{\lambda}) \equiv \frac{1}{M - N\varrho_{\mathrm{active}}}\left\|\boldsymbol{y} - A\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}, A; \hat{\lambda})\right\|_2^2, \tag{67}$$

where $\hat{\lambda}$ is selected by K-fold cross-validation. In [23], it is empirically shown that this estimator robustly estimates the error variance, more so than its competitors.

J. Stat. Mech. (2018) 073405

We use $N_{\rm s}=1000$ different sets of pairs $(A,\boldsymbol{\xi})$ for fixed $\boldsymbol{x}_0$ to evaluate the statistical properties of the observables. We set $\varrho=0.1,\gamma=0.5,\sigma^2=0.02$, and $K=40$, except for the geometric setup. In the geometric setup, we set $\varrho=0.1,\gamma=0.8,\sigma^2=0.02$, and $\kappa=8$.

4.2. Results

4.2.1. Distribution of the local fields and de-biased estimators. First, we examine the statistical properties of the local fields and de-biased estimators. Figure 2 plots the sample quantiles of $\{(h_i-\widehat{Q}x_{0,i})/\sqrt{\widehat{\chi}}\}_i$ versus the theoretical quantiles of the standard normal distribution for one configuration of $(\boldsymbol{x}_0,A,\boldsymbol{\xi})$. It is clear that all the points are close to the line with unit slope and zero intercept. Further, figure 3 plots the average values of the $\widehat{\boldsymbol{x}}^{\rm LASSO}$ and $\widehat{\boldsymbol{x}}^{\rm debiased}$ versus the true parameter $\boldsymbol{x}_0$. In contrast to the LASSO estimators, which are shrunk toward zero by the regularization term, $\widehat{\boldsymbol{x}}^{\rm debiased}$ efficiently reduces the LASSO estimator's bias. The average is taken over N_s realizations of $(A,\boldsymbol{\xi})$. These results validate our theoretical predictions on the local fields and de-biased estimators. Figure 4 plots the constructed de-biased estimators and their 95% confidence intervals. We show only the first 80 components for the sake of clarity.

Although figures 2–4 show the results for one value of λ, the same results are obtained for a wide range of λ. The means of $\{h_i-\widehat{Q}x_{0,i}\}_i$ and $\{\widehat{x}^{\rm debiased}-x_{0,i}\}_i$ are zero in both the i.i.d. Gaussian and the random DCT cases (figures 5(a) and (b)). Further, the variances of $\{h_i-\widehat{Q}x_{0,i}\}_i$ and $\{\widehat{x}_i^{\rm debiased}-x_{0,i}\}_i$ agree with their estimates of $\widehat{\chi}$ and $\widehat{\chi}/\widehat{Q}^2$, respectively for the whole range of the weight of the L_1 regularizer λ (figures 5(c) and (d)).

4.2.2. Hypothesis testing. An important advantage of the proposed scheme over LASSO is that it provides a hypothesis testing method with a null hypothesis that a certain parameter vanishes. Although LASSO provides a parameter selection rule that selects an active component set $\mathcal{A}(\boldsymbol{y},A;\lambda)$ as $\mathcal{A}(\boldsymbol{y},A;\lambda)=\{i|\widehat{x}_i^{\rm LASSO}(\boldsymbol{y},A;\lambda)\neq 0\}$, it cannot measure the statistical significance for finding an active component.

Specifically, we are interested in testing an individual null hypothesis $H_{0,i}:x_{0,i}=0$ versus the alternative hypothesis $H_{1,i}:x_{0,i}\neq 0$, assigning a p-value of P_i for these tests. To this end, we evaluate the p-value of $\{\mathrm{P}_i\}$ by using equation (23) for a two-tailed test. Then, the decision rule is to reject the null hypothesis $H_{0,i}$ if the observed p-value P_i is lower than $\tilde{\alpha}_{\rm sig}$ and to accept the alternative hypothesis otherwise:

$$\widehat{T}_i(\boldsymbol{y},A;\lambda)=\begin{cases}1 & \text{if } \mathrm{P}_i\leqslant\tilde{\alpha}_{\rm sig}\ \text{(reject)}\\ 0 & \text{otherwise (accept)}\end{cases}, \tag{68}$$

where $\tilde{\alpha}_{\rm sig}$ is the significance level. We use $\widehat{T}$ as a rejection flag. This procedure ensures that the type I error probability or the FPR is $\tilde{\alpha}_{\rm sig}$. Here, the FPR is the probability of falsely rejecting the null hypothesis $H_{0,i}$:

$$\mathrm{FPR}\equiv\frac{\left|\left\{i|\widehat{T}_i=1\ \text{and}\ x_{0,i}=0\right\}\right|}{|\{i|x_{0,i}=0\}|}. \tag{69}$$

J. Stat. Mech. (2018) 073405

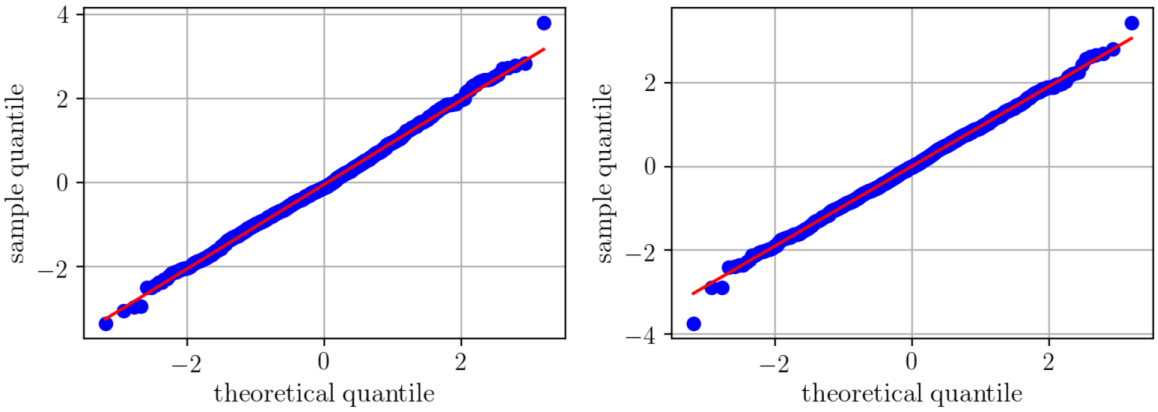

Figure 2. Q-Q plot of $\{(h_i - x_{0,i}\widehat{Q})/\sqrt{\widehat{\chi}}\}_i$. The red line is the unit slope and zero intercept line. Left: the i.i.d. Gaussian case. Right: the random DCT case.

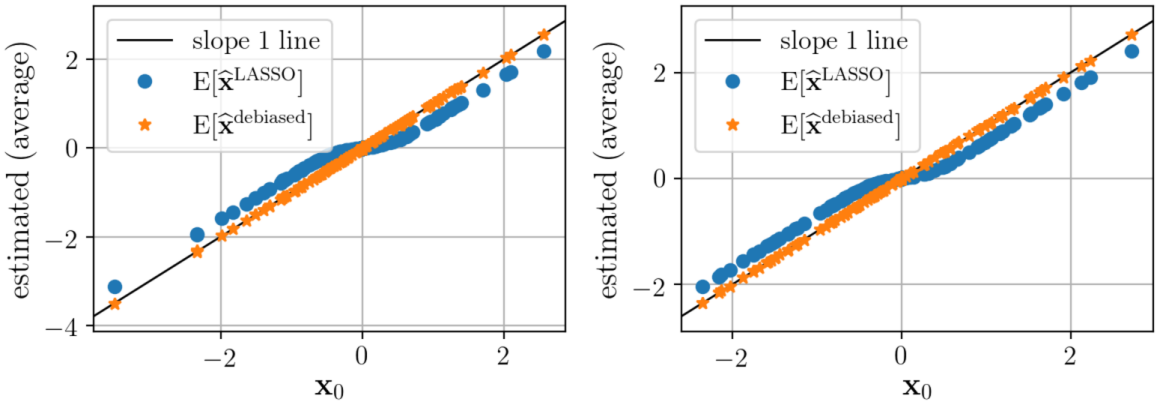

Figure 3. De-biasing effect of $\widehat{\boldsymbol{x}}^{\mathrm{debiased}}$. The blue points stand for the average value of the LASSO solution $\widehat{\boldsymbol{x}}^{\mathrm{LASSO}}$ and orange points stand for the average value of the de-biased estimator $\widehat{\boldsymbol{x}}^{\mathrm{debiased}}$. The black line is the unit slope and zero intercept line. Left: the i.i.d. Gaussian ensemble case. Right: the random DCT ensemble case.

Indeed, figure 6 shows that the significance level $\tilde{\alpha}_{\mathrm{sig}}$ and empirical false positive rate (FPR) are in excellent agreement.

Further, we examine the relation between the FPR and TPR or the statistical power achieved by LASSO and our hypothesis testing procedure. Here, the TPR is the probability that the test correctly rejects the null hypothesis $H_{0,i}$:

$$\mathrm{TPR} \equiv \frac{\left|\left\{i|\widehat{T}_i = 1 \text{ and } x_{0,i} \neq 0\right\}\right|}{|\{i|x_{0,i} \neq 0\}|}. \tag{70}$$

Note that although we can control the FPR by varying the significance level $\tilde{\alpha}_{\mathrm{sig}}$, the TPR cannot be controlled. Thus, a performance measure of the variable selection procedure by hypothesis testing can be given as the TPR for each value of the FPR. We evaluate the performance of hypothesis testing by using the ROC curve, which plots the TPR versus the FPR as an implicit function of $\tilde{\alpha}_{\mathrm{sig}}$. We examine the TPR and FPR by varying the significance level $\tilde{\alpha}_{\mathrm{sig}}$ for each regularization parameter λ. For

J. Stat. Mech. (2018) 073405

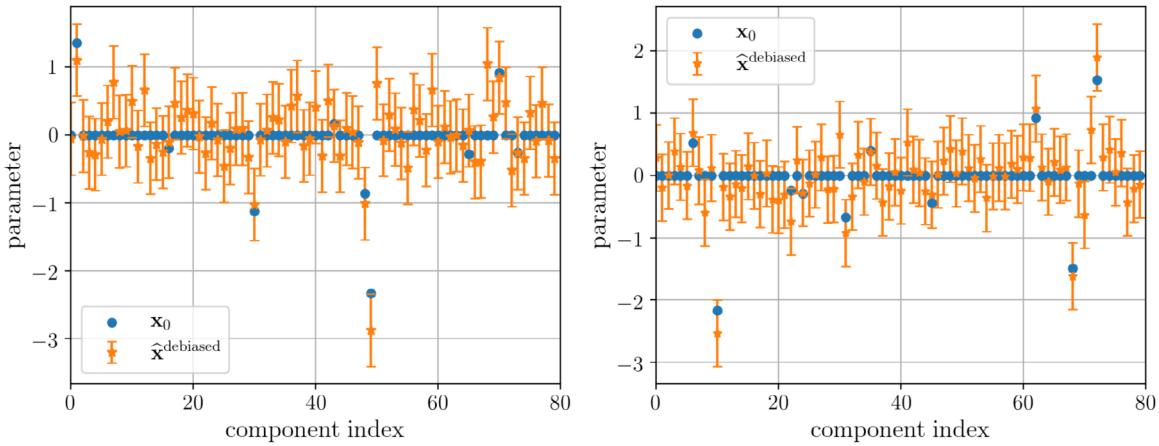

Figure 4. Constructed de-biased estimator $\widehat{\boldsymbol{x}}^{\mathrm{debiased}}$ and its 95% confidence interval. In both the left and the right panels, the blue points stand for the true parameter $\boldsymbol{x}_0$ and orange points are the de-biased estimator $\widehat{\boldsymbol{x}}^{\mathrm{debiased}}$. The orange error bars are the 95% confidence intervals. Left: the i.i.d. Gaussian ensemble case. Right: the random DCT ensemble case.

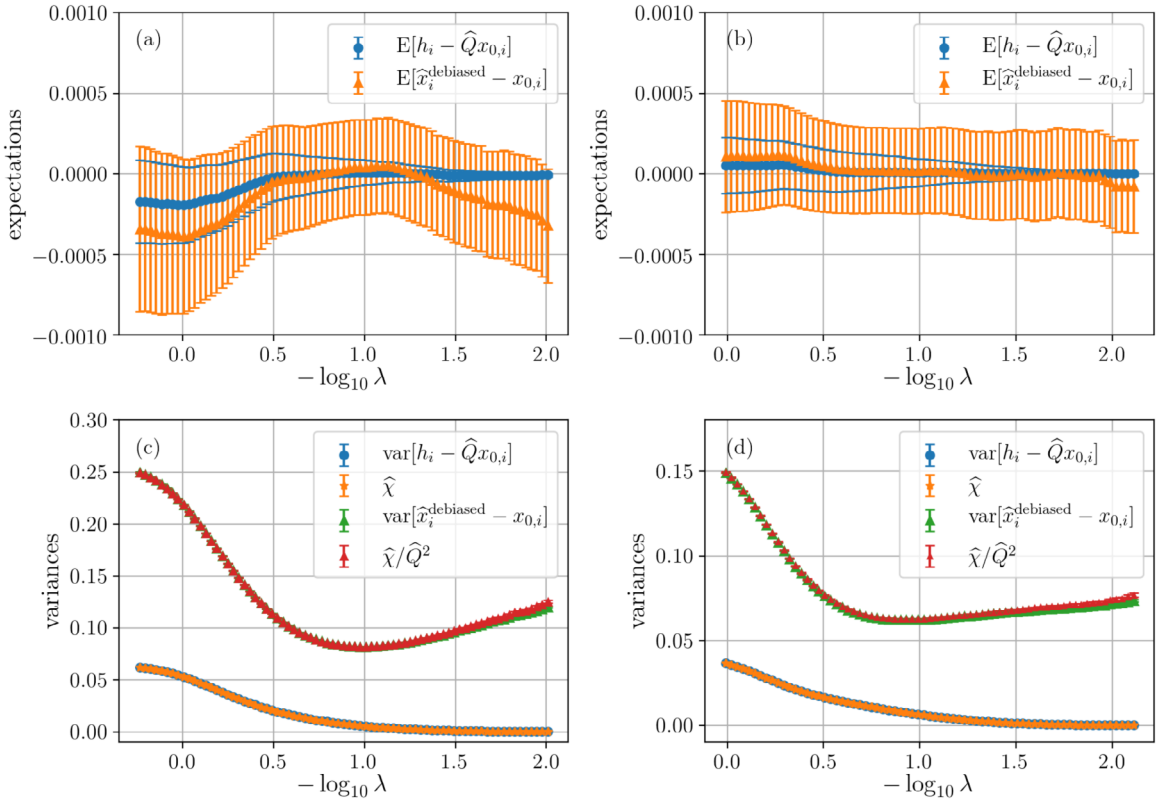

Figure 5. (a) and (b): mean of $\{h_i - \widehat{Q}x_{0,i}\}_i$ and $\{\widehat{x}^{\mathrm{debiased}} - x_{0,i}\}_i$. (c) and (d): comparison of the estimated and empirical values of the variances of $\{h_i - \widehat{Q}x_{0,i}\}_i$ and $\{\widehat{x}_i^{\mathrm{debiased}} - x_{0,i}\}_i$. The orange and red points represent the theoretically estimated values. The blue and green points stand for the empirical ones. (a) and (c) are the i.i.d. Gaussian case. (b) and (d) are the random DCT case.

J. Stat. Mech. (2018) 073405

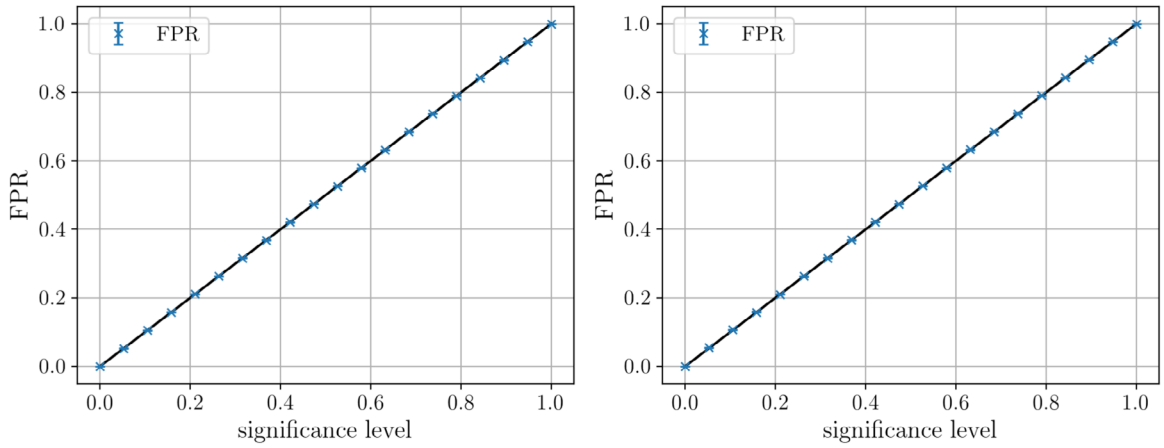

Figure 6. Significance level versus the observed false positive rate (FPR). The black solid line is the unit slope line. Left: the i.i.d. Gaussian case. Right: the random DCT case.

comparison purposes, we also plot the ROC curve for LASSO. For LASSO, the TPR and FPR are examined by changing the regularization parameter λ.

Figure 7 summarizes the results averaged over N_s configurations of $(A, \boldsymbol{\xi})$. It is observed that for some values of λ around which the variance of the de-biased estimator is minimized, our testing procedure performs slightly better than LASSO in the sense that the TPR of the testing method is slightly larger than that of LASSO's one for some values of the FPR. In the case of LASSO, when the measurement ratio γ is sufficiently small, the TPR and FPR do not coincide with $(1, 1)$ for finite $\lambda > 0$, as the consistency property does not hold in such a situation and the number of active components of the LASSO estimator is always smaller than $\min(N, M)$ [9]. On the contrary, as our hypothesis testing procedure always approaches the point $(1, 1)$ from $(0, 0)$, we can examine the TPR for all the values of the FPR $\in [0, 1]$. The superiority of the TPR comes from the fact that we are using the knowledge of the ensemble of the observation matrix. Further, as the hypothesis testing procedure controls the FPR and TPR by varying the significance α_{sig} but not λ, one does not suffer from the shrinkage effect in the variable selection procedure. This is another advantage over variable selection by LASSO. These observations show the utility of our hypothesis testing procedure.

4.2.3. Hyperparameter selection via confidence interval minimization. The issue of hyperparameter selection is noteworthy here. As LASSO has the hyperparameter λ that controls the strength of the regularization, one should choose a value of λ based on some criteria. As shown in figure 5, the estimated variance of the de-biased estimator $\widehat{\chi}/\widehat{Q}^2$ has a minimum value at some $\lambda > 0$. At this point, one can estimate $\boldsymbol{x}_0$ with the highest conviction in the sense that the confidence interval has the smallest width. It is therefore expected that the estimated variance of the de-biased estimator itself serves as a hyperparameter selection criterion. Indeed, in the i.i.d. Gaussian and row-orthogonal/random DCT cases, one can analytically show that minimizing the confidence interval is the equivalent to the minimization of the leave-one-out cross-validation error $\mathcal{C}$:

J. Stat. Mech. (2018) 073405

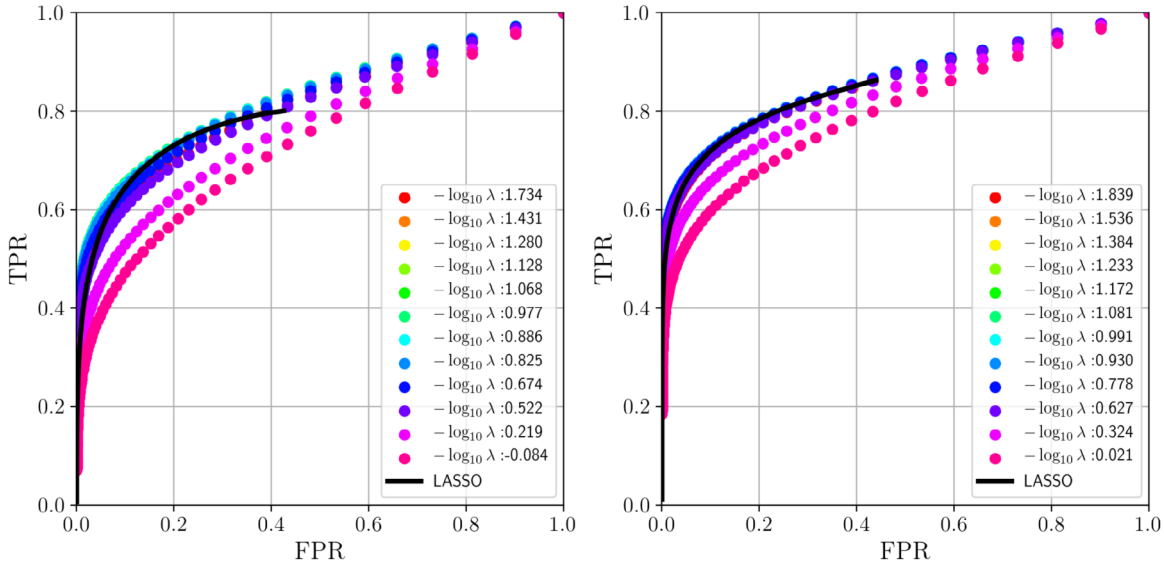

Figure 7. ROC (receiver operating characteristic) curves. The black solid line represents the ROC curve for LASSO obtained by varying the regularization strength λ. The points correspond to the ROC curve for the proposed hypothesis testing method. Left: the i.i.d. Gaussian case. Right: the random DCT case.

$$\frac{\widehat{\chi}}{\widehat{Q}^2} = \begin{cases} \frac{1}{\gamma}\mathcal{C} & \text{the i.i.d. Gaussian case,} \\ \frac{1-\gamma}{\gamma}\mathcal{C} + \sigma^2 & \text{the row-orthogonal or the random DCT cases.} \end{cases} \tag{71}$$

In other words, the leave-one-out cross-validation error and width of the confidence intervals are related with the linear transformation in these cases (figure 8).

Here, the leave-one-out cross-validation error is a widely used hyperparameter selection criterion that evaluates prediction performance, defined as follows:

$$\mathcal{C}(\boldsymbol{y}, A; \lambda) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\left\| y_i - \boldsymbol{a}_i^\top \widehat{\boldsymbol{x}}^{\mathrm{LASSO}}(\boldsymbol{y}_{\setminus i}, A_{\setminus i}; \lambda)\right\|_2^2, \tag{72}$$

where the symbol $\setminus i$ denotes the absence of the ith component (e.g. $\boldsymbol{a}_{\setminus i} = (a_1, ..., a_{i-1}, a_{i+1}, ..., a_N)^\top$) and each term in the summation evaluates the fitness to the ith data when the true signal is inferred from the other data. In the settings considered here, the above leave-one-out cross-validation error is expressed as follows [34, 35]:

$$\mathcal{C} = \left(1 - \frac{\varrho_{\mathrm{active}}}{\gamma}\right)^{-2}\mathrm{RSS} = \left(1 - \frac{2\chi G'(-\chi; J)}{\gamma}\right)^{-2}\mathrm{RSS}. \tag{73}$$

By substituting the expression of the leave-one-out cross-validation error (73) into equation (16), the relations (71) are obtained.

To investigate the validity of the above observation that the confidence interval minimization and leave-one-out cross-validation error minimization provide the same λ, we test the geometric setup case in which $\widehat{\chi}/\widehat{Q}^2$ is not expressed as a linear function of $\mathcal{C}$. Figure 9 compares the variance of $\{\widehat{x}^{\mathrm{debiased}} - x_{0,i}\}_i$ with the leave-one-out

J. Stat. Mech. (2018) 073405

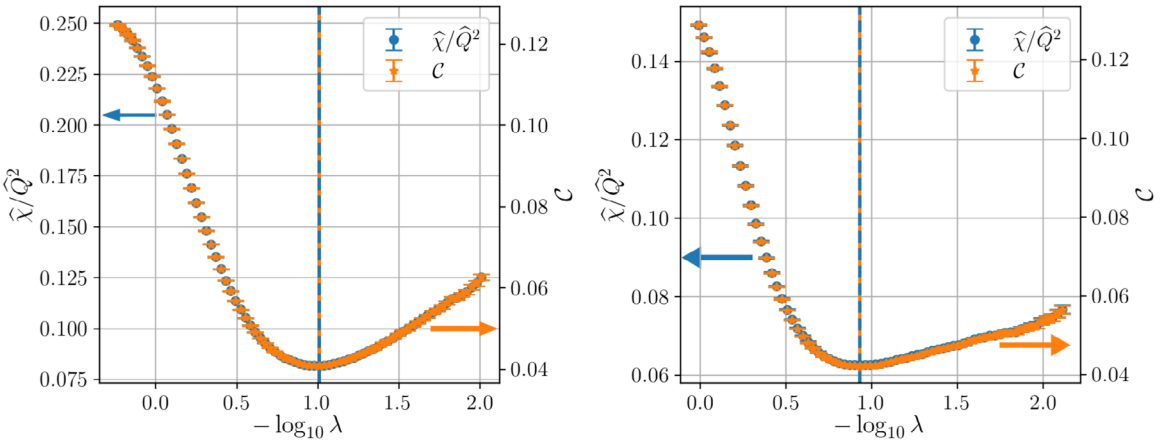

Figure 8. Comparison of the width of the confidence interval versus the leave-one-out cross-validation error. The blue and orange points show the average width of the confidence interval and leave-one-out cross-validation error, respectively. The blue solid line and orange dashed line indicate the value of λ that minimizes the confidence interval and leave-one-out cross-validation error, respectively. The left and right vertical axes represent the values of $\widehat{\chi}/\widehat{Q}^2$ and $\mathcal{C}$, respectively. The axis range for $\mathcal{C}$ is chosen according to equation (71) so that the curves of $\widehat{\chi}/\widehat{Q}^2$ and $\mathcal{C}$ overlap. The values of λ that minimize the width of the confidence interval and cross-validation error perfectly coincide as expected. Left: the i.i.d. Gaussian case. Right: the random DCT case.

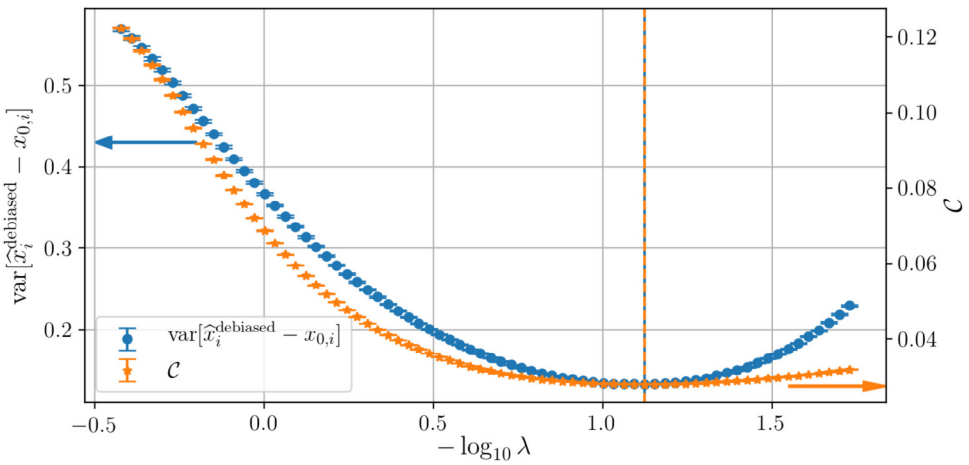

Figure 9. Comparison of the width of the confidence interval versus the leave-one-out cross-validation error for the geometric case. Here, $\widehat{\chi}/\widehat{Q}^2$ is evaluated by $\mathrm{var}[\widehat{x}_i^{\mathrm{debiased}} - x_{0,i}]$. The blue and orange points show the average width of the confidence interval and leave-one-out cross-validation error, respectively. The blue solid line and orange dashed line indicate the value of λ that minimizes the confidence interval and leave-one-out cross-validation error, respectively. The left and right vertical axes represent the values of $\mathrm{var}[\widehat{x}_i^{\mathrm{debiased}} - x_{0,i}]$ and $\mathcal{C}$, respectively. Unexpectedly, the values of λ that minimize the width of the confidence interval and cross-validation error perfectly coincide.

J. Stat. Mech. (2018) 073405

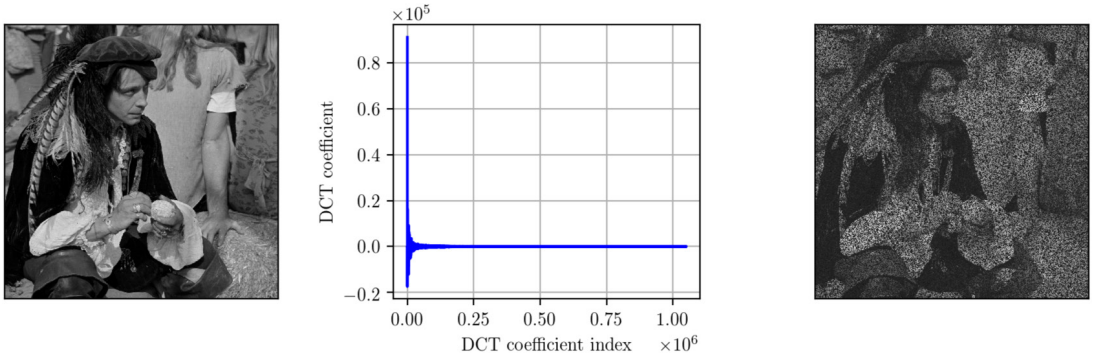

Figure 10. Left: original megapixel grayscale image. Center: its discrete cosine transform (DCT) coefficients. Most of the coefficients are relatively small and hence can be estimated by sparse linear regression. Right: given data. Half of the pixels are masked and the other half are degraded by Gaussian noises.

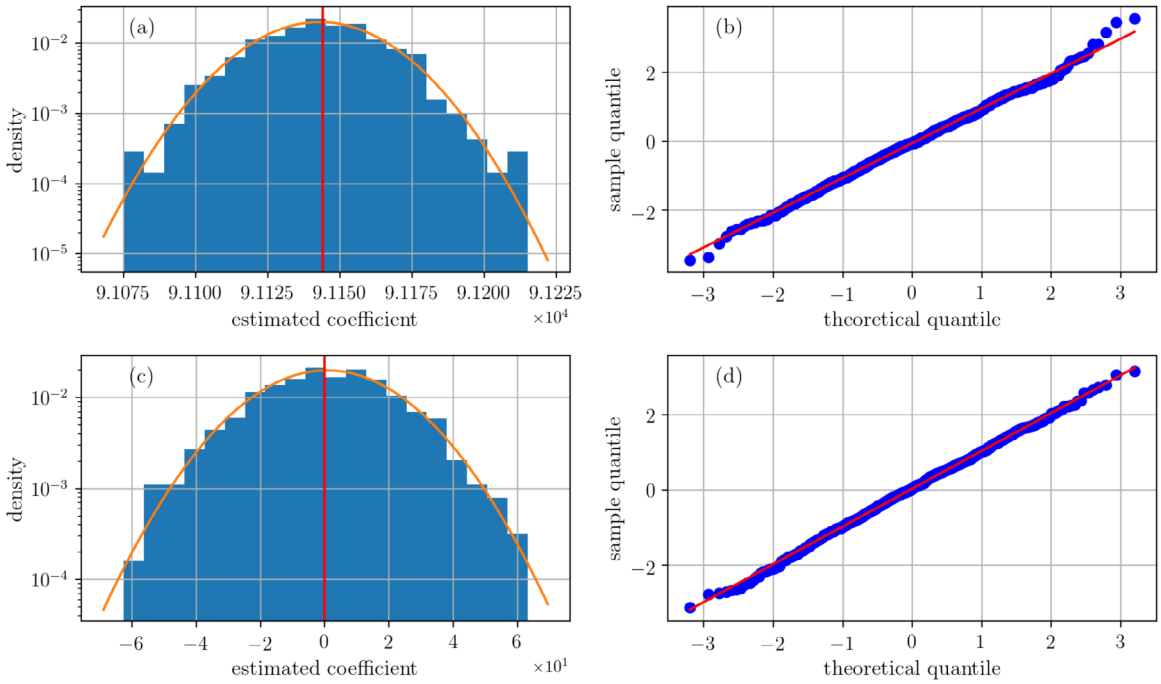

Figure 11. Distribution of some components of the de-biased estimators. The strength of L_1 norm was set to $\lambda = 12.0$. (a) and (c) show the distributions of de-biased estimatosr for the largest and smallest amplitude DCT coefficients, respectively. The red lines show the true values of the DCT coefficients. The orange curves show Gaussians predicted by the developed method. (b) and (d) show the Q-Q plot of $(\widehat{x}_i^{\mathrm{debiased}} - x_{0,i})/\sqrt{\widehat{\chi}}$ for the largest and smallest amplitude DCT coefficient, respectively. The red lines are the unit slope zero-intercept lines.

cross-validation error (73). Surprisingly, the minimization of these two quantities seems to provide the same value of λ, although they do not have a functional relation as (71).

From the above observations, we speculate that the minimization of the confidence interval proposed here and the minimization of the leave-one-out cross-validation error yields the same value of λ for LASSO in general rotationally invariant observation matrices, but further investigation in this direction is still needed.

J. Stat. Mech. (2018) 073405

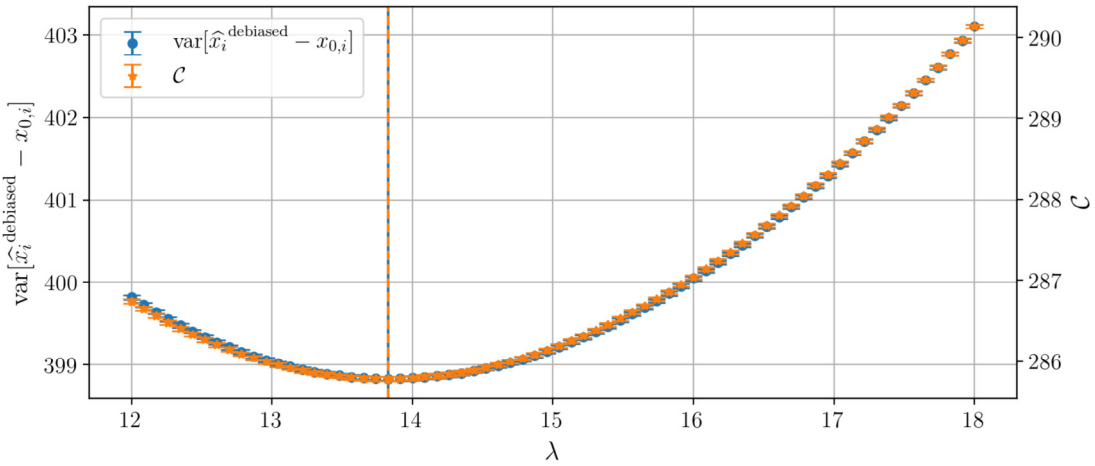

Figure 12. Comparison of the variance of the de-biased estimators versus the leave-one-out cross-validation error. The blue and orange points represents the average variance of the de-biased estimators and leave-one-out cross-validation error, respectively. The blue solid line and orange dashed line indicate the value of λ that minimizes the variance of the de-biased estimators and leave-one-out cross-validation error, respectively. As expected from the property of the measurement matrix, these two quantities are related with linear transformation.

4.3. Demonstration on a real-world data set

We illustrate the practical relevance of the proposed method by application to a problem of inferring Fourier components from down-sized and noisy measurements of real space signals. Although such demands widely arise in various Fourier analyses, we here show a demonstration using an image for ease of visual understanding. The real space signals consist of a partially observed megapixel gray-scale image data $\boldsymbol{y} \in \mathbb{R}^{1024\times1024}$ (figure 10 left). Randomly chosen half of its elements are masked and the other half are degraded by Gaussian noises, variance of which is set to 1% of the average power of the original signal (figure 10 right). Given the data, the problem is to estimate the DCT coefficients $\boldsymbol{x}$ (figure 10 center) of the original image from the observed noisy pixels. As the measurement process can be written as linear measurement model using a partial DCT matrix, the proposed scheme is directly applicable to this inference problem. To investigate the statistical properties of the de-biased estimators, we created 1000 realizations of the random masking and Gaussian noises.

The results are similar to that for synthetic data set in subsection 4.2. Figure 11 shows the distributions of the de-biased estimators for some DCT coefficients. It is verified that the de-biased estimators are normally distributed around the true coefficients. Figure 12 compares the variance of $\{\widehat{x}_i^{\mathrm{debiased}} - x_{0,i}\}_i$ and leave-one-out cross-validation error. As expected from the properties of partial DCT matrices, these two quantities are related with linear transformation. These results imply that the proposed method is of practical relevance for signal processing problems handling real-world data set.

J. Stat. Mech. (2018) 073405

5. Summary

We developed a new computationally feasible scheme for de-biasing and uncertainty estimation in LASSO in the case of rotationally invariant observation matrix ensembles and validated the proposed scheme by using numerical experiments. We focused on the development of a de-biased estimator that has a confidence interval and hypothesis testing scheme for the null hypothesis that a certain parameter vanishes. The numerical experiments showed that the proposed method efficiently constructed de-biased estimators with confidence intervals and p-values for the intended hypothesis testing. We revealed that the proposed hypothesis testing slightly improved the variable selection performance in the sense that the TPR of the testing method achieves a slightly larger value than that of the LASSO's one for some values of the FPR. Further, we examined the utility of the estimator of the confidence interval as a criterion for determining the hyperparameter. Surprisingly, minimizing the width of the confidence interval was equivalent to the minimization of the leave-one-out cross-validation error in our investigation.

Although we only focused on LASSO for linear models, future work could include an extension to other sparse regression methods such as the elastic net [36] as well as generalized linear models.

Acknowledgment

Support by JSPS KAKENHI Nos. 25120013 and 17H00764 (YK) is acknowledged.

Appendix. Derivation of the free energy density

To take the average that appears in (4), we use the replica method [33] based on the identity for $n \in \mathbb{R}$:

$$f = -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{\beta N}\lim_{n\to 0}\frac{\mathbb{E}\left[Z^n\right]_{A,\boldsymbol{\xi}}}{n}. \tag{A.1}$$

In the replica method, we first take the average of the nth power of the partition function over the randomness of $A, \boldsymbol{\xi}$ for the positive integer $n \in \mathbb{N}$, and then analytically continue the obtained expression to real $n \in \mathbb{R}$ to take the limit $n \to 0$, exchanging the order of the limits.

For the general matrix ensembles considered here, it is convenient to first take the average over $\boldsymbol{\xi}$. By taking this average, we obtain the following expression under the replica symmetric ansatz:

$$\mathbb{E}\left[Z^n\right]_{A,\boldsymbol{\xi}} = \int \mathbb{E}\left[\exp\left(\frac{1}{2}\mathrm{Tr}JL\right)\right]_A e^S \mathrm{d}Q\mathrm{d}q\mathrm{d}m, \tag{A.2}$$

where L, $\boldsymbol{u}_a$, and S are defined as follows:

$$L \equiv \frac{\beta^2\sigma^2}{1+\beta n\sigma^2}\left(\sum_a \boldsymbol{u}_a\right)\left(\sum_a \boldsymbol{u}_a\right)^\top - \beta\sum_a \boldsymbol{u}_a\boldsymbol{u}_a^\top, \tag{A.3}$$

$$\boldsymbol{u}_a \equiv \boldsymbol{x}_a - \boldsymbol{x}_0, \tag{A.4}$$

$$e^S \equiv \int\prod_{a=1}^{n}\delta(NQ-\boldsymbol{x}_a^\top\boldsymbol{x}_a)\delta(Nm-\boldsymbol{x}_a^\top\boldsymbol{x}_0) \times \prod_{1\leqslant a<b\leqslant n}\delta(Nq-\boldsymbol{x}_a^\top\boldsymbol{x}_b)\exp\left\{-\frac{N\gamma}{2}\beta n\sigma^2-\beta\lambda\sum_a\|\boldsymbol{x}_a\|_1\right\}\mathrm{d}\boldsymbol{x}, \tag{A.5}$$

where $\boldsymbol{u}_a$ and $\boldsymbol{x}_a$ are the ath replica's variable. In [37], it was shown that under the rotational invariance assumption on the random matrix $J = A^\top A$ for eigenvalue decomposition $J = ODO^\top$ considered in this study, the average over A that appears in equation (A.2) is evaluated by using the eigenvalues $\{s_i\}_i$ of L/N for sufficiently large N:

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\mathrm{Tr}JL\right)\right]_A = \exp\left\{N\sum_i G(s_i;J)\right\}, \tag{A.6}$$

where $G(x;J)$ is the function defined in (9). Under the replica symmetric ansatz, L/N has three types of eigenvalues: $s_1 = \beta\Delta Q - \beta n(q-2m+\varrho) + n\beta^2\sigma^2\Delta Q$, $s_2 = -\beta\Delta Q$, and $s_3 = 0$. The number of degeneracy is $1, n-1$, and $N-n$, respectively. Thus, we obtain the following expression up to the leading order in n:

$$\mathbb{E}\left[\mathrm{e}^{\frac{1}{2}\mathrm{Tr}JL}\right]_A = \exp\left[-Nn\beta\left\{-G(-\beta\Delta Q;J)/\beta + G'(-\beta\Delta Q;J)(q-2m+\varrho-\beta\Delta Q\sigma^2)\right\}\right]. \tag{A.7}$$

On the contrary, by using the Fourier transform of the delta function and Hubbard–Stratonovich transform: $\mathrm{e}^{B^2/2A} = \int \mathrm{e}^{-Ax^2/2+Bx}\sqrt{\frac{A}{2\pi}}dx$ for $A>0, B\in\mathbb{R}$, the factor e^S is given as follows:

$$\mathrm{e}^S = \int\exp\left[Nn\left\{\frac{\gamma\sigma^2}{2}+\frac{q\widetilde{q}}{2}+\frac{Q\widetilde{Q}}{2}-m\widetilde{m} + \frac{1}{N}\sum_{i=1}^{N}\int\ln\phi(x_{0,i},z_i,\widetilde{Q},\widetilde{q},\widetilde{m};\beta,\lambda)Dz_i\right\}\right]\mathrm{d}\widetilde{Q}\mathrm{d}\widetilde{q}\mathrm{d}\widetilde{m}, \tag{A.8}$$

$$\phi(x_{0,i},z_i,\widetilde{Q},\widetilde{q},\widetilde{m};\beta,\lambda) = \int\exp\left\{-\frac{\widetilde{Q}+\widetilde{q}}{2}x_i^2+(\widetilde{m}x_{0,i}+\sqrt{\widetilde{q}}z_i)x_i-\beta\lambda|x_i|\right\}\mathrm{d}x_i. \tag{A.9}$$

For $\beta\to\infty$, the relevant variables scale as $\beta(Q-q)=\chi, \widetilde{Q}+\widetilde{q}=\beta\widehat{Q}, \widetilde{m}=\beta\widehat{m}$, and $\widetilde{q}=\beta^2\widehat{\chi}$ of order unity to ensure an appropriate limit f exists. Finally, by combining equations (A.7)–(A.9) and evaluating the integrals by adopting the saddle point method, we obtain equation (6) for $\beta, N\to\infty$.

J. Stat. Mech. (2018) 073405

References

[1] Donoho D L 2006 Compressed sensing *IEEE Trans. Inf. Theory* **52** 1289
[2] Candes E J and Tao T 2006 Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* **52** 5406
[3] Peng J, Zhu J, Bergamaschi A, Han W, Noh J-Y, Pollack J R and Wang P 2010 Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer *Ann. Appl. Stat.* **4** 53–77
[4] Thibshirani R 1996 Regression shrinkage and selection via the Lasso *J. R. Stat. Soc.* B **58** 267–288
[5] Candes E J and Tao T 2005 Decoding by linear programming *IEEE Trans. Inf. Theory* **51** 4203
[6] Bickej P J, Yitov R and Tsybakov A B 2009 Simultaneous analysis of Lasso and Dantzig selector *Ann. Stat.* **37** 1705–32
[7] Donoho D L, Maleki A and Montanari A 2009 Message-passing algorithms for compressed sensing *Proc. Natl. Acad. Sci.* **106** 18914
[8] Kabashima Y, Wadayama T and Tanaka T 2009 A typical reconstruction limit for compressed sensing based on L_p-norm minimization *J. Stat. Mech.* L09003
[9] Geer S V and Bühlmann P 2011 *Statistics for High-Dimensional Data: Methods, Theory and Applications* (*Springer Series in Statistics*) (Berlin: Springer)
[10] Opper M and Winther O 2001 Tractable approximations for probabilistic models: the adaptive Thouless–Anderson–Palmer mean field approach *Phys. Rev. Lett* **86** 3695
[11] Opper M and Winther O 2001 Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling *Phy. Rev.* E **64** 056131
[12] Opper M and Winther O 2005 Expectation consistent approximate inference *J. Mach. Learn. Resear.* **6** 2177
[13] Minka P T 2001 Expectation propagation for approximate Bayesian inference *Proc. UAI-2001* p 362
[14] Javanmard A and Montanari A 2014 Confidence intervals and hypothesis testing for high-dimensional regression *J. Mach. Learn. Res.* **15** 2869–2909
[15] Geer S V, Bühlmann P, Ritov Y and Dezeure R 2014 On asymptotically optimal confidence regions and tests for high-dimensional models *Ann. Stat.* **42** 1166–202
[16] Zhang C and Zhang S 2014 Confidence intervals for low dimensional parameters in high dimensional linear models *J. R. Stat. Soc. Ser.* B **76** 217–42
[17] Voiculescu D 1986 Addition of certain non-commuting random variables *J. Funct. Anal.* **66** 323
[18] Guo D and Verdú S 2006 Randomly spread CDMA: asymptotics via statistical physics *IEEE Trans. Inf. Theory* **51** 1983
[19] Rangan S, Fletcher A K and Goyal V K 2009 Asymptotics analysis of MAP estimation via the replica method and compressed sensing *Advances in Neural Information Processing Systems 22* (Red Hook, NY: Curran Associates, Inc.) pp 1545–53
[20] Plefka T 1982 Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model *J. Phys. A: Math. Gen.* **15** 1971–8
[21] Opper M and Saad D (ed) 2001 *Advanced Mean-field Methods: Theory and Practice* (Cambridge, MA: MIT Press)
[22] Kabashima Y and Vehkaperä M 2014 Signal recovery using expectation consistent approximation for linear observations *Proc. IEEE Int. Symp. on Information Theory* pp 226–230
[23] Reid S, Tibshirani R and Friedman J 2013 A study of error variance estimation in Lasso regression (arXiv:1311.5274)
[24] Çakmak B and Opper M 2018 Expectation propagation for approximate inference: free probability framework (arXiv:1801.05411)
[25] Efron B *et al* 2004 Least angle regression *Ann. Stat.* **32** 407–99
[26] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent *J. Stat. Softw.* **33** 1
[27] Rangan S 2011 Generalized approximate message passing for estimation with random linear mixing *Proc. IEEE Int. Symp. on Information Theory* (IEEE)
[28] Ma J and Ping L 2017 Orthogonal AMP *IEEE Access* **5** 2020–2033
[29] Rangan S, Scniter P and Fletcher A K 2017 Vector approximate message passing *Proc. IEEE Int. Symp. on Information Theory* (IEEE) pp 1588–1592
[30] Marenko V A and Pastur L A 1967 Distribution of eigenvalues for some sets of random matrices *Sb. Math.* **1** 457
[31] Vehkaperä M, Kabashima Y and Chatterjee S 2016 Analysis of regularized LS reconstruction and random matrix ensembles in compressed sensing *IEEE Trans. Inf. Theory* **62** 2100–24

J. Stat. Mech. (2018) 073405

[32] Rangan S *et al* 2017 Inference for generalized linear models via alternating directions and Bethe free energy minimization *IEEE Trans. Inf. Theory* **63** 676–97
[33] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond: an Introduction to the Replica Method and its Applications* (Singapore: World Scientific)
[34] Obuchi T and Kabashima Y 2016 Cross validation in LASSO and its acceleration *J. Stat. Mech.* 053304
[35] Rad K R and Maleki A 2018 A scalable estimate of the extra-sample prediction error via approximate leave-one-out (arXiv:1801.10243)
[36] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301
[37] Marinari E, Parisi G and Ritort F 1994 Replica field theory for deterministic models: II. A non-random spin glass with glassy behavior *J. Phys. A: Math. Gen* **27** 7647

J. Stat. Mech. (2018) 073405