High-dimensional statistics with a view towards applications in biology

Peter Bühlmann, Markus Kalisch and Lukas Meier
ETH Zürich

May 23, 2013

Abstract

We review statistical methods for high-dimensional data analysis and particular attention is given to recent developments for assessing uncertainties in terms of controlling false positive statements (type I error) and p-values. The main focus is on regression models but we also discuss graphical modeling and causal inference based on observational data. We illustrate the concepts and methods with various packages from the statistical software R for a high-throughput genomic data-set about riboflavin production with Bacillus subtilis, which we make the first time publicly available.

Keywords and phrases: Causal inference, Graphical modeling, Multiple testing, Penalized estimation, Regression

1 Introduction

High-dimensional statistical inference comes into play whenever the number p of unknown parameters is larger than sample size n: typically, we have in mind that p is an order of magnitude larger than n, denoted by $p \gg n$. Most often, we associate a setting where we have more (co-)variables than n, for example in a linear model

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{1}$$

with $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ a univariate response vector, $\mathbf{X}$ the $n \times p$ design matrix whose jth columns contains the covariable $\mathbf{X}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})^T$ and error (noise) term $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ with independent and identically distributed (i.i.d.) components having $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. An intercept term may be implicitly present.[1] Classical statistical methods, like ordinary least squares estimation, cannot be used for estimating β and σ^2 when $p \gg n$ because they would overfit the data, besides severe identifiability issues. A way out of the ill-posedness of estimation in model (1) is given

[1] In the later Sections 2 and 3, we typically would not penalize an intercept.

by assuming a *sparse* structure, typically saying that only few of the components of β are non-zero. We will review concepts for inference in the simple model (1) with $p \gg n$; the approaches can be generalized to more complex scenarios and models, see Section 4.

Many applications in biology nowadays involve high-dimensional data. Typically, high-throughput technology provides large-scale data of e.g. gene expressions (transcriptomics) or peptide and protein abundances (proteomics).

Example: riboflavin production with Bacillus subtilis
As a concrete example, we discuss a data-set about riboflavin (vitamin B_2) production in Bacillus subtilis. The data has been kindly provided by DSM (Kaiseraugst, Switzerland), see also Lee et al. (2001) and Zamboni et al. (2005), and for the first time, we make it publicly available in the Supplemental Section A.1. There is a single real-valued response variable which is the logarithm of the riboflavin production rate; furthermore, there are $p = 4088$ (co-)variables measuring the logarithm of the expression level of 4088 genes: these gene expressions were normalized using the default in the R-package `affy` (Gautier et al., 2004). There is one rather homogeneous data-set from $n = 71$ samples that were hybridized repeatedly during a fed-batch fermentation process where different engineered strains and strains grown under different fermentation conditions were analyzed. This data-set is denoted as `riboflavin` and we make it available (see Supplemental Materials). Another data-set consists of measurements (as above) at different time points (i.e. longitudinal data) with $N = 28$ groups each having 2 to 6 observations at different times. Observations in the same group are from measurements from the same strain of (genetically engineered) Bacillus subtilis while different groups correspond to different strains. The total number of samples is $n = 111$. This data-set is denoted as `riboflavinGrouped` and we make it available (see Supplemental Materials).

The easiest approach is to model the homogeneous riboflavin production data-set with a linear model as in (1) with $4088 = p \gg n = 71$, and this is discussed in Sections 2 and 3. Many questions in biology and other sciences are, however, about causal relationships among variables. They cannot be answered using a linear model as in (1) or extensions of it as presented in Section 4: we will present in Section 6 a method for causal statistical inference which can deal with high-dimensional scenarios.

2 Statistical estimation in a high-dimensional linear model

It is instructive to describe many concepts arising in high-dimensional statistical inference for linear models, as they are simple yet tremendously useful in many applications. Extensions to other regression-type models are discussed in Section 4, and remarks on the radically different marginal approach are given in Section 5.

Estimation of a high-dimensional linear model in (1) with $p \gg n$ requires some regularization. Common approaches include Bayesian or penalized likelihood methods. We largely focus on the latter. In the sequel, we implicitly assume that the (co-)variables

are all (at least roughly) on the same scale: very often, we achieve this by standardizing the columns of the design matrix such that $\|\mathbf{X}^{(j)}\|_2^2 = n$ for every $j = 1, \ldots, p$ (and often also $\sum_{i=1}^n \mathbf{X}_i^{(j)} = 0$ for all j). Ridge regression is defined as follows:

$$\hat{\beta}_{\text{Ridge}} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_2^2), \tag{2}$$

where $\|\cdot\|_2$ is the standard Euclidean norm, and $\lambda > 0$ a regularization parameter which has to be chosen by the user. The Lasso (Tibshirani, 1996) is replacing the ℓ_2-norm penalty by the ℓ_1-norm:

$$\hat{\beta}_{\text{Lasso}} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1), \tag{3}$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, and $\lambda > 0$ is again a regularization parameter (which typically is chosen differently as in (2)).[2] The estimators in (2) and (3) have a simple Bayesian interpretation in terms of maximum a-posteriori (MAP) procedures: assuming that $\beta_1, \ldots, \beta_p$ are i.i.d. with density $f(\cdot)$, the MAP can be easily derived:

$$\begin{aligned}
&\text{If } f(\cdot) \text{ is from } \mathcal{N}(0, \tau^2)\text{: } \hat{\beta}_{\text{MAP}} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \sigma^2/\tau^2\|\beta\|_2^2), \\
&\text{if } f(\cdot) \text{ is from DExp}(\tau)\text{: } \hat{\beta}_{\text{MAP}} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + 2\sigma^2\tau\|\beta\|_1),
\end{aligned}$$

where DExp(τ) is a Double-Exponential distribution with density $f(\beta) = \tau/2 \exp(-\tau|\beta|)$.

Both estimators in (2) and (3) are shrinking the coefficient estimates toward zero, due to the penalty which discourages large values. The Lasso has the special property to shrink some coefficients exactly to zero, because of the geometry of the ℓ_1-norm penalty: i.e., $\hat{\beta}_{\text{Lasso};j} = 0$ depending on the data and λ, and in this sense, the Lasso is doing variable selection. This can be best understood from equivalent optimization problems: the Lasso and Ridge estimators can be expressed as

$$\begin{aligned}
&\hat{\beta}_{\text{Ridge}} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n) \text{ under the constraint that } \|\beta\|_2 \le R, \\
&\hat{\beta}_{\text{Lasso}} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n) \text{ under the constraint that } \|\beta\|_1 \le R,
\end{aligned} \tag{4}$$

with a correspondence (depending on the data) between the value R and the value λ in (2) or (3), respectively. The representations in (4) have a geometric interpretation as displayed in Figure 1. Due to the form of the ℓ_1-norm ball with radius R, $\|\beta\|_1 \le R$, the optimum of the quadratic function $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ (represented by the contour lines in Figure 1) constrained to the set $\|\beta\|_1 \le R$ might occur in the corners of the set such that corresponding components of $\hat{\beta}_{\text{Lasso}}$ are equal to zero. Such a phenomenon does not happen for Ridge estimation.

Many versions of the Lasso have been proposed (Zou, 2006; Meinshausen, 2007; Zou and Li, 2008; van de Geer et al., 2011), and there are other penalized estimators which lead to sparse solutions (Fan and Li, 2001; Zhang, 2010). For further references, see for example Bühlmann and van de Geer (2011).

[2] We note that the factor $1/n$ is irrelevant from a methodological view point: when dropped, we simply use another λ which is n times the original one.

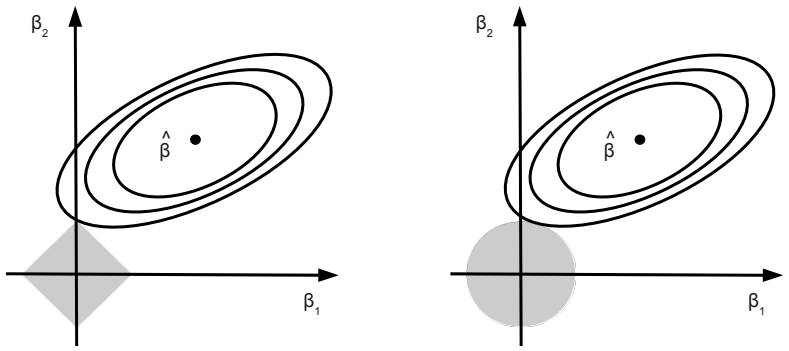

Figure 1: Constrained optimization as in (4) for $p = 2$. The contour lines of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ are shown as ellipses and $\hat{\beta}$ denotes the least squares estimator. Left: ℓ_1-norm constraint corresponding to the Lasso; Right: ℓ_2-norm constraint corresponding to Ridge estimation. The figure is essentially as in Tibshirani (1996) and taken from Bühlmann and van de Geer (2011).

2.1 Identifiability

If $p > n$, the model parameters in (1) are not identifiable because we always find a linear combination of the columns in $\mathbf{X}$ (corresponding to some covariables) which is exactly equal to one other column (one other covariable). Mathematically, the design matrix has not full rank, $\text{rank}(\mathbf{X}) \leq \min(n, p) < p$ for $p > n$, and we can write $\mathbf{X}\beta = \mathbf{X}(\beta + \xi)$ for every ξ in the null-space of $\mathbf{X}$.[3] Therefore, without further assumptions, it is impossible to infer or estimate β from data. We note that the issue is closely related to the classical setting with $p < n$ but $\text{rank}(\mathbf{X}) < p$ (due to linear dependence among covariables) or ill-conditioned design leading to difficulties with respect to identifiability. We note, however, that for prediction or estimation of $\mathbf{X}\beta$ (that is the underlying regression surface), identifiability of the parameters is not necessarily needed. From a practical point of view, high empirical correlations among two or a few other covariables lead to unstable results for estimating β or for pursuing variable selection. Some more details and additional references are given in the Supplemental Section A.2.

2.2 Point estimation without measures of uncertainty

Much progress has been made over the last decade for estimation without assigning uncertainty, confidence or error measures, i.e., so-called point estimation. It is important for further development of methods which quantify uncertainty (as discussed in Section 3). Among the three most important goals in such (point-) estimation are: (i) predicting the regression surface $\mathbf{X}\beta$ or a new response $Y_{\text{new}} = X_{\text{new}}^T\beta$; (ii) estimation of β; and (iii) estimation of the support of β or the so-called active set of relevant variables $S = \{j;\ \beta_j \neq 0\}$.

For the first task (i) of prediction, identifiability of β is not necessarily needed since we are only interested in e.g. $X_{\text{new}}^T\beta$. Thus, from this perspective, prediction is often a much easier problem than estimation of the parameter β or variable selection. Regarding task (ii) of parameter estimation, an identifiability assumption is required on the design $\mathbf{X}$, for example a restricted eigenvalue condition (see Supplemental Section A.2). Finally, for the task (iii) of variable selection, we would like to have an accurate estimator $\hat{S}$ for the active set S. A prime example is the Lasso where we simply use $\hat{S} = \{j;\ \hat{\beta}_{\text{Lasso},j} \neq 0\}$. Ideally, such an estimator would satisfy $\hat{S} = S$ with high probability. Unfortunately, such a property requires the rather strong "beta-min" condition saying that the non-zero regression coefficients must be sufficiently large

$$\min_{j \in S} |\beta_j| > C, \tag{5}$$

where C is typically of the order $\sqrt{\log(p)/n}$ (multiplied by $|S|$ or $\sqrt{|S|}$). Furthermore, e.g. for the Lasso, a stringent (so-called irrepresentability) condition on the design is necessary for variable selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). A less ambitious goal, which does not need such a strong assumption on the design, is

[3] The null-space of $\mathbf{X}$ is the set $\mathcal{N}_{\mathbf{X}} = \{\xi;\ \mathbf{X}\xi = 0\}$, and if $p > n$, the null-space contains other elements than the zero vector.

variable screening: it requires that, at least with high probability, $\hat{S}$ contains all variables from S, i.e.,

$$\hat{S} \supseteq S, \tag{6}$$

where $|\hat{S}|$ is typically much smaller than p. For example with the Lasso, $|\hat{S}| \leq \min(n, p) \ll p$ for high-dimensional settings. Thus, variable screening allows for a drastic dimension reduction in the original covariables which is often a useful first step for many practical applications. The screening property holds (with high probability) when the design is sufficiently well-behaved (i.e., the so-called compatibility condition holds[4]) and assuming the beta-min condition (5). Although variable screening is less ambitious than variable selection, the screening property in (6) is typically hard to be exactly fulfilled.

Reasonable performance of prediction and estimation can be achieved if the underlying truth is sparse. Among the most common notions of sparsity are the size of the active set $|S|$ (so-called ℓ_0-sparsity), but one can also imagine the ℓ_1-norm $\|\beta\|_1$ (or some other norms). If the sparsity is small in relation to sample size n and dimensionality p, then there is hope that some statistical methods exist which perform reasonably well. For example, a typical assumption of such kind is $|S| \ll n/\log(p)$ which shows that the dimensionality can be large as long as $\log(p) \ll n$ (allowing for reasonably large values of $|S|$). If the true underlying model is not sparse, high-dimensional statistical inference is ill-posed and not informative. Good statistical estimators for sparse situations should be sparse themselves. The Lasso (3) is a prime example, and many versions of the Lasso (see references just before Section 2.1) are often reasonable or even better, depending on the problem.[5]

Assessing the accuracy of prediction is relatively straightforward using the tool of cross-validation (Hastie et al., 2009, cf.). Some earlier work points to inaccuracy of cross-validation for measuring the out-of-sample error (Gasser et al., 1991): still, assessing the quality of prediction, e.g. with cross-validation, is a much easier task than measuring the accuracy of parameter estimation, variable selection or screening. Regarding the latter, the traditional thinking in frequentist statistics follows the framework of hypothesis testing where false positive selections, corresponding to type I error, are considered to be worse than false negatives, corresponding to type II error. The challenge in high-dimensional models is the construction of p-values which control some type I error measure while having good power for detecting the alternatives (i.e. avoiding some type II error). This will be discussed in Section 3.

2.2.1 Software in `R`

We use the `R`-Package `glmnet` (Friedman et al., 2010) to illustrate the Lasso estimator on the `riboflavin` data-set.

[4]The compatibility condition is weaker than the irrepresentability condition mentioned in connection with variable selection (van de Geer and Bühlmann, 2009).

[5]Also Ridge estimation (2) can be sparsified by thresholding the estimated coefficients of $\hat{\beta}$, dropping the corresponding covariables and doing, say, a least-squares re-estimation based on fewer variables kept.

```
library(glmnet)

x <- riboflavin[,-1]
y <- riboflavin[,1]

## Check dimensions
dim(x)
##- [1]   71 4088
length(y)
##- [1] 71

## Fit whole solution path for illustration
fit <- glmnet(x=x, y=y)
plot(fit)

## Perform 10-fold cross-validation
set.seed(42)
fit.cv <- cv.glmnet(x=x, y=y)

## Visualize cross-validation error-path
plot(fit.cv)

## Get selected genes
b <- as.matrix(coef(fit.cv))
rownames(b)[b != 0]
## By default, the selected variables are based on the largest value of
lambda such that the cv-error is within 1 standard error of the minimum
```

The resulting model contains 30 genes (plus an intercept term) with corresponding estimated regression coefficients different from zero.

3 Assigning uncertainty in high-dimensional linear models

For the linear model (1), we are interested in two-sided testing of individual hypotheses $H_{0,j}: \beta_j = 0$ versus $H_{A,j}: \beta_j \neq 0$ or corresponding confidence intervals; and we also might consider hypotheses concerning a group of parameters $H_{0,G}: \beta_j = 0$ for all $j \in G$ versus $H_{A,G}: H_{0,G}^c$ (that is, at least one $\beta_j \neq 0$ for some $j \in G$). In addition, we aim for an accurate and not overly conservative correction for multiplicity of testing.

3.1 Why standard bootstrapping and subsampling do not work

As discussed above in Section 2.2, we typically use sparse estimators for high-dimensional data analysis, for example the Lasso; for an exception see Section 3.2.2. The (limiting) distribution of such a sparse estimator is non-Gaussian with point mass at zero, and this is the reason why standard bootstrap or subsampling techniques do not provide

valid confidence regions or p-values. Thus, we have to use other approaches to quantify uncertainty.

3.2 P-values for high-dimensional linear models

3.2.1 Multi sample-splitting

A very generic method for constructing p-values for $H_{0,j}$ or $H_{0,G}$ is based on splitting the sample into two equal parts, where we select the variables using the first and do the statistical inference based on the second half of the data. Such a sample-splitting avoids overly optimistic results based on selecting variables and doing subsequent inference for the selected variables (both based on the full data-set) as if no other variables were present.

To fix ideas, consider the following scheme for multiple testing of $H_{0,j} : \ \beta_j = 0$ among all $j = 1, \ldots, p$. Thereby, we aim for controlling the familywise error rate (FWER) $\mathbb{P}[V > 0]$, where V is the number of false positives. [6]

Algorithm 1 Single sample-splitting for multiple testing of $H_{0,j}$ among $j = 1, \ldots, p$:

1: Split the sample $\{1, \ldots, n\} = I_1 \cup I_2$ with $I_1 \cap I_2 = \emptyset$ and $|I_1| = \lfloor n/2 \rfloor$ and $|I_2| = n - \lfloor n/2 \rfloor$.

2: Based on I_1, select the variables $\hat{S} \subseteq \{1, \ldots, p\}$. Assume (or ensure) that $|\hat{S}| \leq |I_1| = \lfloor n/2 \rfloor \leq |I_2|$.

3: Consider the reduced set of variables with design matrix $\mathbf{X}^{(\hat{S})}$. Based on I_2 with data $(\mathbf{Y}_{I_2}, \mathbf{X}_{I_2}^{(\hat{S})})$, compute p-values P_j for $H_{0,j}$, for $j \in \hat{S}$, from classical least squares estimation assuming Gaussian errors (i.e. t-test which is well-defined since $|\hat{S}| \leq |I_2|$). For $j \notin \hat{S}$, assign $P_j = 1$.

4: Correct the p-values for multiple testing: consider

$$P_{\mathrm{corr},j} = \min(P_j \cdot |\hat{S}|, 1)$$

which is an adjusted p-value for $H_{0,j}$ for controlling the familywise error rate.

The procedure described in Algorithm 1 yields corrected p-values which control the FWER, when assuming the screening property in (6): the whole idea is implicitly contained in the work by Wasserman and Roeder (2009). In practice, the screening property typically does not hold exactly but it is not a necessary condition for constructing valid p-values (Bühlmann and Mandozzi, 2013). We also note that the correction for multiplicity of testing in Step 4 only involves the multiplicative factor $|\hat{S}|$, while a classical Bonferroni-adjustment would multiply the p-values with p: in high-dimensional scenarios, $p \gg n > |\hat{S}|$, and thus, the correction factor employed here is rather small.

[6] A false positive arises when the test-procedure rejects $H_{0,j}$ although $H_{0,j}$ in fact holds true.

A major difficulty of the single sample-splitting method is its sensitivity coming from the choice of how one splits the sample, leading to widely different corresponding p-values. Figure 2 illustrates such a "p-value lottery" phenomenon.

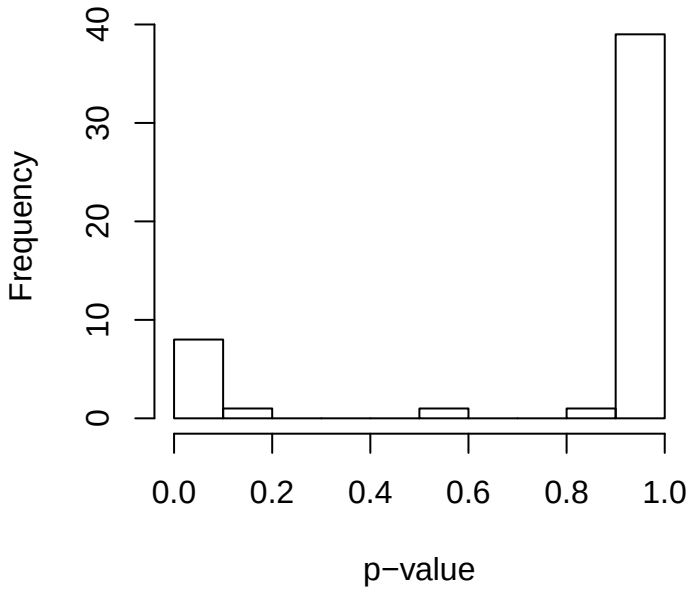

Figure 2: Histogram of p-values $P_{\mathrm{corr},j}$ for a single covariable, in the `riboflavin` data-set, when doing 50 different (random) sample splits.

To overcome this undesirable behavior, one can run the single sample-splitting Algorithm 1 B times, with B large, leading to p-values

$$P_{\mathrm{corr},j}^{[1]}, \ldots, P_{\mathrm{corr},j}^{[B]} \ (j = 1, \ldots, p).$$

The remaining task is then to aggregate these $\{P_{\mathrm{corr},j}^{[b]};\ b = 1, \ldots, B\}$ to a single p-value. Due to dependence among the $P_{\mathrm{corr},j}^{[b]}$'s (since all the different split samples are based on the same full data-set), such an aggregation should be done carefully.[7] A simple but effective solution is to use an empirical γ-quantile:

$$Q_j(\gamma) = \min\left(\text{emp. } \gamma\text{-quantile}\{P_{\mathrm{corr},j}^{[b]}/\gamma;\ b = 1, \ldots, B\}, 1\right).$$

For example, when taking $\gamma = 1/2$, we multiply all $P_{\mathrm{corr},j}^{[b]}$'s by 2 and take the empirical median among them. Furthermore, one can optimize over the best γ-quantile in the

[7] For example, the mean $B^{-1}\sum_{b=1}^{B} P_{\mathrm{corr},j}^{[b]}$ is generally not controlling the FWER.

range $(\gamma_{\min}, 1)$ (for example with $\gamma_{\min}$ equal to 0.05) leading to the aggregated p-value

$$P_j = \min\left((1 - \log(\gamma_{\min})) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)\right) \ (j = 1, \ldots, p). \tag{7}$$

Thereby, the factor $(1 - \log(\gamma_{\min}))$ is the price to be paid for searching for the best $\gamma \in (\gamma_{\min}, 1)$. This multi sample-splitting procedure has been proposed by Meinshausen et al. (2009) and is summarized in Algorithm 2. The multi sample-splitting method

Algorithm 2 Multi sample-splitting for multiple testing of $H_{0,j}$ among $j = 1, \ldots, p$

1: Run the single sample-splitting Algorithm 1 B times leading to p-values $\{P^{[b]}_{\mathrm{corr},j};\ b = 1, \ldots, B\}$. A typical choice is $B = 100$.
2: Aggregate the p-values from Step 1 as in (7) leading to P_j which are adjusted p-values for $H_{0,j}$ $(j = 1, \ldots, p)$, controlling the familywise error rate.

enjoys the property that the resulting p-values are approximately reproducible and not subject to a "lottery" as illustrated in Figure 2, and it controls the familywise error rate. As the single sample-split method, the procedure assumes the screening property (6) (or an approximate version of it). More precise mathematical assumptions for constructing valid p-values are given in Supplemental Section A.4.2.

Testing group hypotheses of the form $H_{0,G}: \ \beta_j = 0$ for all $j \in G$ can be done based on a partial F-test instead of a t-test in Step 3 of Algorithm 1.

3.2.2 Projection and confidence intervals

The multi sample-splitting method assumes (a possibly relaxed form of) the screening property (6), and this in turn necessarily requires a (possibly relaxed) beta-min assumption (5). The methods described here do not rely on such a beta-min assumption.

The general idea is to use a *linear* estimator with subsequent bias correction using the Lasso. Consider for each $j = 1, \ldots, p$ an $n \times 1$ vector $Z^{(j)}$[8] and corresponding estimator

$$\hat{b}_j = \frac{(Z^{(j)})^T Y}{(Z^{(j)})^T X^{(j)}}.$$

Then, by simply using the linear relation between Y and $\{X^{(k)};\ k = 1, \ldots, p\}$ we obtain

$$\mathbb{E}[\hat{b}_j] = \beta_j + \sum_{k \neq j} P_{jk}\beta_k, \quad P_{jk} = \frac{(Z^{(j)})^T X^{(k)}}{(Z^{(j)})^T X^{(j)}}.$$

The second summand is a bias term which can be corrected using the Lasso, and we then obtain the bias-corrected estimator

$$\hat{\beta}_{\mathrm{corr};j} = \hat{b}_j - \sum_{k \neq j} P_{jk}\hat{\beta}_{\mathrm{Lasso};k}. \tag{8}$$

[8] Typically $Z^{(j)}$ is a residual vector when doing a regularized regression of $X^{(j)}$ versus all other variables $\{X^{(k)}; k \neq j\}$), see also Supplemental Section A.3.

Concrete suggestions for score vectors $Z^{(j)}$ are based on Ridge regression (Bühlmann, 2013) or on Lasso regression (Zhang and Zhang, 2011; van de Geer et al., 2013). More details are given in the Supplemental Section A.3. The corresponding estimators in (8) using Ridge or Lasso-based score vectors are denoted by

$$\hat{\beta}_{\text{corr}-\text{Ridge}}, \quad \hat{\beta}_{\text{corr}-\text{Lasso}}.$$

We point out the interesting feature that these estimators $\hat{\beta}_{\text{corr}-\text{Ridge}}$ or $\hat{\beta}_{\text{corr}-\text{Lasso}}$ are not sparse and they have a Gaussian limiting distribution with known covariance matrix, except for the unknown error variance σ_ε^2 (Bühlmann, 2013; Zhang and Zhang, 2011; van de Geer et al., 2013). The latter can be estimated, for example using the scaled Lasso (Sun and Zhang, 2012). We then end up with a statement of the form

$$\sqrt{n}(\hat{\beta}_{\text{corr},j} - \beta_j)/\hat{\sigma}_j \to \mathcal{N}(0,1), \tag{9}$$

where $\hat{\sigma}_j^2 = \hat{\sigma}_\varepsilon^2 \omega_j$ with known ω_j (which is easily computable as function of the design $\mathbf{X}$). As a consequence, we can derive confidence intervals and tests for single parameters β_j, and we can also construct p-values for $H_{0,G}: \ \beta_j = 0$ for all $j \in G$, where $G \subseteq \{1, \ldots, p\}$ is any group (small or large).

Assuming sparsity of the regression vector, but *without* requiring a beta-min assumption as in (5), the method provides valid inference for tests and confidence intervals. When using $\hat{\beta}_{\text{corr}-\text{Lasso}}$, the procedure is optimal and reaches the semiparametric efficiency bound (van de Geer et al., 2013). More precise mathematical assumptions for valid p-values and asymptotic optimality are given in Supplemental Section A.4.3.

When pursuing many tests, we have to adjust for multiple testing. Consider first the scenario when testing $H_{0,j}: \ \beta_j = 0$ for all $j = 1, \ldots, p$. We then obtain p-values $P_1, \ldots, P_p$, and we can use any multiple testing adjustment which is valid for dependent tests (note that $P_1, \ldots, P_p$ are dependent). For example, we can use the Bonferroni-Holm method (Holm, 1979) to control the familywise error rate, or we can use a version of the standard Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) which controls the false discovery rate among dependent tests (Benjamini and Yekutieli, 2001), see also Section 5.1. The R-software package `multtest` provides an array of methods for multiple testing correction, see Section 5.1.1. However, the p-values $P_1, \ldots, P_p$ typically exhibit rather strong dependence which implies that the usual adjustment methods are too conservative.[9] A potential loss of power by avoiding conservative adjustment can be addressed by exploiting the known covariance structure of the problem: more generally than in (9) we have

$$\sqrt{n}(\hat{\beta}_{\text{corr}} - \beta) \approx \mathcal{N}_p(0, \hat{\sigma}_\varepsilon^2 \Omega),$$

where Ω is known. Such a representation allows for efficient adjustment of the p-values $P_1, \ldots, P_p$ (Bühlmann, 2013). We emphasize that such a multiple testing correction can

[9] As an extreme case, suppose that the data for each hypothesis and thus the p-values are the same $P_1 = P_2 = \ldots = P_p$: then the effective number of tests is 1 (instead of the nominal number of tests p) and adjusting the p-values wouldn't be necessary.

be used more generally for m tests with p-values $P_1, \ldots, P_m$ where each P_r corresponds to a hypothesis test $H_{0,G_r}: \ \beta_j = 0$ for all $j \in G_r \subseteq \{1, \ldots, p\}$ (and each G_r can be a small group (e.g. having one element only) or a large group).

3.2.3 Software in R

We use our own R-package `hdi` (Meier, 2013) to analyze the `riboflavin` data-set. The multi-split method yields one significant gene (gene `YXLD_at`), while the Ridge-type projection estimator delivers no significant gene at all.

```
library(hdi)

x <- riboflavin[,-1]
y <- riboflavin[,1]

## Multi-split p-values
set.seed(12)
fit.multi <- hdi(x, y, method = "multi-split", B = 100)
fit.multi

## Ridge p-values
fit.ridge <- hdi(x, y, method = "pval-ridge")
fit.ridge
```

3.3 Stability selection

Stability selection (Meinshausen and Bühlmann, 2010) is a method based on subsampling (or bootstrapping) but rather different from classical approaches. Consider a random subsample $I^* \subset \{1, \ldots, n\}$ of size $|I^*| = \lfloor n/2 \rfloor$. For any variable selection algorithm $\hat{S} \subseteq \{1, \ldots, p\}$, e.g. the Lasso, we consider its subsampled version $\hat{S}(I^*)$ based on the subsample I^*. The subsampled relative selection frequencies are then

$$\hat{\pi}_j = \mathbb{P}^*[j \in \hat{S}(I^*)], \ j = 1, \ldots, p,$$

where $\mathbb{P}^*$ is with respect to the subsample I^*. In practice, this is approximated by a stochastic simulation

$$\hat{\pi}_j \approx B^{-1} \sum_{b=1}^{B} I\left(j \in \hat{S}(I^{*(b)})\right)$$

where $B \approx 500 - 1000$ is large and $I^{*(1)}, \ldots, I^{*(B)}$ are independent random subsamples of size $|I^{*(b)}| = \lfloor n/2 \rfloor$. The set of stable variables is defined as

$$\hat{S}_{\text{stable}} = \{j; \ \hat{\pi}_j \geq \pi_{\text{thres}}\},$$

for some threshold parameter π_{thres}.

The threshold parameter π_{thres} can be linked to some type I error measure about false positive selections. For this purpose we assume that $\hat{S}(I)$ selects at most q variables for every subsample $I \subset \{1, \ldots, n\}$ with $|I| = \lfloor n/2 \rfloor$. As examples we mention the Lasso which selects the q variables entering the regularization path first, as used in Section 3.3.1 below; or the Lasso selecting the top q variables having highest estimated regression coefficients in absolute value. Furthermore, if such an $\hat{S}$ is better than random guessing and if a so-called exchangeability condition holds (which becomes an assumption on the design, implying that false positive selection of any variable is equally likely), we have the following relation: denoting by V the number of false positives,

$$\mathbb{E}[V] \leq \frac{q^2}{(2\pi_{\text{thres}} - 1)p}, \tag{10}$$

see Meinshausen and Bühlmann (2010). Therefore, by pre-specifying that $\mathbb{E}[V]$ should be at most `efp` (say `efp` $= 1$), and assuming `efp` $\geq q^2/p$, we would choose

$$\pi_{\text{thres}} = \frac{1}{2} + \frac{q^2}{2p \cdot \texttt{efp}}$$

which ensures by (10) that the corresponding $\mathbb{E}[V] \leq$ `efp`. The work in Shah and Samworth (2013) extends the result in (10) without requiring the restrictive but not necessary exchangeability assumption.

We note that stability selection can also be used for whole groups $G \subseteq \{1, \ldots, p\}$, instead of single variables $j \in \{1, \ldots, p\}$. For example, in the spirit of a group null-hypothesis $H_{0,G} : \ \beta_j = 0$ for all $j \in G$ and the complementary alternative $H_{0,G}^c$, we would consider the stability that at least one element in a group G has been selected: this is formalized as

$$\hat{\pi}_G = \mathbb{P}^*[G \cap \hat{S} \neq \emptyset].$$

The error bound (10) needs to be adapted by replacing p with $\binom{p}{k}$ and q^2 with $\binom{q}{k}^2$, where k is the group size $|G| = k$ (e.g. considering $k = 2$ for selecting stable groups of variables of size 2).

The beauty of stability selection is its generic applicability to any problem about discrete structure estimation: that is, the selection algorithm $\hat{S}$ does not need to be for variable selection in a linear model but it could for example encode the selection of an edge in a graphical model. Furthermore, in a linear model, we do not need to explicitly estimate the error variance. However, we do not directly obtain p-values for statistical hypothesis testing. More precise mathematical assumptions for the error control as in (10) are given in Supplemental Section A.4.4.

3.3.1 Software in `R`

Again, we use the `R`-package `hdi` (Meier, 2013) to run stability selection on the `riboflavin` data-set. As selector $\hat{S}$, we use the Lasso with the q variables which enter the regular-

ization path[10] first. With $q = 20$, $V = 1$ and $B = 500$ we get 3 stable selected genes: `LYSC_at`, `YOAB_at` and `YXLD_at`.

```
library(hdi)

x <- riboflavin[,-1]
y <- riboflavin[,1]

set.seed(37)
fit.stab <- hdi(x, y, method = "stability", B = 500, EV = 1, q = 20)
fit.stab
```

3.4 Summary of linear model results for riboflavin data-set

For the `riboflavin` data-set with $n = 71$ and $p = 4088$, the results from the different methods vary to a certain extent. The multi sample-splitting Algorithm 2 and the projection estimator (8), here used with Ridge-type score vectors, lead to p-values controlling the very stringent familywise error rate (FWER). At the FWER-adjusted 5% significance level, we find 1 significant variable (gene `YXLD_at`) based on the multi sample-splitting Algorithm 2, while the projection estimator doesn't find a single significant variable or gene. This finding is not surprising: the Ridge-type projection estimator is rather conservative and since it does not require a beta-min assumption, its power for rejection is typically smaller, i.e., it produces typically larger p-values.

Stability selection finds more relevant variables. However, the corresponding error measure is only the expected number of false positive selections $\mathbb{E}[V]$: such an error measure is much less stringent than FWER. Furthermore, the one significant gene found with the multi sample-splitting Algorithm 2 has largest selection frequency in the stability selection approach.

4 Extensions to other models

Much of the work on point estimation carries over from high-dimensional linear models to more complex models. For assigning statistical uncertainties, the multi sample splitting method and stability selection are straightforward to be used for non-linear models while the projection procedure from Section 3.2.2 needs more careful treatment.

4.1 Generalized linear models

Generalized linear models (McCullagh and Nelder, 1989) are very popular for extending the linear model in a unified way. We consider a model with univariate response Y and

[10] The paths (p functions) of estimated coefficients from Lasso $\hat{\beta}_j(\lambda)$ ($j = 1, \ldots, p$) when varying λ from a maximal value to 0^+.

p-dimensional covariables X:

$$\begin{aligned} &Y_1, \ldots, Y_n \text{ independent,} \\ &g(\mathbb{E}[Y_i|X_i = x]) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)}, \end{aligned} \tag{11}$$

where $g(\cdot)$ is a real-valued, known link function, μ is the intercept term and $x^{(j)}$ denotes the jth component of the p-dimensional x. A well-known example is logistic regression for binary response variables $Y_i \in \{0, 1\}$: we denote by $\pi(x) = \mathbb{P}[Y_i = 1|X_i = x](= \mathbb{E}[Y_i|X_i = x])$, and the model employs the logistic link function $g(\pi) = \log(\pi/(1-\pi))$ which maps $(0, 1)$ to the real line. Another example is Poisson regression for count data responses: $Y_i|X_i = x \sim$ Poisson$(\lambda(x))$ and the employed link function is $g(\lambda) = \log(\lambda)$ which maps $\mathbb{R}^+$ to the real line. Obviously, a linear model is a special case of a generalized linear model with the identity link function $g(\theta) = \theta$.

An implicit assumption of the model in (11) is that the (conditional) distribution of Y_i (given X_i) is depending on X_i only through the function $g(\mathbb{E}[Y_i|X_i]) = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)}$. That is, the (conditional) probability or density of $Y|X = x$ is of the form

$$\mathrm{p}(y|x) = \mathrm{p}_{\mu,\beta}(y|x). \tag{12}$$

For generalized linear models, the analogue of the Lasso estimator in (3) is defined by penalizing the negative log-likelihood with the ℓ_1-norm. The negative log-likelihood itself equals

$$-\ell(\mu, \beta;\ \text{data}) = -\sum_{i=1}^{n} \log(\mathrm{p}_{\mu,\beta}(Y_i|X_i)),$$

where $\mathrm{p}_{\mu,\beta}(y|x)$ is as in (12). For many examples and models, e.g. if the (conditional) distribution of $Y|X = x$ is from a sub-class of the exponential family model (see McCullagh and Nelder (1989, Section 2.2)), the negative log-likelihood $\ell(\mu, \beta;\ \text{data})$ is convex in μ, β for all values of the data. Such convexity is not a necessary requirement for ℓ_1-norm penalization introduced below (see for example Section 4.2) but it enables efficient optimization and more elegant mathematical analysis of the property of the estimator. The ℓ_1-norm penalized Lasso estimator is then defined as:

$$\hat{\mu}(\lambda), \hat{\beta}(\lambda) = \arg\min_{\mu,\beta}(-\ell(\mu, \beta;\ \text{data})/n + \lambda\|\beta\|_1) \tag{13}$$

$$= \arg\min_{\mu,\beta}(-n^{-1}\sum_{i=1}^{n} \log(\mathrm{p}_{\mu,\beta}(Y_i|X_i)) + \lambda\|\beta\|_1). \tag{14}$$

Usually, we do not penalize the intercept term.

Similarly to the Lasso (3) for high-dimensional linear models, analogous assumptions are required for estimation of $X_{\text{new}}^T\beta$, for estimation of β and for the active set $S = \{j;\ \beta_j \neq 0\}$: except for estimation of $X_{\text{new}}^T\beta$, we need identifiability assumptions on the

design $\mathbf{X}$ and a condition for the smallest non-zero regression coefficients (as in (5)), see for example van de Geer (2008) and Bühlmann and van de Geer (2011).

Assigning measures of uncertainty and significance can be easily done using the multi sample-splitting method from Section 3.2.1 or stability selection described in Section 3.3. Regarding the former, we can use $\hat{S}$ from the penalized estimator in (13), replace the t-test in Step 3 of Algorithm 1 with the log-likelihood ratio test (McCullagh and Nelder, 1989) and then proceed as in Algorithm 2. For stability selection, we could use $\hat{S}$ from e.g. the ℓ_1-norm penalized maximum likelihood estimator in (13); a related idea with applications to genome-wide association studies is presented in He and Lin (2011). The method based on projection estimators in Section 3.2.2 needs a more elaborate extension and is described in van de Geer et al. (2013) for high-dimensional generalized linear models.

The estimator in (13) can be computed using the `R`-package `glmnet` (Friedman et al., 2010), analogously as in Section 2.2.1.

4.2 Generalized linear mixed models

Mixed effects models are popular for modeling grouped or longitudinal data (Pinheiro and Bates, 2000): the building blocks are fixed effects with corresponding p-dimensional parameter vector β and random effects with corresponding random parameter $b \sim \mathcal{N}_q(0, V)$. From a frequentist point of view, the unknown parameters in the model are β, V and possibly an error variance σ_ε^2.

The high-dimensional scenario typically refers to the case where p is large while the dimension of the covariance matrix V is small (q might still be large but V would have a low-dimensional parameterization). In such a setting, one can again use the ℓ_1-norm penalized maximum likelihood estimator: similarly as in (13), we consider[11]

$$\hat{\beta}(\lambda), \hat{V}(\lambda), \hat{\sigma}_\varepsilon^2(\lambda) = \underset{\beta, V, \sigma_\varepsilon^2}{\arg\min}(-\ell(\beta, V, \sigma_\varepsilon^2;\ \text{data}) + \lambda\|\beta\|_1). \tag{15}$$

The difficulty of this estimator is the fact that the negative log-likelihood $-\ell(\beta, V, \sigma_\varepsilon^2;\ \text{data})$ is a non-convex function in the unknown parameters and in case of non-Gaussian (e.g. generalized linear) mixed models, already the likelihood is difficult to compute. The latter can be addressed by numerical approximations, for example using the Laplace approximation as used in Schelldorfer et al. (2013); the former causes generic computational difficulties as well as more subtle conditions and arguments to establish good properties of the estimator as discussed in Schelldorfer et al. (2011) for Gaussian linear mixed models.

Similarly as for generalized linear models, we can use the multi sample-splitting method from Section 3.2.1 or stability selection described in Section 3.3 for quantifying uncertainties. For the former, we can use the screening method from the penalized estimator in (15), the t-test in Step 3 of Algorithm 1 has to be replaced by a valid procedure

[11] If not in the model, σ_ε^2 should be dropped in the following expressions.

for low-dimensional generalized linear mixed models, and we can then proceed with Algorithm 2. For stability selection for the fixed effects variables, we can use $\hat{S}$ from (15).

Example: grouped data about riboflavin production with Bacillus subtilis
One data-set of riboflavin production with Bacillus subtilis (see also Section 1) consists of measurements ($p = 4088$ gene expressions and the riboflavin production rate) at different time points (longitudinal data) with $N = 28$ groups each having 2 to 6 observations at different times, and the total number of samples is $n = 111$. We refer to the *Example* in Section 1 for further description of the data-set which is denoted as `riboflavinGrouped` and we make it available (see Supplemental Materials).

We can fit a linear mixed model to this data with the 28 different groups. After some preliminary analysis, a reasonable model consists of 2 independent random effects and $p = 4088$ fixed effects. The estimator in (15) with Gaussian distribution can be computed with the `R`-package `lmmlasso` (Schelldorfer, 2011): the results are presented in Schelldorfer et al. (2011).

Computation of the estimator in (15) for generalized mixed effects models can be done with the `R`-package `glmmixedlasso` (Schelldorfer et al., 2013) which is available from `R`-Forge.

4.3 Gaussian graphical models

As a further extension of linear models we mention the Gaussian graphical model:

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}_p(0, \Sigma).$$

Assuming that Σ^{-1} exists, we represent the p-dimensional Gaussian distribution in terms of a graph with a set of nodes or vertices $\{1, \ldots, p\}$ and a set of undirected edges defined as:

there is an undirected edge between node j and k if and only if $\Sigma^{-1}_{jk} \neq 0$.

The distribution then obeys a local and global Markov property with respect to the defined graph (Lauritzen, 1996) and hence, the edges can be interpreted in terms of conditional dependence statements:

there is an undirected edge between node j and k if and only if
$X^{(j)}$ and $X^{(k)}$ are conditionally dependent given all other variables $\{X^{(\ell)};\ \ell \neq j, k\}$.

Estimation of such a graph in the high-dimensional scenario can be done with a nodewise Lasso approach (Meinshausen and Bühlmann, 2006) which is computationally efficient and requires slightly weaker conditions than the ℓ_1-norm penalized maximum likelihood estimation scheme, also called graphical Lasso, or GLasso (Friedman et al., 2007; Banerjee et al., 2008). Assigning uncertainties could be done using the multi sample-splitting

method from Section 3.2.1. In view of the many edges and the multivariate nature of the model, stability selection has been advocated in Meinshausen and Bühlmann (2010).

Extensions for non-Gaussian continuous distributions, based on copula models, are given by Liu et al. (2012) and Xue and Zou (2012); this is exemplified in Section 4.3.1 below using the so-called nonparanormal transformation. Undirected graphical model estimation for the case with mixed-type binary, categorical and continuous variables is considered in Fellinghauer et al. (2013).

4.3.1 Software in `R`

Two major packages dealing with estimation of undirected graphs are `huge` (Zhao et al., 2012) and `glasso` (Friedman et al., 2011). Since `huge` seems to be more elaborate, we only report using this package: for the riboflavin production data, we estimate an undirected graph by the Meinshausen-Bühlmann method (Meinshausen and Bühlmann, 2006) and select the regularization parameter using a variant of stability selection in Section 3.3 termed StARS (Liu et al., 2010). As an illustration and for simplicity (without deeper biological implications), we estimate the undirected graph for the 100 genes with largest empirical variance and the riboflavin production, and we denote this reduced data-set by `riboflavinV100` (the fitting and selection process on the complete data set takes about 2 hours CPU). The resulting graph is shown in Figure 3. We refer the interested reader to the vignette of the `huge` package for more details on the use of this package.

```
library(huge)
set.seed(123)
## For ease of reproduction, we only use the 100 genes
## with largest empirical variance
## The analysis on the full data takes about 2 hours

## Apply nonparanormal transformation
X.npn <- huge.npn(riboflavinV100)

## Estimate undirected graph
out.npn <- huge(X.npn, method = "mb", nlambda=30)

## Select the graph using StARS
npn.stars <- huge.select(out.npn,criterion="stars",stars.thresh=0.05)

## Extract optimal graph
resGraph <- npn.stars$refit

## Plot graph
huge.plot(resGraph)
```

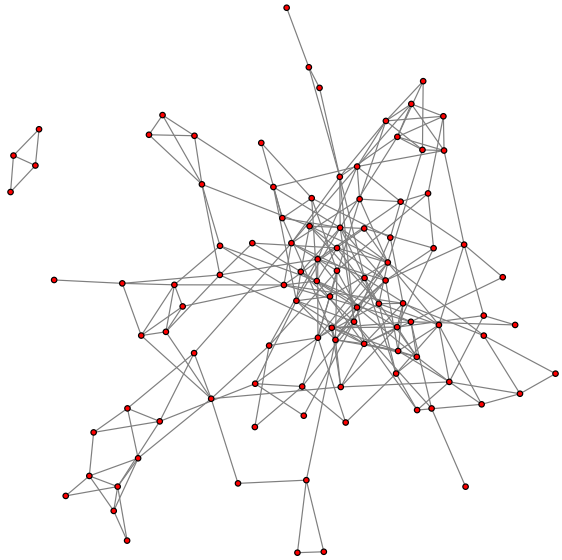

Figure 3: Estimated undirected graph for the `riboflavinV100` data-set. For ease of reproduction, only the 100 genes with largest empirical variance and the amount of riboflavin produced were included in the estimation process. The graph shown was estimated by the Meinshausen-Bühlmann method, after the nonparanormal transformation, and regularized using the StARS criterion.

5 The marginal approach

When having response (or grouping or class) variables Y_i and p-dimensional (co-)variables $X_i = X_i^{(1)}, \ldots, X_i^{(p)}$, with (Y_i, X_i) $(i = 1, \ldots, n)$ independent and identically distributed, the target of interest might be marginal associations between Y and $X^{(j)}$ $(j = 1, \ldots, p)$. For example, marginal association measures are correlations between Y and $X^{(j)}$ or regression parameters α_j in the model $Y = \mu + \alpha_j X^{(j)} +$ noise. Such marginal association parameters are very different from the parameters in e.g. a linear model as in (1) or more general regression models: the latter measure the strength of association which is not explained by all other variables. There are some recent attempts for variable screening in a linear model (1) as in (6) based on marginal correlations. Under some rather strong conditions on the design matrix, the proposed methods provide a superset of S as in (6) (Fan and Lv, 2008; Genovese et al., 2012); an extension of such a purely marginal approach is discussed in Bühlmann et al. (2010).

The dimension p of the (co-)variables X_i is not really a disturbing issue when estimating marginal association parameters, even if $p \gg n$.[12] The only drawback comes in when adjusting tests (and confidence intervals) with respect to multiplicity, especially when considering all $p \gg n$ marginal associations.

Genome-wide association studies (GWAS) are examples where oftentimes, only marginal associations are considered. For example, if Y_i is binary encoding healthy and diseased status of an individual and $X_i^{(j)}$ a categorical variable with three levels describing a single-nucleotide polymorphism (SNP) at position j in the genome, we obtain p-values from two-sample tests (the two samples are encoded by the binary response) for a location shift at each genomic position $j = 1, \ldots, p$. A typical value of p is $\approx 10^6$ while sample size is in the hundreds or low thousands.

5.1 Multiple testing adjustment

In the GWAS example above we have p-values $P_1, \ldots, P_p$ where p is very large, and adjusting for multiplicity is crucial (Roeder and Wasserman, 2009). Common type I error measures for multiple testing are the familywise error rate (FWER; the probability of at least one false positive selection) or the false discovery rate (FDR; the proportion of false positive selections among the significant tests). The Bonferroni-Holm procedure (Holm, 1979) leads to FWER control under any dependence structure among the tests, and due to such generality, the method is often overly conservative; the Westfall-Young method (Westfall and Young, 1989) offers an alternative, at least for some cases, which often has better power (Meinshausen et al., 2011). The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) leads to FDR control for independent hypotheses and a modification allowing for arbitrary dependence among the tests, but again being conservative, is given in Benjamini and Yekutieli (2001). If p is very large, it is often hard to detect a single significant marginal association, because of the large multiple

[12] For example, the empirical correlation between Y and $X^{(j)}$ is not depending on whether there are none, few or many other variables $X^{(k)}$ $(k \neq j)$.

testing adjustment factor. A hierarchical approach where statistical tests are pursued in a top-down fashion from large groups of correlated test-statistics to smaller groups and individual hypotheses is presented in Meinshausen (2008): it is an interesting route to deal with the problem of very high multiplicity in testing.

5.1.1 Software in R

Several methods for multiple testing adjustment are implemented in the R-package `multtest` (Pollard et al., 2012). In the following we show for the `riboflavin` dataset how to select genes controlling the FWER at 0.05 and using simple linear regression as marginal test.

```
## Installing this package from Bioconductor:
## source("http://bioconductor.org/biocLite.R")
## biocLite("multtest")

library(multtest)

## compute marginal regressions and extract p-values
p <- ncol(riboflavin)-1
pval <- vector("numeric", p)
for (i in 1:p) {
  fit <- lm(riboflavin[,1] ~ riboflavin[,i+1])
  tab <- summary(fit)$coefficients
  pval[i] <- tab[2,4]
}

## Holm to control FWER (53 genes selected)
resHolm <- mt.rawp2adjp(rawp = pval, proc = "Holm")
head(resHolm$adjp)
## extract the column index of those variables
## with adjusted p-values less than 0.05
idx <- resHolm$index[which(resHolm$adjp[,"Holm"] < 0.05)] + 1
## names of corresponding genes
colnames(riboflavin)[idx]

## Benjamini-Hochberg to control FDR (375 genes selected)
resBH <- mt.rawp2adjp(rawp = pval, proc = "BH")
head(resBH$adjp)
## extract the column index of those variables
## with adjusted p-values less than 0.1
idx <- resBH$index[which(resBH$adjp[,"BH"] < 0.1)] + 1
## names of corresponding genes
colnames(riboflavin)[idx]
```

53 genes are selected when controlling the FWER at 0.05. Finally, we show how to select genes controlling the FDR at 0.1 and, again, using simple linear regression as marginal test.

```
## Benjamini-Hochberg to control FDR
resBH <- mt.rawp2adjp(rawp = pval, proc = "BH")
head(resBH$adjp)
## extract the column index of those variables
## with adjusted p-values less than 0.1
idx <- resBH$index[which(resBH$adjp[,"BH"] < 0.1)] + 1
## names of corresponding genes
colnames(riboflavin)[idx]
```

375 genes are selected when controlling the FDR at 0.1.

Thus, with the marginal approach many more genes are selected than in the conditional approach using the Lasso and sample splitting, projection estimators or stability selection as discussed in Section 3. This is expected since the marginal approach measures total association which could possibly be explained away by taking information of the remaining variables into account. In contrast, the conditional approach measures only direct association which cannot be explained away by conditioning on the remaining variables. In this sense, the conditional approach uses a stricter criterion for selection and thus has the tendency of yielding a (much) smaller amount of selected variables.

6 Causal inference based on Directed Acyclic Graphs (DAGs)

In the previous sections, we largely focused on estimating a regression or marginal association parameter: in many applications, based on such estimated parameters, we would then assign strength or importance to a variable. For example, if a parameter estimate $|\hat{\beta}_j|$ is large in the linear model (1), we assign a high importance to the covariable $X^{(j)}$ for explaining the response Y.

Often though, a much more interesting (and ambitious) goal is to infer the causal strength of a variable $X^{(j)}$ on a response of interest Y. Causal strength can be described as an outside intervention on the variable $X^{(j)}$ and measuring the size of its effect on the response Y: this can be formalized, for example using Pearl's do-operator calculus (Pearl, 2000).

To illustrate the difference to regression, we consider the situation where the response Y and the covariables $X = (X^{(1)}, \ldots, X^{(p)})$ have a $(p+1)$-dimensional Gaussian distribution. We can then always relate Y to X with a linear model as in (1):

$$Y = \sum_{j=1}^{p} \beta_j X^{(j)} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is independent from X. The parameter β_j measures the effect on Y when changing $X^{(j)}$ by one unit and *keeping all other covariables fixed*. In many practical applications though, when we make an intervention at say variable $X^{(j)}$, we cannot keep the other covariables $\{X^{(k)};\ k \neq j\}$ fixed: for example, if we make a perturbation at gene j with corresponding $X^{(j)}$ measuring e.g. its expression, the expression

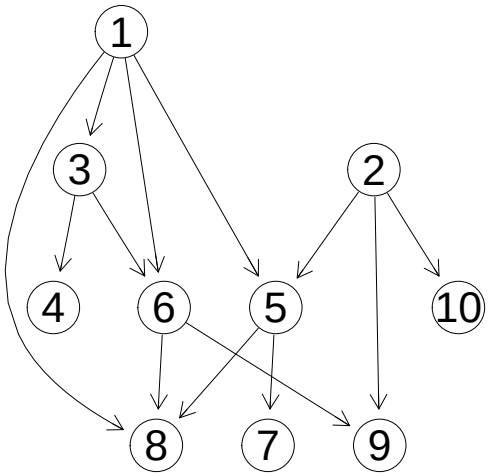

Figure 4: Example of a causal DAG on $p = 10$ nodes. Imagine a corresponding linear structural equation model with Gaussian errors that produces the data. To estimate the causal effect of node 3 on node 9, we would regress variable 9 on variable 3 and variable 1 (since node 1 is the only parent node of node 3; variable 1 is a so-called adjustment variable).

of the other genes $\{X^{(k)};\ k \neq j\}$ will change as well (and hence cannot be kept fixed). Causal inference and intervention analysis often aim to quantify the total effect on Y when making an intervention at variable $X^{(j)}$, including all indirect effects of $X^{(j)}$ on Y which are caused by the chain of events that an intervention at $X^{(j)}$ changes many other $X^{(k)}$'s ($k \neq j$) which in turn have an influence on the response Y. A common way to formalize the causal structure is given by a directed acyclic graph (DAG) which has no directed cycles. Such a total effect of an intervention at $X^{(j)}$ to the response Y, denoted by γ_j, can then be quantified using the do-calculus (Pearl, 2000): in the Gaussian case, γ_j equals the regression parameter for covariable $X^{(j)}$ in a linear model when regressing Y on $X^{(j)}$ and the variables $\{X^{(k)};\ k \in \mathrm{pa}(j)\}$ where $\mathrm{pa}(j)$ denotes the parental set of nodes of vertex j, i.e., $\mathrm{pa}(j) = \{k;$ there is a directed edge from k to $j\}$ (and $\mathrm{pa}(j)$ are sometimes called the adjustment variables). See Figure 4 for an example.

6.1 Bounds for causal effects based on observational data

As discussed above, estimation of a causal or intervention effect γ_j can be based on linear regression and an estimate of the parental set $\mathrm{pa}(j)$. The latter is a structure estimation problem of inferring a true underlying DAG. In general, however, the DAG is not identifiable from the observational distribution (i.e. the distribution from non-intervention data) and we can only infer a so-called Markov equivalence class of DAGs (Spirtes et al., 2000; Pearl, 2000). The latter can be estimated, for example by the PC-algorithm

(Spirtes et al., 2000) or ℓ_0-penalized maximum likelihood estimation (Chickering, 2002): in the high-dimensional setting where $p \gg n$ but the true underlying DAG is sparse, consistency of the estimation has been established for both the PC-algorithm (Kalisch and Bühlmann, 2007) and the ℓ_0-penalized maximum likelihood estimator (van de Geer and Bühlmann, 2013).

Because we can only identify a Markov equivalence class from observational data, we cannot infer a causal or intervention effect γ_j from observational data. However, it is still possible to identify lower bounds for $|\gamma_j|$ with the so-called IDA (Inference when Dag is Absent) procedure (Maathuis et al., 2009, 2010). These lower bounds can be used for ranking the importance of variables $X^{(j)}$ in terms of their absolute value of the intervention effect on a response variable Y, and such a ranking can be used in practice to prioritize variables with respect to causal strength, as demonstrated in Maathuis et al. (2010).

Assigning uncertainties for such lower bound estimates of causal effects can be pursued with stability selection from Section 3.3 where the selection algorithm $\hat{S}$ is given by the (top q) highest lower bound estimates, see Section 6.2. A related procedure is advocated in Stekhoven et al. (2012).

The IDA method is based on several strong assumptions, most notably that the true underlying influence diagram is a DAG, which does not allow for feedback mechanism, and that all relevant variables in the causal system are observed. Some relaxations of these conditions have been worked out: the FCI algorithm allows for hidden variables (Spirtes et al., 2000; Colombo et al., 2012) while graphs with directed cycles are considered in Spirtes (1995), Richardson (1996) and Mooij et al. (2011).

6.2 Software in `R`

Software for fitting the causal effect using IDA is provided in the `R`-package `pcalg` (Kalisch et al., 2012). As an illustrative example, we use IDA to estimate the causal effect of gene `YCIC_at` on the riboflavin production. For ease of reproduction, only the 100 genes with highest empirical variances and the response variable of riboflavin production were included in the estimation process (i.e., using the `riboflavinV100` data-set).

```
## Estimate causal effect of YCIC_at on Riboflavin production
library(pcalg)

## For ease of reproduction, we only use the 100 Genes
## with largest empirical variance
## Full data with model selection takes more than 2 hours
n <- nrow(riboflavinV100) ## n = 71 samples
p <- ncol(riboflavinV100) ## p = 1+100 variables in total

## position of explanatory variable in data frame
xPos <- 2 ## Activity of YCIC_at is in column 2
```

```
## position of goal variable in data frame
yPos <- 1 ## Riboflavin production is in first column

## Estimate covariance matrix of all involved variables
covMat <- cov(riboflavinV100)
corMat <- cov2cor(covMat)

## Estimate causal stucture
suffStat <- list(C = corMat, n = n) ## prepare input data
pc.fit <- pc(suffStat, indepTest = gaussCItest, p = p,
             alpha = 0.01) ## fit causal structure
pcEst <- pc.fit@graph ## extract estimated graph object

## Estimate causal effects of YCIC_at on Riboflavin production
res <- ida(x.pos = xPos, y.pos = yPos, mcov = covMat, graphEst = pcEst)
```

The resulting estimated lower bound for the causal effect in this example is 0.26. Actually, the obtained value turns out to be not only a lower bound but in fact an estimate of the causal effect $\gamma_{\text{YCIC\_at}}$ (due to so-called uniqueness within an estimated Markov equivalence class). This suggests that gene `YCIC_at` has a causal effect for the riboflavin production; in particular, if one increases the expression of `YCIC_at` by one unit, the riboflavin production is expected to increase by 0.26 units. For completeness, the effect of gene `YCIC_at` on the riboflavin production rate was also computed based on the full `riboflavin` data set with $p = 4088$ genes. Then, the causal effect $\gamma_{\text{YCIC\_at}}$ is estimated as non-identifiable (since the estimated Markov equivalence class leads to different causal effects): however, it is possible to obtain an estimated lower bound for the absolute value of the causal effect $|\gamma_{\text{YCIC\_at}}|$. This estimated lower bound equals 0.08 which still allows for the interpretation that an increase of the expression of `YCIC_at` by one unit leads to a change of the riboflavin production rate of at least 0.08 units. We refer the interested reader to Kalisch et al. (2012) for more details on the use of this package.

We can easily use `pcalg` in connection with stability selection from Section 3.3. For the `riboflavinV100` data-set, using $q = 5$ and $\pi_{\text{thres}} = 0.54$ resulting in $\mathbb{E}[V] \leq 3$, and based on $B = 100$ random splits, we find `XHLA_at` as a stable gene for having (potentially) a causal effect on the riboflavin production rate.

7 Summary Points

1. Extracting information (including assigning uncertainty) from high-dimensional data is possible using appropriate modern statistical methods.
2. Software implementations of most methods are readily available in `R` (R Core Team, 2012).
3. Two additional main assumptions are usually required to guarantee reasonable performance, besides the standard conditions for low-dimensional settings: (i)

sparsity for the underlying structure and (ii) identifiability of the model. An exception is the marginal approach which does not necessarily require such conditions.

4. Regarding point 3: sparsity is a fundamental and basic assumption on the unknown parameter vector, and a typical way to ensure identifiability is given by imposing conditions on the design matrix. For (bounds of) causal inference statements, which is a much more ambitious task than regression or classification, further assumptions are required.

5. Typically and unfortunately, the main conditions in point 3 are difficult (or impossible) to check, and powerful diagnostic tools for corresponding model assumptions are largely missing.

6. In view of point 5., drawing confirmatory conclusions from high-dimensional data should only be done with great care.

7. Some areas in biology allow for experimental validation of hypotheses which are derived or prioritized using statistical methods. Such validation is of major importance not only for the field of application but also for further understanding or appropriateness of statistical assumptions and techniques.

Sidebar

	H_0 holds	H_A holds
Declared significant	false positives (FP)	true positives (TP)
Declared non-significant	true negatives (TN)	false negatives (FN)

Table 1: Terminology of different error types. A false positive is a type I error, a false negative is a type II error.

	H_0 holds	H_A holds	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 2: Possible outcome of a total of m different hypothesis tests. The number of false positives is denoted by V and the number of false negatives by T.

- **Familywise error rate (FWER)**: The probability of making at least one false positive selection, i.e.
$$\text{FWER} = \mathbb{P}[V > 0].$$
- **False discovery rate (FDR)**: The expected value of the proportion of incorrectly rejected null hypotheses ("false discoveries") among all rejections ("discoveries"), i.e.
$$\text{FDR} = \mathbb{E}\left[\frac{V}{R}\right].$$

Acknowledgments

The data set about riboflavin production in Bacillus subtilis has been kindly provided by DSM (Kaiseraugst, Switzerland). We express our gratitude to Markus Wyss and Hans-Peter Hohmann for generously agreeing to make the data publicly available, and we thank Andrea Muffler for data collection and Sabine Arnold for acting as scientific contact person at DSM. We also thank an anonymous reviewer for detailed and constructive comments.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*. To appear.

Bühlmann, P., Kalisch, M., and Maathuis, M. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika*, 97:261–278.

Bühlmann, P. and Mandozzi, J. (2013). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. Preprint.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Colombo, D., Maathuis, M., Kalisch, M., and Richardson, T. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40:294–321.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*, 70:849–911.

Fellinghauer, B., Bühlmann, P., Ryffel, M., von Rhein, M., and Reinhardt, J. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9:432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Friedman, J., Hastie, T., and Tibshirani, R. (2011). *glasso: Graphical lasso- estimation of Gaussian graphical models*. R package version 1.7.

Gasser, T., Kneip, A., and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, 86:643–652.

Gautier, L., Cope, L., Bolstad, B., and Irizarry, R. (2004). affy - analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20:307–315.

Genovese, C., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the Lasso and marginal regression. *Journal of Machine Learning Research*, 13:2107–2143.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York, second edition.

He, Q. and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics*, 27:1–8.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Lee, J.-M., Zhang, S., Saha, S., S.S.Anna, Jiang, C., and Perkins, J. (2001). RNA expression analysis using an antisense Bacillus subtilis genome array. *Journal of Bacteriology*, 183:7371–7380.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*, 40:2293–2326.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selction (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems*.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248.

Maathuis, M., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37:3133–3164.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Meier, L. (2013). *hdi: High-Dimensional Inference*. R package version 0.0-1/r2.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, 52:374–393.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95:265–278.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection (with discussion). *Journal of the Royal Statistical Society Series B*, 72:417–473.

Meinshausen, N., Maathuis, M., and Bühlmann, P. (2011). Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *Annals of Statistics*, 39:3369–3391.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681.

Mooij, J., Janzing, D., Heskes, T., and Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24, 24nd Annual Conference on Neural Information Processing Systems (NIPS 2011)*.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ. Press.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., and Dudoit, S. (2012). *multtest: Resampling-based multiple hypothesis testing*. R package version 2.14.0.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI-1996)*, pages 454–461.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 24:398–413.

Schelldorfer, J. (2011). *lmmlasso: Linear mixed-effects models with Lasso*. R package version 0.1-2.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38:197–214.

Schelldorfer, J., Meier, L., and Bühlmann, P. (2013). GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using L1-penalization. *Journal of Computational and Graphical Statistics*. To appear.

Shah, R. and Samworth, R. (2013). Variable selection with error control: another look at Stability Selection. *Journal of the Royal Statistical Society Series B*, 75:55–80.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 491–499.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, second edition.

Stekhoven, D., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M., and Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28:2819–2823.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99:879–898.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. 3:1360–1392.

van de Geer, S. and Bühlmann, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*. To appear.

van de Geer, S., Bühlmann, P., and Ritov, Y. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. Preprint.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37:2178–2201.

Westfall, P. and Young, S. (1989). P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84:780–786.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics*, 40:2541–2571.

Zamboni, N., Fischer, E., Muffler, A., Wyss, M., Hohmann, H.-P., and Sauer, U. (2005). Transient expression and flux changes during a shift from high to low riboflavin production in continuous cultures of Bacillus subtilis. *Biotechnology and Bioengineering*, 89:219–232.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942.

Zhang, C.-H. and Zhang, S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. arXiv:1110.2563v1.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The `huge` package for high-dimensional undirected graph estimation in `R`. *Journal of Machine Learning Research*, 13:1059–1062.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics*, 36:1509–1566.

Peter Bühlmann
Markus Kalisch
Lukas Meier
Seminar for Statistics, ETH Zürich
ETH-Zentrum, HG, CH-8092 Zürich, Switzerland
e-mail: {buhlmann,kalisch,meier}@stat.math.ethz.ch

A Supplement

A.1 Riboflavin production data

The data-set, introduced in Section 1, about riboflavin (vitamin B_2) production by Bacillus subtilis has been kindly provided by DSM (Switzerland), see also Lee et al. (2001) and Zamboni et al. (2005). The log-transformed riboflavin production rate is the single real-valued response variable, and there are $p = 4088$ (co-)variables measuring the logarithm of the expression level of 4088 genes. We make the data available (see Supplemental Materials).

There is a homogeneous data-set from $n = 71$ samples, denoted as `riboflavin`. For ease of reproduction in some examples, we also provide a data set containing only the 100 genes with largest empirical variances and the response variable of riboflavin production, denoted as `riboflavinV100`.

Another data-set consists of measurements as above at different time points with $N = 28$ groups each having 2 to 6 observations at different times. The total number of samples is $n = 111$. This data-set is denoted as `riboflavinGrouped`.

A.2 Some notes on identifiability

For linear models, strong conditions on maximal pairwise empirical correlations among covariables (columns of $\mathbf{X}$) are checkable and lead to (approximate) identifiability of the parameter β. Weaker conditions on $\mathbf{X}$ are often formulated in terms of the compatibility constant or restricted eigenvalues of $\mathbf{X}^T\mathbf{X}/n$ (Bickel et al., 2009; van de Geer and Bühlmann, 2009; Bühlmann and van de Geer, 2011). Unfortunately, they are uncheckable in practice: although we observe $\mathbf{X}$, we typically cannot check the conditions because essentially, we would have to consider all subsets of variables having a certain cardinality, and this becomes very quickly computationally infeasible. An interesting exception is some recent work by Juditsky et al. (2013) who present checkable and "weak" conditions which lead to performance guarantees for the Lasso.

A.3 Score vectors for the projection estimators in Section 3.2.2

For the score vectors in Section 3.2.2, we consider two proposals based on Ridge and Lasso regression. Regarding the former, consider

$$Z^{(j)}_{\text{Ridge}} = ([(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T]_{j\cdot})^T,$$

where $A_{j\cdot}$ denotes the jth row of a matrix A. The corresponding estimator becomes

$$\hat{b}_{\text{Ridge};j} = \frac{(Z^{(j)}_{\text{Ridge}})^T Y}{(Z^{(j)}_{\text{Ridge}})^T X^{(j)}}$$

$$= \quad \text{usual Ridge estimator for } \beta_j \text{ standardized with the factor } 1/(Z^{(j)}_{\text{Ridge}})^T X^{(j)}.$$

The corrected estimator is denoted by

$$\hat{\beta}_{\text{Ridge}-\text{corr};j} \text{ as in (8) based on } Z^{(j)}_{\text{Ridge}},$$

which has been proposed in Bühlmann (2013).

The second approach using Lasso regression is motivated from classical least-squares methodology in the $p < n$ settings: we can obtain the jth coefficient $\hat{\beta}_{\text{OLS},j}$ by running a least-squares regression of Y versus the residuals from (least squares) regressing $X^{(j)}$ against all other variables $\{X^{(k)};\ k \neq j\}$. In the high-dimensional setting with $p > n$, we construct the residuals from a Lasso-regression of $X^{(j)}$ versus $\{X^{(k)};\ k \neq j\}$, and this residual vector is our vector $Z^{(j)}_{\text{Lasso}}$. The corresponding estimator is then

$$\hat{\beta}_{\text{Lasso}-\text{corr};j} \text{ as in (8) based on } Z^{(j)}_{\text{Lasso}},$$

as proposed in Zhang and Zhang (2011) and further analyzed in van de Geer et al. (2013).

A.4 Mathematical assumptions for methods providing measures of uncertainty

We summarize, on a superficial level, the mathematical assumptions underlying various methods.

A.4.1 P-values from single sample splitting: Algorithm 1

For asymptotically valid p-values, in the scenario where $p \gg n$, the single sample splitting Algorithm 1 requires an identifiability assumption for the design $\mathbf{X}$, for example a compatibility or restricted eigenvalue condition. Also a rather standard sparsity condition is required, saying that $|S| \log(p)/n \to 0$, where $|S|$ denotes the cardinality of the active set, i.e., the number of non-zero regression coefficients. Furthermore, the method is justified when assuming a beta-min condition (5) (Meinshausen et al., 2009): such a beta-min condition can be slightly weakened to a "zonal" assumption for the underlying regression coefficients (Bühlmann and Mandozzi, 2013): the non-zero coefficients need to be either sufficiently large in absolute value (i.e. $|\beta_j| > \text{ const.}\sqrt{\log(p)/n}$ if $\beta_j \neq 0$) or sufficiently small (i.e. $|\beta_j| < L$ where L is depending on various characteristics).

Although rigorous theoretical justification has been given for linear models with Gaussian errors only (Meinshausen et al., 2009), the method should provide asymptotically valid p-values for other models (e.g. generalized linear models) when requiring a condition on the design and on the unknown coefficients (i.e. a beta-min or zonal assumption).

A.4.2 P-values from multiple sample splitting: Algorithm 2

The multiple sample splitting Algorithm 2 requires exactly the same assumptions as the single splitting technique discussed above. Thus, the additional reproducibility, avoiding a "p-value lottery", and often also additional power come for free.

A.4.3 P-values from projection estimators in Section 3.2.2

The main difference to the sample splitting methods, regarding underlying mathematical assumptions, is that the projection estimators do not make a beta-min or zonal assumption for the unknown regression coefficients, except that the regression coefficient vector should be sparse with $|S| \log(p) n^{-1/2} \to 0$ (van de Geer et al., 2013). As usual (and e.g. for the sample splitting procedures), the projection estimators also require some identifiability conditions on the design, for example a compatibility or restricted eigenvalue assumption.

To achieve optimality in terms of semiparametric efficiency, when using $\hat{\beta}_{\mathrm{corr-Lasso}}$, some sparsity assumption on the design is made: for linear models, the work in van de Geer et al. (2013) requires that regressing one covariable against all others is a sparse problem where the number of nonzero coefficients is of small order $o(n/\log(p))$.

A.4.4 Stability selection

Stability selection does not require an explicit condition on e.g. the regression coefficient: it only assumes that the selection method $\hat{S}$ performs better than random guessing, and this seems indeed a rather weak condition. The restriction, however, comes in terms of a so-called exchangeability condition (Meinshausen and Bühlmann, 2010): for linear models, it essentially means that the selection of noise covariables is equally likely among all inactive (noise) variables.

References

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Juditsky, A., Kilinç-Karzan, F., Nemirovski, A., and Polyak, B. (2013). Accuracy guarantees for ℓ_1 recovery of block-sparse signals. *Annals of Statistics*, 40:3077–3107.