Two-sample testing in high dimensions

Nicolas Städler
Netherlands Cancer Institute, Amsterdam, The Netherlands

and Sach Mukherjee
German Centre for Neurodegenerative Diseases, Bonn, Germany

[Received January 2013. Final revision November 2015]

Summary. We propose new methodology for two-sample testing in high dimensional models. The methodology provides a high dimensional analogue to the classical likelihood ratio test and is applicable to essentially any model class where sparse estimation is feasible. Sparse structure is used in the construction of the test statistic. In the general case, testing then involves non-nested model comparison, and we provide asymptotic results for the high dimensional setting. We put forward computationally efficient procedures based on data splitting, including a variant of the permutation test that exploits sparse structure. We illustrate the general approach in two-sample comparisons of high dimensional regression models (‘differential regression’) and graphical models (‘differential network’), showing results on simulated data as well as data from two recent cancer studies.

Keywords: Differential network; Differential regression; Gaussian graphical models; High dimensional regression; High dimensional two-sample testing; Non-nested hypotheses; Restricted log-likelihood-ratio statistic; Sparsity

1. Introduction

We consider two-sample testing in high dimensional settings. Our work is motivated by current applied problems where interest focuses on comparing high dimensional data between conditions (we consider examples from biomedicine below). Consider a model class $f$ with parameter $\theta$. Given data from two conditions (or groups; we use both terms interchangeably), assumed to be distributed according to $f$ with (unknown) $p$-dimensional parameters $\theta^{(1)}$ and $\theta^{(2)}$, we seek to test the hypothesis

$$H_0 : \theta^{(1)} = \theta^{(2)} \quad \text{against} \quad H_A : \theta^{(1)} \neq \theta^{(2)}. \quad (1)$$

In the low dimensional case, this is a classical problem and the likelihood ratio test (LRT) offers a general solution. However, the classical LRT breaks down in high dimensions. Permutation tests offer an alternative, but in precisely the high dimensional settings of interest the computational demands of multiple rounds of estimation can render such tests burdensome—often prohibitively so—and furthermore it may be non-obvious how a suitable test statistic should be constructed. A number of specific tests have been developed for high dimensional data (which are reviewed below), but a general approach to solving the testing problem (1) in high dimensions has remained lacking.

Address for correspondence: Nicolas Städler, Pharmaceutical Division, Global Pricing and Market Access, F. Hoffmann-La Roche Ltd, CH 4070 Basel, Switzerland.
E-mail: staedler.nn@gmail.com

© 2016 Royal Statistical Society
The aim of this paper is to address the general two-sample problem in high dimensions. For this we put forward a high dimensional analogue to the classical LRT. We focus on settings that are in some sense sparse, allowing \( f \) to represent a model class where sparse estimation is possible. Our methodology is motivated by the twin observations that sparse structure can be useful for effective and efficient inference and in general such structure must be estimated from the data. These observations suggest a data adaptive approach. The main idea in our methodology is to exploit estimated sparsity patterns in the construction of a test statistic and in \( p \)-value calculation. This is an intuitive idea, but it leads to non-nested testing that requires careful treatment and new results for the high dimensional setting. The two key elements of the methodology are as follows.

(a) **Construction of a data-driven restricted log-likelihood-ratio statistic**: estimated sparsity patterns are used in the construction of the test statistic. This renders testing feasible by restricting attention to parameter subspaces.

(b) **Computation of \( p \)-values**: use of sparse structure in the construction of the test statistic turns out to lead to non-nested model comparison (because in general the resulting parameter subspaces need not be nested). We obtain a null distribution for the proposed statistic via new results that extend the seminal work of Vuong (1989) on non-nested testing to the high dimensional setting where the model dimension grows with sample size. Furthermore, we show how required computations can be carried out efficiently. In addition, inspired by an insightful suggestion from a referee, we introduce a variant of the permutation test that exploits sparsity to improve performance relative to a standard permutation test.

We embed these ideas within a data splitting approach that gives a simple and practically applicable procedure. The data are partitioned into two halves with the first half of the data used to estimate sparsity patterns and the second half used to calculate \( p \)-values. \( p \)-values from multiple, random data splits are aggregated to give a final result. Our use of data splitting is inspired by previous work in variable selection (Wasserman and Roeder, 2009; Meinshausen et al., 2009).

Estimation for high dimensional models has been widely studied but many aspects of quantifying uncertainty remain open. Our results bridge the gap between high dimensional estimation and testing by showing how sparse estimators can be leveraged for testing. The test is applicable to any model class where sparse estimation is possible, and essentially any available estimator can be used in the construction of the statistic (provided that it can reasonably be supposed to satisfy the sparsity and screening assumptions that we discuss below). Given the rich literature in sparse estimation, by now spanning many model classes, this is appealing because it means that the test can be readily instantiated in these settings. In addition, taking advantage of sparse structure leads to a computationally efficient procedure. Furthermore, the approaches that we put forward require no tuning parameters beyond those required for the sparse estimators themselves.

Recent years have seen increasing interest in high dimensional testing. Some relevant references include Meinshausen et al. (2009), Bühlmann (2013) and Zhang and Zhang (2014) on testing in the high dimensional regression model, Bai and Saranadasa (1996), Chen and Qin (2010), Lopes et al. (2011), Chang et al. (2014) and Thulin (2014) on testing high dimensional means, Cai et al. (2013) and Li and Chen (2012), who developed two-sample tests for high dimensional covariance matrices, and Rosenbaum (2005), Gretton et al. (2012), Wei et al. (2015) and Biswas and Ghosh (2014), who proposed non-parametric two-sample tests in the medium to large dimensional setting. During revision of this paper, we learned of the work of Charbonnier...
et al. (2015) who proposed a homogeneity test for high dimensional linear regression. However, their method is customized to linear regression and is less general than our test.

We illustrate the methodology in two specific settings for which we provide detailed instantiations of the general approach.

(a) High dimensional regression where model $f$ is the usual linear regression model and the parameter $\theta$ comprises the coefficients $\beta$ and noise variance $\sigma^2$: here, the idea is that the two conditions may differ with respect to regression models.

(b) Gaussian graphical modelling where $f$ is a zero-mean multivariate normal distribution with $\theta$ denoting the concentration matrix. Here, the two conditions may differ with respect to concentration matrices and the associated graphical models.

By analogy with the term ‘differential expression’ as widely used for testing means in gene expression studies, we call these problems differential regression and differential network respectively. We emphasize that the methodology is not restricted to these models, but rather offers a general, likelihood-based framework. We note also that, although in both examples we use $l_1$-penalization to estimate sparse structure, $l_1$-penalization per se is not central to our approach.

Biomedical applications provided the original motivation for our work. Scientific interest often focuses on potential differences between conditions or groups (such as disease types, cell types and environmental conditions). As an example we consider below protein data from cancer samples (from The Cancer Genome Atlas ‘pan-cancer’ study; see Akbani et al. (2014)). Interplay between molecules plays a central role in biology and an emerging hypothesis is that patterns of interplay may differ between groups. However, it has remained challenging to quantify the significance of such differences and more generally of differences between group-specific multivariate models. Our work provides a general and principled framework for problems of this type.

The R package nethet implements our methods (specifically differential regression and differential network) and is available as part of the ‘Bioconductor’ project at http://www.bioconductor.org/packages/nethet.

The data and the programs that were used in the paper can be obtained from http://wileyonlinelibrary.com/journal/rss-datasets

2. Methodology

2.1. Preliminaries

Consider a model class with density function

$$f(y|x; \theta), \quad y \in \mathbb{R}^k, \quad x \in \mathbb{R}^l, \quad z = (y, x) \quad \text{and} \quad \theta \in \Theta \subset \mathbb{R}^p,$$

where $l > 0$ gives a conditional model that allows for explanatory variables $x$, $l = 0$ gives a marginal model (for $y$) and $\theta$ is understood to include all model parameters.

Let $Z_i^{(r)} = (Y_i^{(r)}, X_i^{(r)}) \in \mathbb{R}^{k+l}$, $i = 1, \ldots, n_r$, denote independently and identically distributed (IID) samples following model $f$, with $r \in \{1, 2\}$ indexing the groups under comparison and $n_r$ being the group-specific sample size. The group-specific data sets $Z^{(1)}$ and $Z^{(2)}$ are taken to be drawn independently with (unknown) parameters $\theta^{(1)}$ and $\theta^{(2)}$ respectively.

2.2. The procedure in outline

The goal is to address the two-sample problem (1) in the high dimensional setting. In outline, the procedure is as follows.
(a) **Data splitting:** data $Z^{(1)}$ are randomly divided into two parts $Z^{(1)}_{in}$ and $Z^{(1)}_{out}$ of equal size and similarly data $Z^{(2)}$ into halves $Z^{(2)}_{in}$ and $Z^{(2)}_{out}$. Throughout, the subscripts ‘in’ and ‘out’ will indicate first- and second-half data respectively.

(b) **Screening:** a screening procedure $\hat{I}$ is applied to data to estimate an active set $\hat{I}(\cdot) \subseteq \{1 \ldots p\}$ that induces a parameter subspace in the sense of

$$\Theta_j = \{(\theta_1, \ldots, \theta_p) \in \Theta : \theta_j = 0 \text{ for all } j \notin \hat{I}\}.$$ 

Procedure $\hat{I}(\cdot)$ is applied to $Z^{(1)}_{in}$ and $Z^{(2)}_{in}$ separately, but also to pooled data $Z^{(12)}_{in} = (Z^{(1)}_{in}, Z^{(2)}_{in})$, yielding corresponding estimated active sets $\hat{I}^{(1)}_{in}, \hat{I}^{(2)}_{in}$ and $\hat{I}^{(12)}_{in}$.

(c) **p-value calculation:** p-values are calculated by using the second half of the data, using sparsity patterns encoded in the active sets obtained by using the first half of the data. Specifically, active sets $\hat{I}^{(1)}_{in}$ and $\hat{I}^{(2)}_{in}$ give constraints for each group individually and are used to obtain an *individual model* with log-likelihood $L^{\text{ind}}$, whereas the pooled active set $\hat{I}^{(12)}_{in}$ gives a *joint model* with log-likelihood $L^{\text{joint}}$. The two models are compared by using the restricted log-likelihood-ratio statistic $LR = 2(L^{\text{ind}} - L^{\text{joint}})$. Importantly, provided that the active sets are not large, p-value calculation using LR becomes a low dimensional problem that, as we discuss below, is amenable to asymptotic analysis, or to a computationally efficient permutation-based approach.

### 2.3. Screening

The screening procedure $\hat{I}$ is applied to group-specific data $Z^{(1)}_{in}$ and $Z^{(2)}_{in}$ separately, and also to pooled data $Z^{(12)}_{in}$, to obtain estimated active sets $\hat{I}^{(1)}_{in}, \hat{I}^{(2)}_{in}$ and $\hat{I}^{(12)}_{in}$. For notational simplicity, in this section we omit the superscript and use $Z_{in} = (Y_{in}, X_{in})$ to refer generically to the first-half data sets (i.e. $Z_{in}$ is one of $Z^{(1)}_{in}, Z^{(2)}_{in}$ or $Z^{(12)}_{in}$) and $\hat{I}$ to refer to the corresponding estimated active set. The sample size is $n_{in}$. We consider the high dimensional setting, $p > n_{in}$, where the model is assumed to be sparse with respect to parameter $\theta$, with non-zero components of $\theta$ indexed by a (true, unknown) active set $I \subset \{1 \ldots p\}$. The estimated active set $\hat{I}$ is taken to satisfy the following assumptions:

(a) $|\hat{I}(Z_{in})|$ is small compared with $n_{in}$ (sparsity assumption);

(b) $\lim_{n_{in} \to \infty} \mathbb{P}\{I \subset \hat{I}(Z_{in})\} = 1$ (screening assumption).

Satisfying the sparsity assumption ensures that the procedure $\hat{I}$ reduces dimensionality to a sufficient extent. Satisfying the screening assumption ensures that all non-zero parameter components are retained. Any screening procedure that satisfies these could be used within our framework. In examples below, we use $l_1$-penalized estimators of the form

$$\hat{\theta}_\lambda = \arg\max_{\theta \in \Theta} L(\theta; Y_{in}|X_{in}) - \lambda \|\theta\|_1,$$

(3)

where $L$ denotes the conditional log-likelihood, $\| \cdot \|_1$ the $l_1$-norm of its argument and $\lambda$ is a non-negative regularization parameter. Setting $I^{\lambda}(Z_{in}) = \{j : \hat{\theta}^{\lambda}_{j} \neq 0\}$ defines a screening procedure. For linear regression estimator (3) is the lasso (Tibshirani, 1996). $l_1$-regularization has been shown to be useful for variable screening (Wasserman and Roeder, 2009; Meinshausen et al., 2009). Indeed, for the lasso, screening requires milder assumptions than exact recovery of the active set and may be a more appropriate goal (Bühlmann, 2013).

### 2.4. Constructing the restricted log-likelihood-ratio statistic

Screening using $\hat{I}$ applied to $Z^{(1)}_{in}$ and $Z^{(2)}_{in}$ separately gives individual active sets $\hat{I}^{(1)}_{in}$ and $\hat{I}^{(2)}_{in}$
and applying \( \tilde{I} \) to the pooled data \( Z_{in}^{(12)} = (Z_{in}^{(1)}, Z_{in}^{(2)}) \) gives a joint active set \( \tilde{I}_{in}^{(12)} \). These then define active parameter spaces \( \Theta_{\tilde{I}_{in}^{(1)}} \) and \( \Theta_{\tilde{I}_{in}^{(2)}} \). The resulting individual model is

\[
M_{\text{ind}}^{\text{ind}} = \{ f(y^{(1)}|x^{(1)}; \theta^{(1)}) f(y^{(2)}|x^{(2)}; \theta^{(2)}); (\theta^{(1)}, \theta^{(2)}) \in \Theta_{\tilde{I}_{in}^{(1)}} \times \Theta_{\tilde{I}_{in}^{(2)}} \}
\]

and the joint model \( M_{\text{joint}} \) is

\[
M_{\text{joint}} = \{ f(y^{(1)}|x^{(1)}; \theta^{(12)}) f(y^{(2)}|x^{(2)}; \theta^{(12)}); \theta^{(12)} \in \tilde{I}_{in}^{(12)} \}.
\]

Given second-half data \( Z_{out}^{(1)} \) and \( Z_{out}^{(2)} \) with sample sizes \( n_{1,\text{out}} \) and \( n_{2,\text{out}} \) respectively, the log-likelihood functions with respect to \( M_{\text{ind}}^{\text{ind}} \) and \( M_{\text{joint}} \) are

\[
L_{\text{ind}}(\theta^{(1)}, \theta^{(2)}) = \sum_{i=1}^{n_{1,\text{out}}} \log \{ f(Y_{out,i}^{(1)}|X_{out,i}^{(1)}; \theta^{(1)}) \} + \sum_{i=1}^{n_{2,\text{out}}} \log \{ f(Y_{out,i}^{(2)}|X_{out,i}^{(2)}; \theta^{(2)}) \}
\]

and

\[
L_{\text{joint}}(\theta^{(12)}) = \sum_{i=1}^{n_{1,\text{out}}} \log \{ f(Y_{out,i}^{(1)}|X_{out,i}^{(1)}; \theta^{(12)}) \} + \sum_{i=1}^{n_{2,\text{out}}} \log \{ f(Y_{out,i}^{(2)}|X_{out,i}^{(2)}; \theta^{(12)}) \}.
\]

We test hypothesis (1) on the basis of the restricted log-likelihood-ratio statistic

\[
LR = 2 \{ L_{\text{ind}}(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) - L_{\text{joint}}(\hat{\theta}^{(12)}) \},
\]

where \( (\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) \) and \( \hat{\theta}^{(12)} \) denote maximum likelihood estimates obtained under constraints imposed by the corresponding active sets. In this way, sparsity patterns estimated from the first half of the data are used in the construction of the test statistic.

### 2.5. Computation of p-values

We put forward two different routes to testing using statistic (4): the first is an asymptotic approach and the second is permutation based.

#### 2.5.1. Asymptotic approach (‘split-asym’)

Statistic (4) compares the models \( M_{\text{ind}}^{\text{ind}} \) and \( M_{\text{joint}} \). The two models are in general non-nested as the screening procedure applied to pooled data will select components of \( \theta \) that are different from those obtained via screening performed individually on data from each condition. In other words, typically \( \tilde{I}_{in}^{(12)} \not\subset \tilde{I}_{in}^{(1)} \cap \tilde{I}_{in}^{(2)} \). Non-nestedness renders the null distribution not easily accessible, as discussed in the seminal paper of Vuong (1989), who showed (in a low dimensional setting) that for correctly specified models the asymptotic null is a weighted sum of \( \chi^2 \)-distributions. In Section 3 we extend Vuong’s result to the high dimensional setting in which model dimensions diverge with sample size. Using those results here, we arrive at a \( p \)-value

\[
p_{\text{split-asym}} = 1 - \Psi_w(LR),
\]

where \( \Psi_w(\cdot) \) denotes the distribution function of a weighted sum of independent \( \chi^2 \)-distributions with weights \( w \) (computation of the weights is discussed in Section 3). Since sparse structure is used in the statistic LR, the finite sample applicability of equation (5) hinges not on the raw dimension \( p \), but rather on the size of the active sets.
Algorithm 1: multisplitting for high dimensional two-sample testing

**Input**: number of splits $B$; screening procedure $\hat{I}();$ data $Z^{(1)}, Z^{(2)}$

**Step 1**: for $b = 1, \ldots, B$ do

**Step 2**: randomly split the data into two halves, $(Z_{in}^{(1)}, Z_{out}^{(1)})$ and $(Z_{in}^{(2)}, Z_{out}^{(2)})$

**Step 3**: screening on the first half, compute active sets $\hat{I}_{in}^{(1)} = \hat{I}(Z_{in}^{(1)})$, $\hat{I}_{in}^{(2)} = \hat{I}(Z_{in}^{(2)})$ and $\hat{I}_{in}^{(12)} = \hat{I}(Z_{in}^{(12)})$

**Step 4**: p-value calculation on the second half, compute test statistic $LR = 2(L_{ind}^{in} - L_{joint}^{in})$, and obtain p-value $p_b$ by either equation (5) or (6)

**Step 5**: end for

**Output**: aggregate p-values $p_1, \ldots, p_B$ and output $p_{agg}$ (using equation (7) or (8))

### 2.5.2. Permutation approach (‘split-perm’)

Use of the asymptotic p-value (5) is computationally favourable but relies on certain additional assumptions (see theorem 1). We consider also an alternative, permutation-based approach that is free from these additional assumptions, but that also exploits sparsity. Denote by $LR_1^*, \ldots, LR_S^*$ the restricted log-likelihood-ratio statistics evaluated at $S$ independent permutations of the group labels. Then a permutation p-value is given by

$$p_{split-perm} = \frac{1}{S} \left\{ 1 + \sum_{s=1}^{S} I(LR_s^* \geq LR) \right\}. \quad (6)$$

In the split-perm procedure, the statistics $LR_1^*, \ldots, LR_S^*$ are evaluated only on the second half of the data with fixed active sets $\hat{I}_{in}^{(1)}, \hat{I}_{in}^{(2)}$, and $\hat{I}_{in}^{(12)}$, and thus, for each permutation $s$, only (constrained, low dimensional) maximum likelihood estimates are needed to calculate $p_{split-perm}$. This renders split-perm feasible even in settings where the computational demands of a conventional permutation test would be prohibitive.

### 2.5.3. p-value aggregation

So far, in both the asymptotic and the permutation approaches, a p-value is computed based on a single, random data split. Following Meinshausen et al. (2009) we consider aggregating results over multiple random data splits indexed by $b = 1, \ldots, B$. Proceeding for each split as described above we obtain p-values $p_1, \ldots, p_B$ which are aggregated to arrive at a final p-value $p_{agg}$. We consider aggregation using the formula

$$p_{agg}^{meinsh} = \min\{(1 - 0.05) \inf_{\gamma \in (0,0.05)} q_{\gamma}(\{p_b/\gamma; b = 1, \ldots, B\}), 1\}, \quad (7)$$

given by Meinshausen et al. (2009), or by taking the median

$$p_{agg}^{median} = \text{median}(p_1, \ldots, p_B). \quad (8)$$

The former guarantees type I error control conditionally on the correctness of the individual p-values, whereas the latter requires additional assumptions on the joint distribution of the p-values (van de Wiel et al., 2009). A generic ‘multisplit’ approach is summarized in algorithm 1 (Table 1).

### 3. Theory

This section studies the asymptotic properties of the restricted log-likelihood-ratio statistic that appears above. We also provide details on the weights and their computation.
3.1. Asymptotic results

3.1.1. Set-up and regularity assumptions

We deviate slightly from the notation that has been used so far and equip several quantities with a subscript $n$ to make dependence on sample size explicit. We treat the active sets $\hat{I}_{\text{in}}^{(1)}$, $\hat{I}_{\text{in}}^{(2)}$ and $\bar{I}_{\text{in}}$ as non-random and omit the circumflex symbol in what follows. However, we allow the size of the active sets to depend on sample size. Without loss of generality we take $n := n_1 = n_2$ and consider for groups $r = 1, 2$ the following ‘triangular array’ set-up: for every $n$, the observations

$$\{z_{ni} = (y_{ni}^{(r)}, x_{ni}^{(r)}); i = 1, 2, \ldots, n\}$$

are IID with conditional density $f_n(y_n|x_n; \theta_n^{(r)})$, $y_n \in \mathbb{R}^{k_n}$, $x_n \in \mathbb{R}^{l_n}$ and $\theta_n^{(r)} \in \mathbb{R}^{p_n}$.

Consider competing models $\mathcal{M}_{\text{ind}}^{n}$ and $\mathcal{M}_{\text{joint}}^{n}$ with dimensions $d_{\text{ind}}^{n} = |I_{\text{in}}^{(1)}| + |I_{\text{in}}^{(2)}|$ and $d_{\text{joint}}^{n} = |\bar{I}_{\text{in}}|$ respectively, and let $(\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)})$ and $\bar{\theta}_n$ be the corresponding pseudotrue values:

$$\hat{\theta}_n^{(r)} = \arg\max_{(\theta_n^{(r)}), \theta_n^{(2)} \in \Theta_{\text{in}}^{(1)} \times \Theta_{\text{in}}^{(2)}} \mathbb{E}[\log\{f_n(y_{n1}^{(r)}|x_{n1}^{(r)}; \theta_n^{(1)}) + \log\{f_n(y_{n2}^{(r)}|x_{n2}^{(r)}; \theta_n^{(2)})\}],

\bar{\theta}_n = \arg\max_{\theta_n^{(12)} \in \Theta_{\text{in}}^{(12)}} \mathbb{E}[\log\{f_n(y_{n1}^{(1)}|x_{n1}^{(1)}; \theta_n^{(12)}) + \log\{f_n(y_{n2}^{(2)}|x_{n2}^{(2)}; \theta_n^{(12)})\}],

Introduce the notation

$$S_J(y_{n1}^{(r)}|x_{n1}^{(r)}, \bar{\theta}) = \nabla_J \log\{f_n(y_{n1}^{(r)}|x_{n1}^{(r)}; \theta)\} \quad \text{for some } J \subset \{1, \ldots, p_n\}$$

and set

$$S_{I_{\text{in}}^{(r)}}^{(1)} = S_{I_{\text{in}}^{(r)}}^{(1)}(y_{n1}^{(r)}|x_{n1}^{(r)}; \hat{\theta}_n^{(1)}),

S_{I_{\text{in}}^{(r)}}^{(2)} = S_{I_{\text{in}}^{(r)}}^{(2)}(y_{n1}^{(r)}|x_{n1}^{(r)}; \hat{\theta}_n^{(2)}),

S_{\bar{I}_{\text{in}}^{(r)}}^{(12)} = S_{\bar{I}_{\text{in}}^{(r)}}^{(12)}(y_{n1}^{(r)}|x_{n1}^{(r)}; \bar{\theta}_n^{(12)}).$$

Then, the score functions for models $\mathcal{M}_{\text{ind}}^{n}$ and $\mathcal{M}_{\text{joint}}^{n}$ are given by

$$S_{\text{ind}}^{n} = (S_{I_{\text{in}}^{(r)}}^{(1)}, S_{I_{\text{in}}^{(r)}}^{(2)})$$

and

$$S_{\text{joint}}^{n} = S_{\bar{I}_{\text{in}}^{(r)}}^{(12)} + S_{I_{\text{in}}^{(r)}}^{(12)}$$

respectively. Moreover, introducing the notation $B(U) = \mathbb{E}(UU^T)$ and $B(U, V) = \mathbb{E}(UV^T)$ (where $U$ and $V$ denote random vectors), we obtain

$$B(S_{\text{ind}}^{n}) = \begin{pmatrix} B(S_{I_{\text{in}}^{(r)}}^{(1)}) & 0 \\ 0 & B(S_{I_{\text{in}}^{(r)}}^{(2)}) \end{pmatrix},

B(S_{\text{joint}}^{n}) = B(S_{\bar{I}_{\text{in}}^{(r)}}^{(12)}) + B(S_{I_{\text{in}}^{(r)}}^{(12)})$$

and

$$B(S_{\text{joint}}, S_{\text{ind}}^{n}) = (B(S_{\bar{I}_{\text{in}}^{(r)}}^{(12)}, S_{I_{\text{in}}^{(r)}}^{(1)}), B(S_{I_{\text{in}}^{(r)}}^{(2)}, S_{I_{\text{in}}^{(r)}}^{(2)})).$$
The following results are based on regularity assumptions (L1)–(L4), in the on-line supplementary material, which are assumed to hold for each of the two models. In essence we take the same approach as Fan and Peng (2004) and impose higher order moments of the likelihood function to facilitate derivations. Precise statements of these assumptions can be found in the on-line supplementary material. In brief, assumption (L1) assumes that the model-specific information matrix is positive definite and its eigenvalues uniformly bounded (with respect to sample size $n$). Assumption (L2) imposes fourth moments of the likelihood function. Assumption (L3) states that third-order partial derivatives of the log-likelihood function are dominated by integrable functions and assumption (L4) assumes that the maximal eigenvalue of the combined information matrix is uniformly bounded.

### 3.1.2. Main results

The following proposition which follows from theorem 1.1 in the on-line supplementary material guarantees existence of root $\hat{\theta}_{\text{ind}}^{d}/n$ and root $\hat{\theta}_{\text{joint}}^{d}/n$ consistent estimators for both models.

**Proposition 1.** Assume regularity assumptions (L1)–(L3) in the on-line supplementary material. If $\sqrt{d_{\text{ind}}^{1}}/n = o(1)$, then there are local maximizers $\hat{\theta}_{\text{ind}}^{d} / n$ and $\hat{\theta}_{\text{joint}}^{d} / n$ of $\mathcal{L}_{\text{ind}}(\theta^{(1)}, \theta^{(2)})$ and $\mathcal{L}_{\text{joint}}(\theta^{(12)})$ respectively, such that

$$\left\| \begin{pmatrix} \hat{\theta}_{n}^{(1)} - \tilde{\theta}_{n}^{(1)} \\ \hat{\theta}_{n}^{(2)} - \tilde{\theta}_{n}^{(2)} \end{pmatrix} \right\| = O_p \{ \sqrt{d_{\text{ind}}^{1}/n} \}$$

and

$$\|\hat{\theta}_{n}^{(12)} - \tilde{\theta}_{n}^{(12)}\| = O_p \{ \sqrt{d_{\text{joint}}^{d}/n} \}.$$ 

Consider the restricted log-likelihood-ratio statistic

$$R_n = 2 \{ \mathcal{L}_{\text{ind}}(\hat{\theta}_{n}^{(1)}, \hat{\theta}_{n}^{(2)}) - \mathcal{L}_{\text{joint}}(\hat{\theta}_{n}^{(12)}) \},$$

and define the matrix

$$W = \begin{pmatrix} I_{d_{\text{ind}}^{1}} & B(S_{\text{ind}}^{-1}, S_{\text{joint}}) B(S_{\text{joint}}^{-1}) \\ -B(S_{\text{joint}}^{-1}, S_{\text{ind}}) B(S_{\text{ind}}^{-1}) & -I_{d_{\text{joint}}^{d}} \end{pmatrix}.$$  \hspace{1cm} (9)

Set $d_{n} := d_{\text{ind}}^{d} + d_{\text{joint}}^{d}$ and let $w = (w_{1}, \ldots, w_{d_{n}})$ be the vector of eigenvalues of $W$, which are all real (see proposition 3). Define the weighted sum of $\chi^{2}$ random variable

$$R_n := \sum_{j=1}^{d_{n}} w_{nj} U_{j}^{2}, \quad U_{1}, \ldots, U_{d_{n}} \overset{\text{IID}}{\sim} \mathcal{N}(0, 1).$$

We write $\Psi_{w}(\cdot)$ for the distribution function of $R_n$ and set

$$m_{n} = \text{tr}(W),$$

$$\tau_{n}^{2} = \text{tr}(W^{2}),$$

$$r_{n} = \text{rank}(W).$$

The following proposition shows that $R_n$, appropriately centred and scaled, can be asymptotically approximated by a standard normal distribution.

**Proposition 2.** If $\tau_{n}^{2} \to \infty$ and $\max_{j}|w_{nj}| = O(1)$, then

$$\left( R_n - m_{n} \right)/ \sqrt{2\tau_{n}^{2}} \overset{D}{\to} \mathcal{N}(0, 1).$$
Proof. The result follows from Lindeberg’s central limit theorem: consider the triangular array

\[ V_{nj} := \frac{w_{nj}}{\sqrt{2\tau_n^2}} (U_j^2 - 1), \quad j = 1, \ldots, d_n. \]

For each \( n \), \( V_{n1}, \ldots, V_{nd_n} \) are independent, \( \mathbb{E}(V_{nj}) = 0 \), and \( \sum_{j=1}^{d_n} \mathbb{E}(V_{nj}^2) = 1 \). Furthermore, we have

\[
\sum_{j=1}^{d_n} \mathbb{P}\left\{ V_{nj}^2 \mathbf{1}(|V_{nj}| > \epsilon) \right\} = \sum_{j=1}^{d_n} \frac{w_{nj}^2}{2\tau_n^2} \mathbb{E}\left[ (U_j^2 - 1)^2 \mathbf{1}\left\{ \frac{\epsilon \sqrt{2\tau_n^2}}{|w_{nj}|} \right\} \right] \\
\leq \frac{1}{2} \mathbb{E}\left[ (U_1^2 - 1)^2 \mathbf{1}\left\{ \frac{\epsilon \sqrt{2\tau_n^2}}{\max_j |w_{nj}|} \right\} \right].
\]

Thus, \( \tau_n^2 \rightarrow \infty \) and \( \max_j |w_{nj}| = O(1) \) imply the Lindeberg condition.

We now consider the \( H_0 \)-scenario and set \( I_0 = \{ j : \theta_{0,j} \neq 0 \} \), where \( \theta_0 := \theta_{01} = \theta_{02} \). The following theorem establishes the asymptotic properties of the restricted log-likelihood-ratio statistic under the null hypothesis.

Theorem 1. Consider the \( H_0 \)-scenario. Suppose assumptions (L1)–(L4) in the on-line supplementary material, \( \tau_n^2 \approx r_n \) and \( r_n \rightarrow \infty \), and assume that

(a) \( I_0 \subset I_{\text{in}}^{(1)} \cap I_{\text{in}}^{(2)} \cap I_{\text{in}}^{(12)} \) and
(b) \( d_n^5/n \rightarrow 0 \).

Then, we have

(i) \( (\text{LR}_n - m_n)/\sqrt{(2\tau_n^2)} \rightarrow^D \mathcal{N}(0, 1) \) and
(ii) \( \sup_{x \in \mathbb{R}} |\mathbb{P}(\text{LR}_n \leq x) - \Psi_w(x)| \rightarrow 0. \)

Proof. Part (i) follows from a more general result which we derive in the on-line supplementary material (theorem 1.2). The proof involves a martingale central limit theorem and uses techniques that were described in Fan and Peng (2004) and Peng and Schick (2012). Part (ii) is a consequence of proposition 2 and part (i) of theorem 1.

The screening assumption (a) in theorem 1 guarantees that the competing models \( \mathcal{M}_n^{\text{ind}} \) and \( \mathcal{M}_n^{\text{joint}} \) are correctly specified. The rate of \( o(n^{1/5}) \) in the sparsity assumption (b) is in line with Fan and Peng (2004), which is not surprising as we impose similar regularity assumptions. Improvements could be obtained by considering exponential families (see Portnoy (1988)). Besides assumptions (L1)–(L4) in the on-line supplementary material and conditions (a) and (b) in theorem 1 we further require \( \tau_n^2 \approx r_n \) and \( r_n \rightarrow \infty \). The quantity \( r_n \) can be seen as the effective dimensionality in the model comparison, e.g. in the nested case where \( \mathcal{M}_n^{\text{joint}} \subset \mathcal{M}_n^{\text{ind}} \), \( r_n = d_n^{\text{ind}} - d_n^{\text{joint}} \) (this follows from proposition 3 in Section 3.2).

The statistic \( \text{LR}_n \) is based on the local maximizers from proposition 1. The assertion of proposition 1 is easily misunderstood. It does not guarantee the existence of root \( d_n^{\text{ind}}/n \) and root \( d_n^{\text{joint}}/n \) consistent sequences of maximum likelihood estimators. It only states that a clairvoyant statistician (with preknowledge of the pseudotrue values) can choose local maximizers with the desired properties (van der Vaart, 2007).

3.1.3. Discussion for random active sets

In the above discussion we treated the active sets as non-random. However, in practice the active
sets are random as they are obtained via sparse estimation (applied to first-half data). Theorem 1 does not directly translate to this setting, partly because of the additional assumptions that we impose (besides screening and sparsity). It may be possible to obtain results for specific parametric families where these assumptions are implicitly fulfilled, but we do not pursue this direction here. Instead, we sketch how the general result of theorem 1 can lead to type I error control under data splitting. Consider the following screening and sparsity assumptions:

\[
\begin{align*}
(a') & \Pr(I_0 \subset \hat{I}^{(1)}_{in} \cap \hat{I}^{(2)}_{in} \cap \hat{I}^{(12)}_{in}) \to 1; \\
(b') & d^5_{in}/n \to 0 \text{ almost surely.}
\end{align*}
\]

Define the sequence of events \(A_n = \{I_0 \subset \hat{I}^{(1)}_{in} \cap \hat{I}^{(2)}_{in} \cap \hat{I}^{(12)}_{in}, \lim_{n \to \infty} d^5_{in}/n = 0\} \) and note that assumptions (a') and (b') imply that \(\Pr(A^c_n) \to 0\). Therefore, we have

\[
\Pr(p_{\text{split-asym}} \leq \alpha) = \Pr(p_{\text{split-asym}} \leq \alpha, A_n) + o(1). \tag{10}
\]

Conditionally on the first-half data we have non-random active sets. Therefore,

\[
\Pr(p_{\text{split-asym}} \leq \alpha, A_n | Z_{in}^{(1)}, Z_{in}^{(2)}) = 1 - \Pr\{LR_n \leq \Psi^{-1}_w (1 - \alpha), A_n | Z_{in}^{(1)}, Z_{in}^{(2)}\} = \alpha - \Pr\{LR_n \leq \Psi^{-1}_w (1 - \alpha), A_n | Z_{in}^{(1)}, Z_{in}^{(2)}\} - \Psi_w \{\Psi_w^{-1} (1 - \alpha)\}
\]

\[= \alpha + o(1), \tag{11}\]

where the last equality follows from theorem 1, part (ii), and independence of first- and second-half data. Finally, combining equations (10) and (11), we arrive at a statement on type I error control, i.e. \(\Pr(p_{\text{split-asym}} \leq \alpha) = \alpha + o(1)\). We emphasize that this sketch ignores other assumptions besides (a) and (b) when invoking theorem 1.

### 3.2. Characterization and estimation of weights \(w\)

The weights \(w\) (these are eigenvalues of the matrix \(W\); see equation (9)) are required to obtain the asymptotic null distribution; here we provide results that clarify their interpretation and show how they can be efficiently computed in practice. Let \(I^0 = I^{(1)}_{in} \cap I^{(2)}_{in} \cap I^{(12)}_{in}\) be the intersection of all three active sets and denote with \(\hat{I}^{(1)}_{in} = I^{(1)}_{in} - I^0_{in}, \hat{I}^{(2)}_{in} = I^{(2)}_{in} - I^0_{in}\) and \(\hat{I}^{(12)}_{in} = I^{(12)}_{in} - I^0_{in}\) their complements. Partition the score functions of \(\mathcal{M}^{\text{ind}}_{in}\) and \(\mathcal{M}^{\text{joint}}_{in}\) according to

\[
S^{\text{ind}}_{in} = (S^{\text{ind}}_{\theta^0}, S^{\text{ind}}_{\rho^0}, S^{\text{ind}}_{\rho^0}), \quad S^{\text{joint}}_{in} = (S^{\text{joint}}_{\rho^0}, S^{\text{joint}}_{\rho^0}, S^{\text{joint}}_{\rho^0}).
\]

Denoting \(d^0 = |I^0_{in}|, d^{(1)} = |\hat{I}^{(1)}_{in}|, d^{(2)} = |\hat{I}^{(2)}_{in}|\) and \(d^{(12)} = |\hat{I}^{(12)}_{in}|\), we prove in Appendix A the following proposition.

**Proposition 3.** Consider the \((d^{(12)} - d^0) \times (d^{(12)} - d^0)\) matrix

\[
\tilde{W} = \mathbf{1}_{d^{(12)}-d^0} - Q(S^{\text{joint}}_{\rho^0} | S^{\text{ind}}_{\rho^0})Q(S^{\text{joint}}_{\rho^0} | S^{\text{ind}}_{\rho^0})^{-1}, \tag{12}
\]

where we have set \(Q(U | V) = B(U) - B(U, V)B(V)^{-1}B(V, U)\).

Then, the weights \(w\) can be characterized as follows.

(a) The weights are real numbers which lie between \(-1\) and 1.
(b) \(2d^0\) weights are 0.
(c) \(d^{(1)} + d^{(2)} - d^{(12)} + (d^{(12)} - d^0 - \text{rank} (\tilde{W}))\) weights are 1.
(d) \(d^{(12)} - d^0 - \text{rank} (\tilde{W})\) weights are \(-1\).
(e) The remaining weights equal $\pm \sqrt{(1 - \tilde{w}_j)}$, where $\tilde{w}_j$, $j = 1, \ldots$, rank($\hat{W}$), are the non-zero eigenvalues of $\hat{W}$.

Proposition 3 provides some interesting insights. For example, it turns out that the sum of all weights equals $d^{(1)} + d^{(2)} - d^{(12)}$. Similarly, in the nested case where $M_{\text{joint}} \subset M_{\text{ind}}$ we have $d^{(12)} - d^0 = 0$ and the null distribution reduces to the familiar $\chi^2_{d^{(1)} + d^{(2)} - d^{(12)}}$-distribution. In most cases, however, we deal with non-nested model comparison where the matrix $\hat{W}$ determines some of the weights. Interestingly, this matrix has some similarity with the coefficient of determination ($R^2$). Specifically, $Q(S_{\rho^0_{\text{ind}}}^{\text{joint}}, S_{\rho^0_{\text{ind}}}^{\text{ind}})$ is the variance of $S_{\rho^0_{\text{ind}}}^{\text{joint}}$ after projecting onto $S_{\rho^0_{\text{ind}}}^{\text{ind}}$ and $Q(S_{\rho^0_{\text{joint}}}^{\text{joint}}, S_{\rho^0_{\text{ind}}}^{\text{ind}})$ quantifies the (partial) variance of $S_{\rho^0_{\text{ind}}}^{\text{joint}}$ not explained by $S_{\rho^0_{\text{ind}}}^{\text{ind}}$.

In practice some of the weights $\theta$ are unknown and must be estimated. One possibility is to estimate $\hat{W}$ and then to compute eigenvalues to obtain $\tilde{w}$. Alternatively, one could estimate $\hat{W}$ and invoke proposition 3 to arrive at $\hat{w}$. We recommend the latter as the number of weights to be determined is much smaller. Finally, computation of $p_{\text{split-asym}}$ requires evaluation of the distribution function $\Psi_{\rho}()$. We use the approach of Davies (1980) (based on numerical inversion of the characteristic function), as implemented in the R-package CompQuadForm.

Full details can be found in the on-line supplementary material.

4. Examples: differential regression and differential network

Here we provide two applications of our methodology, to regression and Gaussian graphical models (GGMs), that we call differential regression and differential network respectively. These examples illustrate how the general approach can be instantiated for specific models and screening procedures.

4.1. Differential regression (DiffRegr)

Consider the linear regression model

$$Y = X \beta + \epsilon, \quad (13)$$

with $Y \subset \mathbb{R}$ (i.e. $k = 1$ in the general conditional model (2)), $X \subset \mathbb{R}^l$, random error $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\theta = (\beta, \sigma^2) \in \Theta \subset \mathbb{R}^p$, $p = l + 1$. $l_1$-penalized estimation in the screening step coincides with the lasso (Tibshirani, 1996). We use the implementation in the R-package glmnet (Friedman et al., 2010) and choose $\lambda$ by tenfold cross-validation, giving a value $\lambda_{cv}$. For finite sample sizes the active sets that are obtained from $\hat{\beta}^{\lambda_{cv}}$ may be too large. We therefore consider the modified procedure

$$\hat{\beta}_{\text{DiffRegr}, \tau}^{\lambda_{cv}}(Z_{\text{in}}) = \{ j : |\hat{\beta}_j^{\lambda_{cv}}| \text{ is among the } \tau \text{ largest of all non-zero coefficients} \},$$

where $Z_{\text{in}} = (Y_{\text{in}}, X_{\text{in}})$ has $n_{\text{in}}$ samples and is one of the data sets $Z_{\text{in}}^{(1)}, Z_{\text{in}}^{(2)}$ or $Z_{\text{in}}^{(12)}$. The tuning parameter $\tau$ upper-bounds the size of the active sets. We take $\tau = \lceil n_{\text{in}}/5 \rceil$ in all examples, corresponding to a typical rule of thumb in linear regression of having at least five samples per predictor. Evaluation of LR requires constrained maximum likelihood estimation. In the case of linear regression these are simply least squares problems and computation of the permutation-based $p$-value $p_{\text{split-perm}}$ is therefore feasible. Asymptotic $p$-values ($p_{\text{split-asym}}$) are obtained following Section 3.2 and using the following lemma whose proof appears in the on-line supplementary material.

Lemma 1. Assume that $Y|X = x$ is distributed according to regression model (13) with $\theta = (\beta, \sigma^2)$. Given $J_\alpha$, $J_\beta \subset \{ 1, \ldots, l \}$ and $\theta_\alpha = (\beta_\alpha, \sigma^2_\alpha)$ and $\theta_\beta = (\beta_\beta, \sigma^2_\beta)$ we have
4.2. Differential network (DiffNet)

Let $Y \in \mathbb{R}^k$ ($l = 0$) have a Gaussian distribution with zero mean and covariance $\Sigma$. The undirected conditional independence graph $G$ of the corresponding GGM is defined by locations of non-zero entries in the inverse covariance matrix $\theta = \Sigma^{-1}$. For $l_1$-penalized estimation of GGMs we use the graphical lasso as implemented in the R package glasso (Friedman et al., 2008). We set the regularization parameter to a value $\lambda_{cv}$ by using tenfold cross-validation and by analogy with the regression case above consider the thresholding procedure

$$
I_{\text{DiffNet}, \tau}(\mathbf{Z}_{in}) = \{ (j, j') : \hat{\theta}_{jj'}^{\lambda_{cv}} \text{ is among the } \tau \text{ largest of all non-zero (off-diagonal) entries} \}.
$$

Similarly to the regression case, we choose $\tau = \lfloor p \times n_{in}/5 \rfloor$ (estimation of the concentration matrix $\theta$ in a GGM with graph $G$ is related to regression of each variable (or node) on all neighbouring variables). We consider asymptotic $p$-values and obtain $\tilde{w}$ by using the following lemma whose proof appears in the on-line supplementary material.

Lemma 2. Assume that $Y$ follows the GGM with $\theta = \Sigma^{-1}$. Consider two pairs of nodes $(j, j')$ and $(l, l')$, and concentration matrices $\theta_a = \Sigma_a^{-1}$ and $\theta_b = \Sigma_b^{-1}$. Then, we have

$$
\mathbb{E}_\theta \{ S(j, j')(Y; \theta_a)S(l, l')(Y; \theta_b) \} = \Sigma_{jj'}\Sigma_{ll'} + \Sigma_{jl}\Sigma_{l'j'} + \Sigma_{j'l}\Sigma_{j'l'} - \Sigma_{jj'}\Sigma_{l'l'} - \Sigma_{jl}\Sigma_{j'l'} - \Sigma_{j'l}\Sigma_{jl'} + \Sigma_{a, jj'}\Sigma_{b, ll'}.
$$

5. Numerical results

In this section we study the performance of differential regression, DiffRegr, and differential network, DiffNet, on simulated and biological data. Calculations are as described in Section 4. For the multisplit procedure MS (with $B = 50$ for DiffRegr and $B = 25$ for DiffNet) we use the median for $p$-value aggregation (the median outperformed the alternative approach using formula (7) in most examples). $p$-value calculations are based either on the asymptotic approximation, asym, or on $S = 250$ permutations, perm. We also show results from the corresponding single-split variant SS (i.e. $B = 1$). For DiffNet we considered only asymptotic $p$-values since, although constrained maximum likelihood estimation for GGMs is a convex, low dimensional optimization task, currently available implementations remain too slow for obtaining permutation-based $p$-values.

5.1. Differential regression

We sampled $n_1 = n_2 = 100$ observations according to a linear regression model (as described under DiffRegr above) with $X^{(1)}, X^{(2)} \sim \mathcal{N}(0, \Sigma)$ and $\Sigma_{jj'} = 0.5^{|j-j'|}$. Coefficients $\beta^{(r)}$ were generated such that under the null hypothesis, i.e. $\beta^{(1)} = \beta^{(2)}$, there are $s$ non-zero entries at random locations, each with magnitude 1. Under the alternative, the regression parameters $\beta^{(1)}$ and $\beta^{(2)}$ have $s$ non-zero entries at random locations; $s - 2$ of them are shared between conditions and have magnitude 1 and the remaining two non-zero entries are at different locations and have magnitude 0.5. Thus, $s$ controls the sparsity level, with smaller $s$ corresponding to a sparser setting. The alternative case is intended to challenge the analyses by providing a scenario where the two conditions differ in a subtle manner with most coefficients remaining the same. The noise variances $\sigma^{(r)}$ were adjusted to control the signal-to-noise-ratio snr at predefined values. We considered three settings.
(a) Setting 1: vary the number of predictors $l = 20, 40, 80, 160$, fixing $s = 5$ and $\text{snr} = 5$.

(b) Setting 2: vary the signal-to-noise ratio $\text{snr} = 1.5, 3, 5, 10$, fixing $l = 80$ and $s = 5$.

(c) Setting 3: vary the sparsity level $s = 3, 5, 7, 10$, fixing $l = 80$ and $\text{snr} = 5$.

For each of the three settings we performed 200 simulation runs for the null and the alternative case. We used DiffRegr and report $p$-values that were obtained by using single splitting and multisplitting. We compared DiffRegr with the classical LRT (asymptotic $\chi^2$ null distribution, $p=l+1$). We considered also a permutation test based on the symmetric Kullback–Leibler divergence $D_{\text{symm}}(\hat{\beta}_1^{(1)} , \hat{\sigma}_1^{(1)} || \hat{\beta}_1^{(2)} , \hat{\sigma}_1^{(2)} )$, where $(\hat{\beta}_1^{(1)} , \hat{\sigma}_1^{(1)} )$ and $(\hat{\beta}_1^{(2)} , \hat{\sigma}_1^{(2)} )$ are lasso estimates obtained from group-specific data and $\lambda$ is chosen by tenfold cross-validation. $p$-values were obtained via $S = 250$ group label permutations and we call this test KLPerm. For all settings and all methods we performed a receiver operating characteristic (ROC) analysis by considering true positive (TPR) and false positive rates (FPR) at the 5% cut-off. This observation raises the question of whether an improved aggregation method could yield better (less conservative) results. The single-split approach SS shows similar performance when comparing asymptotic (asym) and permutation-based $p$-values (perm). As expected the latter shows better type I error control (at the 5% level) but rather tend to be conservative.

5.2. Differential network

Data were generated from a zero-mean GGM with $k = 30$ nodes. The sample size $n (= n_1 = n_2)$ was 35 or 70 and we considered the following conditional independence structures.

(a) Model 1 (random networks): networks have $s \times k$ edges, between randomly selected pairs of nodes. The simulation parameter $s$ controls sparsity, with small $s$ giving a sparser graph. For model 1, we set $s = 1$. Under the alternative, $G^{(1)}$ and $G^{(2)}$ have 80% of the edges in the same locations and 20% of the edges in different locations.

(b) Model 2 (sparse random networks): model 2 is like model 1 but with $s = 0.2$. Again, under the alternative, 80% of the edges are in the same locations and 20% in different locations.

(c) Model 3 (hub networks): networks consist of disjoint star subgraphs with 10 nodes, in which a hub connects to all other nodes in its subgraph. Under the null hypothesis, the networks have two 10-star subgraphs; under the alternative, $G^{(1)}$ and $G^{(2)}$ have two common 10-star subgraphs and one of the networks has an additional third 10-star subgraph.

Given graphs $G^{(1)}$ and $G^{(2)}$ we generate concentration matrices $\theta^{(1)}$ and $\theta^{(2)}$ as follows.

(a) Set $B^{(r)} = \delta_d A^{(r)} + (\delta_c - \delta_d) A^{(1)} \ast A^{(2)}$. Here, $A^{(r)}$ denotes the adjacency matrix of $G^{(r)}$ and $A^{(1)} \ast A^{(2)}$ is the elementwise product of the two matrices ($\delta_d$, the magnitude of non-zero
Fig. 1. Simulation results for differential regression, DiffRegr—area under the ROC curve as a function of dimensionality \( l \), signal-to-noise ratio \( \text{snr} \) and sparsity \( s \) (the ROC results were obtained by considering true positive and false positive rates for various \( p \)-value cut-offs where rates are with respect to the number of null and alternative cases in the simulation set-up) (—–, MS(asym) (multisplit DiffRegr, asymptotic \( p \)-value); – –, SS(asym) (single-split DiffRegr, asymptotic \( p \)-value); – –Δ– –, MS(perm) (multisplit DiffRegr, permutation-based \( p \)-value); – –Δ– –, SS(perm) (single-split DiffRegr, permutation-based \( p \)-value); +, KLPerm (Kullback–Leibler divergence between group-specific lasso estimates); ×, LRT): (a) performance when varying dimensionality \( l \) (with \( s = 5 \) and \( \text{snr} = 5 \); see the text for details of the data-generating set-up); (b) varying signal-to-noise ratio \( \text{snr} \) (with \( l = 80 \) and \( s = 5 \)); (c) varying sparsity \( s \) (with \( l = 80 \) and \( \text{snr} = 5 \); smaller \( s \) corresponds to a sparser data-generating model).

entries that differ between groups; \( \delta_c \), the magnitude of non-zero entries that are common between groups; we take \( \delta_d = 0.8 \) and \( \delta_c = 0.9 \) in all experiments).

(b) Set \( \theta^{(r)} = B^{(r)} + (e_{\text{min}} + 0.1) I_k \); \( e_{\text{min}} \) is the smallest eigenvalue of \( B^{(r)} \) and \( I_k \) denotes the identity matrix.

As in Section 5.1 we performed 200 simulation runs under the null and alternative hypotheses. We use single-split and multisplit versions of DiffNet and report the area under the ROC curve, AUC, as well as true positive and false positive rates TPR and FPR at the \( \alpha = 5\% \) cut-off. We compare our approach with the classical LRT (asymptotic \( \chi^2_p \) null distribution, \( p = k(k + 1)/2 \))
Table 2. Simulation results for differential regression, DiffRegr: true positive rate TPR and false positive rate FPR at $\alpha = 5\%$ cut-off (with FPR in parentheses), as a function of dimensionality $l$, signal-to-noise ratio $\text{snr}$ and sparsity $s^\dagger$

<table>
<thead>
<tr>
<th>Method</th>
<th>Results for TPR (FPR) at 5% level: vary dimensionality $l$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$l = 20$</td>
</tr>
<tr>
<td>MS(asym)</td>
<td>0.86 (0)</td>
</tr>
<tr>
<td>SS(asym)</td>
<td>0.78 (0.12)</td>
</tr>
<tr>
<td>MS(perm)</td>
<td>0.86 (0)</td>
</tr>
<tr>
<td>SS(perm)</td>
<td>0.77 (0.06)</td>
</tr>
<tr>
<td>KLPerm</td>
<td>0.84 (0.07)</td>
</tr>
<tr>
<td>LRT</td>
<td>0.97 (0.14)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Results for TPR (FPR) at 5% level: vary signal-to-noise ratio $\text{snr}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{snr} = 1.5$</td>
</tr>
<tr>
<td>MS(asym)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>SS(asym)</td>
<td>0.08 (0.06)</td>
</tr>
<tr>
<td>MS(perm)</td>
<td>0.01 (0.01)</td>
</tr>
<tr>
<td>SS(perm)</td>
<td>0.06 (0.06)</td>
</tr>
<tr>
<td>KLPerm</td>
<td>0.06 (0.04)</td>
</tr>
<tr>
<td>LRT</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Results for TPR (FPR) at 5% level: vary sparsity $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s = 3$</td>
</tr>
<tr>
<td>MS(asym)</td>
<td>1 (0)</td>
</tr>
<tr>
<td>SS(asym)</td>
<td>1 (0.06)</td>
</tr>
<tr>
<td>MS(perm)</td>
<td>1 (0)</td>
</tr>
<tr>
<td>SS(perm)</td>
<td>1 (0.04)</td>
</tr>
<tr>
<td>KLPerm</td>
<td>0.92 (0.04)</td>
</tr>
<tr>
<td>LRT</td>
<td>1 (1)</td>
</tr>
</tbody>
</table>

$^\dagger$Rates are with respect to the number of null and alternative cases in the simulation set-up. Top: performance when varying dimensionality $l$ (with $s = 5$ and $\text{snr} = 5$; see the text for details of the data-generating set-up). Middle: varying signal-to-noise ratio $\text{snr}$ (with $l = 80$ and $s = 5$). Bottom: varying sparsity $s$ (with $l = 80$ and $\text{snr} = 5$; note that smaller $s$ corresponds to a sparser data-generating model). Methods: multisplit DiffRegr, MS; single-split DiffRegr, SS; $p$-values from data splitting and asymptotic results, asym; $p$-values from data splitting and permutation, perm; permutation test based on the Kullback–Leibler divergence between group-specific lasso estimates, KLPerm; classical LRT.

and with a recently proposed two-sample test for high dimensional covariance matrices due to Li and Chen (2012), method LiChen.

Results are shown in Table 3. The conclusions are similar to the regression case: in terms of AUC, multisplitting is very competitive in all scenarios. MS outperforms all other methods in the $n = 70$ scenario. As above we find that MS yields much too conservative results at the 5\% level: the $n = 35$ setting is an extreme case with very low TPR despite good AUC-scores. This again suggests that better approaches for $p$-value aggregation are needed. Method LiChen shows good performance in most scenarios. MS seems to perform better in sparse settings (model 2) whereas LiChen works well in denser scenarios (model 1).
Table 3. Simulation results for differential network, DiffNet†

<table>
<thead>
<tr>
<th>Method</th>
<th>Results for model 1</th>
<th>Results for model 2</th>
<th>Results for model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>TPR</td>
<td>FPR</td>
</tr>
<tr>
<td><strong>n = 35</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS(asym)</td>
<td>0.98</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>SS(asym)</td>
<td>0.87</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td>LiChen</td>
<td>0.96</td>
<td>0.86</td>
<td>0.10</td>
</tr>
<tr>
<td>LRT</td>
<td>0.57</td>
<td>0.68</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>n = 70</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS(asym)</td>
<td>I</td>
<td>0.94</td>
<td>0</td>
</tr>
<tr>
<td>SS(asym)</td>
<td>0.99</td>
<td>0.84</td>
<td>0.02</td>
</tr>
<tr>
<td>LiChen</td>
<td>I</td>
<td>0.99</td>
<td>0.08</td>
</tr>
<tr>
<td>LRT</td>
<td>I</td>
<td>1</td>
<td>0.93</td>
</tr>
</tbody>
</table>

†ROC performance of p-values thresholded at various α-levels: area under the curve AUC, true positive rate TPR and false positive rate FPR at α = 5% cut-off are shown for network models with k = 30 nodes and sample sizes n (n1 = n2 =) 35 and 70. Model 1, random network; model 2, sparse random network; model 3, hub network. Methods: multisplit DiffNet, MS; single-split DiffNet, SS; the test proposed by Li and Chen (2012), LiChen; classical LRT.

5.3. Application to biological data

5.3.1. Differential regression

We used data from the Cancer Cell Line Encyclopaedia (the data are publically available from http://www.broadinstitute.org/ccle/home). Barretina et al. (2012) used this data set to predict response to therapy by using a lasso-type approach. They described that the histone deacetylase inhibitor panobinostat shows increased sensitivity in haematological cancers compared with solid cancers.

We used DiffRegr (multisplit, with B = 50) to compare the three cancer subtypes with the largest sample size: lung cancer (89 cell lines), skin cancer (71 cell lines) and cancer with haematopoietic and lymphoid tissue origin (40 cell lines). As the response we took experimentally determined sensitivity to panobinostat and as covariates gene expressions of the l = 250 genes showing highest Pearson correlation over all samples with panobinostat sensitivity. Fig. 2 shows histograms of all 50 p-values obtained by multisplit DiffRegr. This illustrates the sensitivity of p-values to a single data split but also the information that is contained in the entire distribution of p-values obtained via multisplitting. For example in the comparison of skin against haematopoietic cancer we see a wide range of (single-split) p-values. However, the distribution of these p-values is heavily skewed towards zero which is reflected in an aggregated p-value of 0.047. For each pair of conditions we additionally carried out ‘back-testing’ by pooling data from the two conditions and dividing randomly into two groups. All p-values obtained from applying DiffRegr under back-testing are non-significant (see Figs 2(d)–2(f)).

5.3.2. Differential network

We considered protein data (expression of k = 51 phosphoproteins) from samples spanning 11 cancer types (data from The Cancer Genome Atlas and available from http://www.cancer genome.nih.gov). Differential network permits hypothesis testing regarding multivariate
Two-sample Testing

![Histograms over p-values](image)

**Fig. 2.** Differential regression, cancer data example, histograms over p-values: (a)–(c) histograms of individual p-values obtained from the multisplit method with 50 splits comparing the different cancer subtypes (a) lung–skin, (b) lung–haematopoietic and (c) skin–haematopoietic; (d)–(f) corresponding histograms obtained from back-testing, where for each comparison the two conditions are pooled and then randomly divided into two (pseudo)groups.
Fig. 3. Comparison of cancer types by using differential network applied to protein data from patient samples ($O$, $n_2 = 100$; $\Delta$, $n_2 = 75$; $+$, $n_2 = 50$; $\times$, $n_2 = 25$) (LUSC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; HNSC, head and neck squamous cell carcinoma; UCEC, uterine corpus endometrial carcinoma; BLCA, bladder urothelial carcinoma; READ, rectal adenocarcinoma; COAD, colon adenocarcinoma; KIRC, renal clear cell carcinoma; OVCA, high grade serious ovarian cystadenocarcinoma; BRCA, breast cancer; GBM, glioblastoma multiforme): we compared $n_1 = 195$ LUSCs with samples from the other cancer types; for each of the other types we considered a random $n_2$-subset ($n_2 = 25, 50, 75, 100$) of samples such that observations sampled for smaller $n_2$ are included in data sets obtained for larger $n_2$; the results were obtained by using DiffNet with $B = 50$ data splits and median aggregation.

We put forward a new approach for two-sample testing that uses estimated sparsity patterns to give a general—yet tractable and effective—high dimensional analogue to the classical LRT. Some notion of sparsity is essential for much of high dimensional statistics. Increasingly sparse estimation is being extended to a wider class of low dimensional structures beyond the familiar regression set-up. Our results show how to exploit sparse structure for testing within an other-
A limitation of our analysis is the screening assumption that the estimators are required to satisfy. This is a rather general issue in sparse estimation and in practice it remains difficult to be assured that all relevant parameter components will be selected. For the lasso it has been shown that screening requires a ‘beta-min’ condition; this is arguably undesirable in testing as a test should control type I error regardless of the size of the coefficients (see Bühlmann (2013)). However, we note that our empirical results suggested that asymptotic $p$-values do not behave erratically in challenging settings where the screening assumption is violated. Rather, the $p$-values tend to be conservative. Nevertheless, further work is needed to understand the conditions under which such behaviour can be expected.

We found empirically that the statistic proposed was effective in an ROC sense, outperforming, for example, a permutation test based on the Kullback–Leibler divergence that also made use of sparse estimates. However, our theoretical results concerned type I error control only and we did not study power or optimality. This will be an important direction for future theoretical work that could shed light on how sparsity patterns can optimally be exploited for testing. The multisplit procedure improved ROC performance over single splitting. Multisplitting also improves stability and this is a desirable feature in practical problems. Our empirical results—and in particular the observation that the ROC gains of multisplitting over single splitting disappear when considering the 5% cut-off—suggest that improved schemes for $p$-value aggregation may be useful.

Although the focus of this paper was not on applications, the methodology has immediate utility in biology. Differential network can be used to test a number of hypotheses of current scientific interest. For example, an emerging notion is that differences between disease subtypes may in some cases be manifested also at the level of regulatory patterns as encoded in biological networks. However, it has remained challenging to test such hypotheses in a principled way. Our approach can be used to test such hypotheses directly, as illustrated in the example that we presented. A further application of differential network is in gene set testing. Currently, gene set tests (Subramanian et al., 2005) are not truly multivariate. We have recently adapted our approach to test differences in gene sets at the level of not only means but also (gene-set-specific) covariance structure (Städler and Mukherjee, 2015).

7. Supplementary material

In the on-line supplementary document we provide details on the asymptotic theory of Section 3 and some aspects of the methodology. Additionally, we present an empirical analysis investigating the asymptotic properties of DiffRegr in a variety of settings. The R package nethet, incorporating DiffNet and DiffRegr, is available from http://bioconductor.org/packages/nethet.

Acknowledgements

The authors thank Peter Bühlmann, Hans Rudolf Künsch, Jacopo Mandozzi and Rajen Shah for discussions and comments and the Joint Editor and Associate Editor and referees for some insightful suggestions that greatly improved the paper, including inspiration for the ‘split-perm’ approach. This work was supported in part by the UK Medical Research Council. SM was a recipient of a Royal Society Wolfson Research Merit Award.
Appendix A: Proof of proposition 3

Let $U_1$, $U_2$ and $V$ be random vectors of arbitrary lengths. We write $B(U_1, U_2) = \mathbb{E}(U_1 U_2^T)$ and $B(U_1) = B(U_1, U_1)$. Further we set

$$Q(U_1, U_2|V) = B(U_1, U_2) - B(U_1, V)B(V)^{-1}B(V, U_2)$$

and

$$Q(U_1|V) = Q(U_1, U_1|V).$$

We consider the following auxiliary lemma which is a consequence of Schur’s complement and some matrix algebra.

**Lemma 3.** Let $\tilde{U}_1 = (V, U_1)$ and $\tilde{U}_2 = (V, U_2)$. Then we have

$$B(\tilde{U}_1, \tilde{U}_2)B(\tilde{U}_2)^{-1}B(\tilde{U}_2, \tilde{U}_1) = \begin{pmatrix} B(V) & B(U_1, V) \\ B(U_1, V) & Q(U_1, U_2|V)Q(U_2|V)^{-1}Q(U_2, U_1|V) + B(U_1, V)B(V)^{-1}B(V, U_1) \end{pmatrix}.$$ 

We now prove proposition 3. The eigenvalues of $W$ are solutions to

$$\det(W - wI) = \left[ -B(S^{\text{joint}}, S^{\text{ind}})B(S^{\text{joint}})^{-1} - 1_{d(12)} - w \right] = 0.$$ 

If $w \neq 1$, and setting $\tilde{w} = (1 - w^2)$ we obtain

$$\det(W - wI) = (1 - w)^{d(1)}d(2) \det \{ B(S^{\text{joint}}, S^{\text{ind}})B(S^{\text{joint}})^{-1} - \tilde{w}B(S^{\text{joint}}) \}.$$ (14)

The quantity on the right-hand side of equation (14) has $d(12)$ roots. Therefore, we conclude that $\dim(W) - 2d(12) = d(1) + d(2) - d(12)$ of the weights equal 1.

Denote $A := B(S^{\text{joint}}, S^{\text{ind}})B(S^{\text{ind}})^{-1}B(S^{\text{ind}}, S^{\text{joint}})$ and write

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}})B(S_{\text{ind}})^{-1} & B(S_{\text{ind}}, S_{\text{joint}}^{\text{ind}}) \\ B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}})B(S_{\text{ind}})^{-1} & 0 \end{pmatrix}.$$ 

By definition we have $A_{11} = -Q(S_{\text{joint}}^{\text{ind}}|S_{\text{ind}})$ and $B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}}^{\text{ind}})$. Now,

$$B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}}^{(1)})B(S_{\text{ind}}^{(1)})^{-1}B(S_{\text{ind}}^{(1)}, S_{\text{joint}}^{(1)}) = B(S_{\text{ind}}^{(1)}, S_{\text{ind}}^{(1)})B(S_{\text{joint}}^{(1)})^{-1}B(S_{\text{joint}}^{(1)}, S_{\text{ind}}^{(1)} + S_{\text{joint}}^{(2)}) = B(S_{\text{ind}}^{(1)}, S_{\text{joint}}^{(1)}B(S_{\text{ind}}^{(1)})^{-1}B(S_{\text{ind}}^{(1)}, S_{\text{joint}}^{(1)}) = B(S_{\text{ind}}^{(1)}).$$

where the second equality follows from independence of the samples $(Y^{(1)}, X^{(1)})$ and $(Y^{(2)}, X^{(2)})$ and the third equality from lemma 3. Similarly,

$$B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}}^{(2)})B(S_{\text{ind}}^{(2)})^{-1}B(S_{\text{ind}}^{(2)}, S_{\text{joint}}^{(2)}) = B(S_{\text{ind}}^{(2)}).$$

We therefore obtain

$$A_{22} = B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}}^{(2)})B(S_{\text{ind}}^{(2)})^{-1}B(S_{\text{ind}}^{(2)}, S_{\text{joint}}^{(2)}) + B(S_{\text{joint}}^{\text{ind}}, S_{\text{ind}}^{(2)})B(S_{\text{ind}}^{(2)})^{-1}B(S_{\text{ind}}^{(2)}, S_{\text{joint}}^{(2)})$$

$$= B(S_{\text{ind}}^{(1)}) + B(S_{\text{ind}}^{(2)}).$$

$$= B(S_{\text{ind}}^{(2)}).$$
where the last equality is due to independence of the two conditions. A similar argument gives $A_{12} = B(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{ind})$ and $A_{21} = B(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{joint})$. Putting together the foregoing results, we have

$$
\text{det} \{ B(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{ind}) B(S_{\text{in}}^\text{ind})^{-1} B(S_{\text{in}}^\text{ind}, S_{\text{in}}^\text{joint}) - \tilde{w} B(S_{\text{in}}^\text{joint}) \}
$$

$$
= \text{det} \left( -Q(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{ind}) + (1 - \tilde{w}) B(S_{\text{in}}^\text{ind}, S_{\text{in}}^\text{joint}) B(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{ind})^{-1} B(S_{\text{in}}^\text{ind}, S_{\text{in}}^\text{joint}) \right)
$$

$$
= (1 - \tilde{w})^d \text{det} \{-Q(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{ind}) + (1 - \tilde{w}) \{ B(S_{\text{in}}^\text{ind}, S_{\text{in}}^\text{joint}) B(S_{\text{in}}^\text{ind}, S_{\text{in}}^\text{joint})^{-1} B(S_{\text{in}}^\text{ind}, S_{\text{in}}^\text{joint}) \} \}
$$

$$
= (1 - \tilde{w})^d \text{det}(\tilde{W} - \tilde{w} I). \tag{15}
$$

From equations (14) and (15) we conclude that $2d^0$ eigenvalues are 0, $d^{(1)} + d^{(2)} - d^{(12)} - \{ d^{(12)} - d^0 - \text{rank}(W) \}$ equal 1, $d^{(12)} - d^0 - \text{rank}(W)$ are −1 and the remaining equal $\pm \sqrt{(1 - \tilde{w})}$, where $\tilde{w}_j, j = 1, \ldots, \text{rank}(W)$, are the non-zero eigenvalues of $W$.

From Lemma 3 and the derivation above we find that

$$
Q(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{ind}) = Q(S_{\text{in}}^{(1)}, S_{\text{in}}^{(1)}) - Q(S_{\text{in}}^{(1)}, S_{\text{in}}^{(2)}|S_{\text{in}}^{(1)}) - Q(S_{\text{in}}^{(2)}, S_{\text{in}}^{(1)}|S_{\text{in}}^{(1)}) - Q(S_{\text{in}}^{(1)}, S_{\text{in}}^{(2)}|S_{\text{in}}^{(1)})
$$

$$
+ Q(S_{\text{in}}^{(1)}, S_{\text{in}}^{(2)}|S_{\text{in}}^{(1)}) - Q(S_{\text{in}}^{(2)}, S_{\text{in}}^{(2)}|S_{\text{in}}^{(2)}) - Q(S_{\text{in}}^{(2)}, S_{\text{in}}^{(2)}|S_{\text{in}}^{(2)}) - Q(S_{\text{in}}^{(1)}, S_{\text{in}}^{(2)}|S_{\text{in}}^{(1)}) \tag{16}
$$

and

$$
Q(S_{\text{in}}^\text{joint}, S_{\text{in}}^\text{joint}) = Q(S_{\text{in}}^{(1)}, S_{\text{in}}^{(1)}|S_{\text{in}}^{(1)} + S_{\text{in}}^{(2)}|S_{\text{in}}^{(2)}). \tag{17}
$$

References


---

**Supporting information**

Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to “Two-sample testing in high dimensions”’.