High-Dimensional Statistics with a View Toward Applications in Biology

Peter Bühlmann, Markus Kalisch, and Lukas Meier

Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland;
email: buhlmann@stat.math.ethz.ch, kalisch@stat.math.ethz.ch, meier@stat.math.ethz.ch

Keywords

Abstract

We review statistical methods for high-dimensional data analysis and pay particular attention 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 using a high-throughput genomic data set about riboflavin production with *Bacillus subtilis*, which we make publicly available for the first time.

1. INTRODUCTION

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad\qquad 1.$$

with $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ a univariate response vector, $\mathbf{X}$ the $n \times p$ design matrix whose jth column contains the covariable $\mathbf{X}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})^T$, and the 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 $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$. An intercept term may be implicitly present.[1] Classical statistical methods, like ordinary least squares estimation, cannot be used for estimating $\boldsymbol{\beta}$ and σ_ε^2 when $p \gg n$ because they would overfit the data, besides causing severe identifiability issues. A way out of the ill-posed estimation in the model in Equation 1 is by assuming a sparse structure, typically by saying that only a few of the components of $\boldsymbol{\beta}$ are nonzero. We review concepts for inference in the simple model in Equation 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). A concrete example is given next.

> **Example 1 (riboflavin production with *Bacillus subtilis*):** As a concrete example, we discuss a data set about riboflavin (vitamin B$_2$) production with *B. subtilis*. DSM (Kaiseraugst, Switzerland) (see also Lee et al. 2001 and Zamboni et al. 2005) has kindly provided these data, and for the first time, we make the data set publicly available in **Supplemental Section A.1** (follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**). There is a single real-valued response variable, which is the logarithm of the riboflavin production rate. Furthermore, $p = 4{,}088$ (co)variables measure the logarithm of the expression level of 4,088 genes; these gene expressions were normalized using the default in the R package `affy` (Gautier et al. 2004). One rather homogeneous data set exists from $n = 71$ samples that were hybridized repeatedly during a fed-batch fermentation process in which different engineered strains and strains grown under different fermentation conditions were analyzed. This data set is denoted as `riboflavin`, and we make it available in the **Supplemental Material**. Another data set consists of measurements (as above) at different time points (i.e., longitudinal data), with $N = 28$ groups, each having two to six observations at different times. Observations in the same group are from measurements from the same strain of (genetically engineered) *B. subtilis*, as 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 as well in the **Supplemental Material**.

The easiest approach is to model the homogeneous riboflavin production data set with a linear model, as in Equation 1, with $4{,}088 = p \gg n = 71$, and we discuss this approach 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 Equation 1, or using extensions of it, as presented in Section 4. We present in Section 6 a method for causal statistical inference that can deal with high-dimensional scenarios.

---

[1]In Sections 2 and 3, we do not penalize an intercept.

Supplemental Material

2. STATISTICAL ESTIMATION IN A HIGH-DIMENSIONAL LINEAR MODEL

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

Estimation of a high-dimensional linear model in Equation 1 with $p \gg n$ requires some regularization. Common approaches include Bayesian or penalized likelihood methods. We largely focus on the latter. From now on, we implicitly assume that the (co)variables are all (at least roughly) on the same scale: Very often, we achieve this assumption by standardizing the design matrix, such that $\sum_{i=1}^{n} X_i^{(j)} = 0$ and $||\mathbf{X}^{(j)}||_2^2 = n$ for every $j = 1, \ldots, p$. Ridge regression is defined as follows:

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

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

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

where $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$, and $\lambda > 0$ is again a regularization parameter (which is typically chosen differently as for Equation 2).[2] The estimators in Equations 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:

If $f(\cdot)$ is from $\mathcal{N}(0, \tau^2)$, then $\hat{\boldsymbol{\beta}}_{\text{MAP}} = \text{argmin}_{\beta}(||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \sigma_\varepsilon^2/\tau^2||\boldsymbol{\beta}||_2^2)$;
if $f(\cdot)$ is from $\text{DExp}(\tau)$, then $\hat{\boldsymbol{\beta}}_{\text{MAP}} = \text{argmin}_{\beta}(||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + 2\sigma_\varepsilon^2\tau||\boldsymbol{\beta}||_1)$,

where $\text{DExp}(\tau)$ is a double-exponential distribution with density $f(\beta) = \tau/2 \exp(-\tau|\beta|)$.

Both estimators in Equations 2 and 3 are shrinking the coefficient estimates toward zero because the penalty discourages large values. The Lasso has the special property of shrinking 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; in this sense, the Lasso is doing variable selection. This property can be best understood from equivalent optimization problems: The Ridge and Lasso estimators can be expressed as

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \text{argmin}_{\beta}(||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2/n) \text{ under the constraint that } ||\boldsymbol{\beta}||_2 \leq R,$$
$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \text{argmin}_{\beta}(||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2/n) \text{ under the constraint that } ||\boldsymbol{\beta}||_1 \leq R, \qquad 4.$$

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

Many versions of the Lasso have been proposed (Zou 2006, Meinshausen 2007, Zou & Li 2008, van de Geer et al. 2011), and other penalized estimators that lead to sparse solutions exist

---

[2]The factor $1/n$ is irrelevant from a methodological viewpoint: When dropped, we simply use another λ, which is n times the original λ.
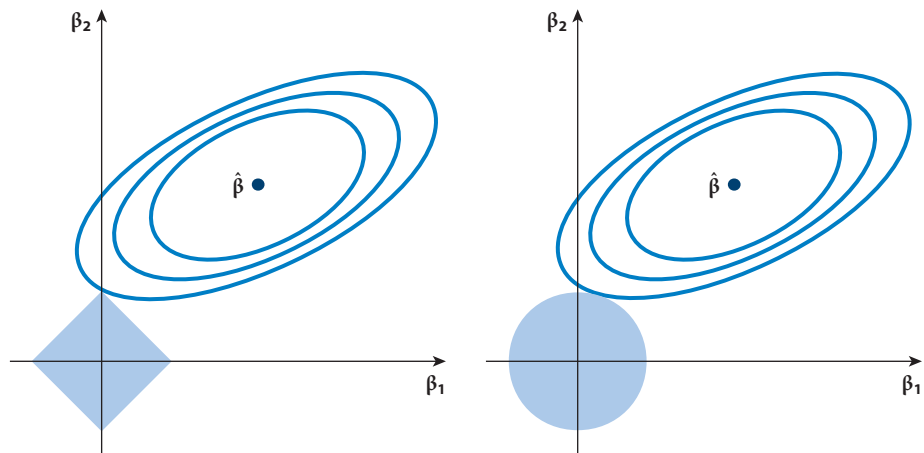
Figure 1

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

(Fan & Li 2001, Candès & Tao 2007, Zhang 2010). For further information, readers may wish to refer to, for example, Bühlmann & van de Geer (2011).

2.1. Identifiability

If $p > n$, then the model parameters in Equation 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 does not have full rank, $\mathrm{rank}(\mathbf{X}) \leq \min(n, p) < p$ for $p > n$, and we can write $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\xi})$ for every $\boldsymbol{\xi}$ in the null space of $\mathbf{X}$.[3] Therefore, without further assumptions, inferring or estimating $\boldsymbol{\beta}$ from data is impossible. The issue is closely related to the classical setting with $p < n$ but $\mathrm{rank}(\mathbf{X}) < p$ (owing to linear dependence among covariables) or with ill-conditioned designs, both leading to difficulties with respect to identifiability. For prediction or estimation of $\mathbf{X}\boldsymbol{\beta}$ (the underlying regression surface), however, 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 $\boldsymbol{\beta}$ or for pursuing variable selection. Some more details and additional references are given in **Supplemental Section A.2**.

2.2. Point Estimation Without Measures of Uncertainty

Much progress has been made over the past decade for estimating without assigning uncertainty, confidence, or error measures, i.e., point estimation. Further development of methods that quantify uncertainty is important (as discussed in Section 3). Among the three most important goals in such (point) estimation are (*a*) predicting the regression surface, $\mathbf{X}\boldsymbol{\beta}$, or a new response, $Y_{\mathrm{new}} = \mathbf{X}_{\mathrm{new}}^T \boldsymbol{\beta}$;

---

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

(*b*) estimation of β; and (*c*) estimation of the support of β or the active set of relevant variables, $S = \{j; \beta_j \neq 0\}$.

For the first task (prediction), identifiability of β is not necessary because we are interested in only, e.g., $\mathbf{X}_{\text{new}}^T \beta$. Thus, from this perspective, prediction is often a much easier problem than estimation of the parameter β or variable selection. Regarding the second task task (parameter estimation), an identifiability assumption on the design $\mathbf{X}$ is required, for example, a restricted eigenvalue condition (see **Supplemental Section A.2**). Finally, for the third task (variable selection), we would like to have an accurate estimator, $\hat{S}$, for the active set, S. A prime example is the Lasso, for which 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, which says that the nonzero regression coefficients must be sufficiently large,

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

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 condition (irrepresentability) on the design is necessary for variable selection (Meinshausen & Bühlmann 2006, Zhao & Yu 2006). A less ambitious goal, which does not need such a strong assumption on the design, is variable screening: It requires that, at least with high probability, $\hat{S}$ contains all variables from S, i.e.,

$$\hat{S} \supseteq S, \qquad\qquad 6.$$

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 compatibility condition holds[4]) and when the $\beta_{\min}$ condition is assumed (Equation 5). Although variable screening is less ambitious than variable selection, the screening property in Equation 6 is typically hard to fulfill exactly.

Reasonable 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, $|S|$ (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 that perform reasonably well exist. 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, then high-dimensional statistical inference is ill posed and uninformative. Good statistical estimators for sparse situations should be sparse themselves. The Lasso (Equation 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 (cf. Hastie et al. 2009). Some earlier work points to the 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 much easier than measuring the accuracy of parameter estimation, variable selection, or screening. Regarding the latter two, the traditional thinking in

---

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

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

frequentist statistics follows the framework of hypothesis testing, in which 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 challenge is 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 for the riboflavin data set:

```
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 tenfold 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 in Equation 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. 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 that 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.

Table 1   Terminology of different error types

	H_0 holds	**H_A holds**
Declared significant	False positives[a]	True positives
Declared nonsignificant	True negatives	False negatives[b]

[a]A false positive is a type I error.
[b]A false negative is a type II error.

Table 2   Possible outcome of a total of *m* different hypothesis tests

	H_0 holds	**H_A holds**	**Total**
Declared significant	V[a]	S	R
Declared nonsignificant	U	T[b]	$m - R$
Total	m_0	$m - m_0$	M

[a]The number of false positives is denoted by V.
[b]The number of false negatives is denoted by T.

3.2. *p*-Values for High-Dimensional Linear Models

We discuss two different methods for constructing *p*-values. Both are useful tools. A comparison among the methods is briefly mentioned in Section 3.4.

3.2.1. Multisample 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. We select the variables using the first and do the statistical inference based on the second half of the data. Such 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 $j = 1, \ldots, p$. We thereby aim for controlling the familywise error rate (FWER), $\mathbb{P}[V > 0]$, where V is the number of false positives.[6] We refer readers to **Table 1** and **Table 2** for the terminology of different error types.

<div style="border-left:3px solid;padding-left:8px;">

Familywise error rate (FWER): the probability of making at least one false positive selection, i.e., FWER = $\mathbb{P}[V > 0]$

</div>

> **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 $I_1 \cap I_2 = \emptyset$, where $|I_1| = \lfloor n/2 \rfloor$ 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 $H_{0,j}$, where $j \in \hat{S}$, using classical least squares estimation assuming Gaussian errors (i.e., the t-test that is well defined because $|\hat{S}| \leq |I_2|$). 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 FWER.

---

[6]A false positive arises when the test procedure rejects $H_{0,j}$, although $H_{0,j}$ in fact holds true.
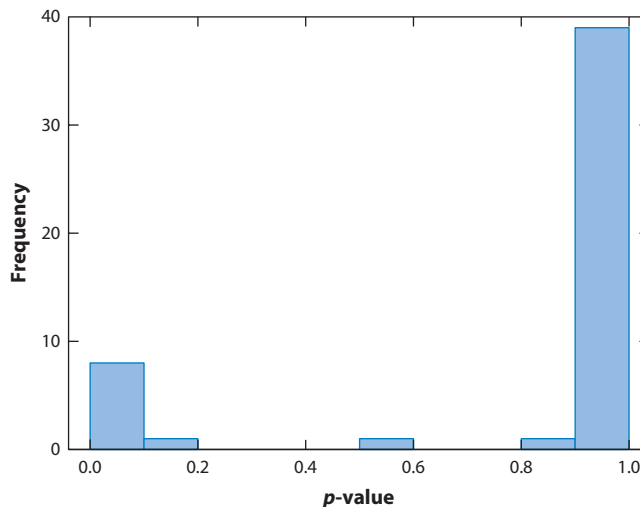
Figure 2

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

The procedure described in Algorithm 1 yields corrected p-values that control the FWER, when assuming the screening property in Equation 6: The whole idea is implicit in Wasserman & Roeder (2009). In practice, the screening property typically does not hold exactly, but that is not a necessary condition for constructing valid p-values (Bühlmann & Mandozzi 2013). Also, the correction for multiplicity of testing in step 4 involves only the multiplicative factor $|\hat{S}|$, whereas 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.

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

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

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

The remaining task is then to aggregate these $\{P_{\text{corr},j}^{[b]}; b = 1, \ldots, B\}$ to a single p-value. Owing to dependence among the $P_{\text{corr},j}^{[b]}$'s (because 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:

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

For example, when taking $\gamma = 1/2$, we multiply all $P_{\text{corr},j}^{[b]}$'s by two and take the empirical median among them. Furthermore, one can optimize over the best γ-quantile in the range $(\gamma_{\min}, 1)$ (e.g., with $\gamma_{\min}$ equal to 0.05), leading to the aggregated p-value

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

---

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

Thereby, the factor $(1 - \log(\gamma_{\min}))$ is the price to be paid for searching for the best $\gamma \in (\gamma_{\min}, 1)$. This multisample-splitting procedure has been proposed by Meinshausen et al. (2009) and is summarized in Algorithm 2. The multisample-splitting method results in p-values that are approximately reproducible and not subject to a lottery as illustrated in **Figure 2**, and it controls the FWER. As with the single-sample-splitting method, the procedure assumes the screening property (Equation 6) (or an approximate version of it). More precise mathematical assumptions for constructing valid p-values are given in **Supplemental Section A.4.2**.
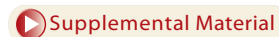
> **Algorithm 2 (multisample splitting for multiple testing of $H_{0,j}$ among $j = 1, \ldots, p$):**
>   1. Run the single-sample-splitting Algorithm 1 B times, leading to p-values $\{P_{\text{corr},j}^{[b]}; \quad b = 1, \ldots, B\}$. A typical choice is $B = 100$.
>   2. Aggregate the p-values from step 1 as in Equation 7, leading to P_j, which are adjusted p-values for $H_{0,j}(j = 1, \ldots, p)$, controlling the FWER.

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 multisample-splitting method assumes (a possibly relaxed form of) the screening property (Equation 6). This property in turn necessarily requires a (possibly relaxed) $\beta_{\min}$ assumption (Equation 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 $\mathbf{Z}^{(j)}$ and a corresponding estimator[8]:

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

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

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

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

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

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

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

Interestingly, these estimators ($\hat{\boldsymbol{\beta}}_{\text{corr-Ridge}}$ and $\hat{\boldsymbol{\beta}}_{\text{corr-Lasso}}$) are not sparse, and they have a Gaussian limiting distribution with a known covariance matrix, except for the unknown error variance, σ_ε^2 (Bühlmann 2013, Zhang & Zhang 2011, van de Geer et al. 2013). The latter can be estimated,

---

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

for example, by using the scaled Lasso (Sun & Zhang 2012). We then end up with a statement
of the form

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

where $\hat{\sigma}_j^2 = \hat{\sigma}_\varepsilon^2 \omega_j$ with a known ω_j (which is easily computable as a 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
Equation 5, the method provides valid inference for tests and confidence intervals. When using
$\hat{\beta}_{\text{corr-Lasso}}$, 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 that is valid for dependent tests ($P_1, \ldots, P_p$ are dependent).
For example, we can use the Bonferroni–Holm method (Holm 1979) to control the FWER, or we
can use a version of the standard Benjamini–Hochberg procedure (Benjamini & Hochberg 1995),
which controls the false discovery rate (FDR) among dependent tests (Benjamini & 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, implying that the usual adjustment methods are too conservative.[9] We
can address the potential loss of power from avoiding conservative adjustment by exploiting the
known covariance structure of the problem; more generally than in Equation 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 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 multisplit method yields one significant gene (gene YXLD_at), whereas 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
```

[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 one (instead of the nominal number of tests, p), and adjusting the p-values would not be
necessary.

3.3. Stability Selection

Stability selection (Meinshausen & Bühlmann 2010) is a method based on subsampling (or bootstrapping) but is 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

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

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

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

where $B \approx 500$–$1,000$ 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 $|I| = \lfloor n/2 \rfloor$. For example, we use the Lasso and select the q variables entering the regularization path[10] first, as used in Section 3.3.1 below, or the Lasso selects the top q variables having the highest estimated regression coefficients in absolute value. Furthermore, if such an $\hat{S}$ is better than random guessing and if an exchangeability condition holds (which becomes an assumption on the design, implying that false positive selection of any variable is equally likely), then we have the following relation, with V denoting the number of false positives:

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

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

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

which ensures by Equation 10 that the corresponding $\mathbb{E}[V] \leq$ `efp`. Shah & Samworth (2013) extend the result in Equation 10 without requiring the restrictive but unnecessary exchangeability assumption.

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 of at least one element in a group G being selected. This consideration is formalized as

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

The error bound in Equation 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., $k = 2$ for selecting stable variable groups of two).

---

[10]The regularization paths are 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^+.

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 Equation 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 for the `riboflavin` data set. For the selector $\hat{S}$, we use the Lasso with the q variables that enter the regularization path first. With $q = 20$, $V = 1$, and $B = 500$, we get three stable selected genes, specifically, `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, in which $n = 71$ and $p = 4,088$, the results from the different methods vary to a certain extent. The multisample-splitting Algorithm 2 and the projection estimator (Equation 8), here used with Ridge-type score vectors, lead to p-values controlling the very stringent FWER. At the FWER-adjusted 5% significance level, we find one significant variable (gene `YXLD_at`) based on the multisample-splitting Algorithm 2, whereas the projection estimator does not find a single significant variable or gene. This finding is not surprising: The Ridge-type projection estimator is rather conservative, and because it does not require a β_{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 the FWER. Furthermore, the one significant gene found with the multisample-splitting Algorithm 2 has the 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 multisample-splitting method and stability selection are straightforward for nonlinear models. The projection procedure from Section 3.2.2, in contrast, needs more careful treatment.

4.1. Generalized Linear Models

Generalized linear models (McCullagh & Nelder 1989) are very popular for extending the linear model in a unified way. We consider a model with univariate response Y and p-dimensional covariables $\mathbf{X}$:

$$Y_1, \ldots, Y_n \text{ are independent,}$$
$$\text{and } g(\mathbb{E}[Y_i|\mathbf{X}_i = \mathbf{x}]) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)}, \qquad 11.$$

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 $\mathbf{x}$. A well-known example is logistic regression for binary response variables, $Y_i \in \{0, 1\}$; we denote the conditional probability by $\pi(x) = \mathbb{P}[Y_i = 1|\mathbf{X}_i = \mathbf{x}](=\mathbb{E}[Y_i|\mathbf{X}_i = \mathbf{x}])$. 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: $Y_i|\mathbf{X}_i = \mathbf{x} \sim \text{Poisson}(\lambda(\mathbf{x}))$, 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 Equation 11 is that the (conditional) distribution of Y_i (given $\mathbf{X}_i$) depends on $\mathbf{X}_i$ only through the function $g(\mathbb{E}[Y_i|\mathbf{X}_i]) = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)}$. That is, the (conditional) probability or density of $Y|\mathbf{X} = \mathbf{x}$ is of the form

$$p(y|\mathbf{x}) = p_{\mu,\beta}(y|\mathbf{x}). \qquad 12.$$

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

$$-\ell(\mu, \beta; \text{ data}) = -\sum_{i=1}^{n} \log(p_{\mu,\beta}(Y_i|\mathbf{X}_i)),$$

where $p_{\mu,\beta}(y|\mathbf{x})$ is as in Equation 12. For many examples and models, e.g., if the (conditional) distribution of $Y|\mathbf{X} = \mathbf{x}$ is from a subclass of the exponential family model (see McCullagh & 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 the ℓ_1-norm penalization introduced below (see, e.g., 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

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

$$= \underset{\mu,\beta}{\arg\min}\left(-n^{-1}\sum_{i=1}^{n} \log(p_{\mu,\beta}(Y_i|\mathbf{X}_i)) + \lambda||\beta||_1\right). \qquad 14.$$

Usually, we do not penalize the intercept term.

Assumptions analogous to those for Lasso (Equation 3) for high-dimensional linear models are required for estimation of $\mathbf{X}_{\text{new}}^T\beta$, for estimation of β, and for the active set, $S = \{j; \beta_j \neq 0\}$. Except for estimation of $\mathbf{X}_{\text{new}}^T\beta$, we need identifiability assumptions on the design $\mathbf{X}$ and a condition for the smallest nonzero regression coefficients (as in Equation 5) (see, e.g., van de Geer 2008 and Bühlmann & van de Geer 2011). Assigning measures of uncertainty and significance is easily done using the multisample-splitting method from Section 3.2.1 or the stability selection method from Section 3.3. Regarding the former, we can use $\hat{S}$ from the penalized estimator in Equation 13, replace the t-test in step 3 of Algorithm 1 with the log-likelihood ratio test (McCullagh & 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 Equation 13; a related idea with applications to genome-wide association studies (GWAS) is presented in He & 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 Equation 13 can be computed using the R package `glmnet` (Friedman et al. 2010), as in Section 2.2.1.

4.2. Generalized Linear Mixed Models

Mixed effects models are popular for modeling grouped or longitudinal data (Pinheiro & Bates 2000). The building blocks are fixed effects with a corresponding p-dimensional parameter vector,

β, and random effects with a corresponding random parameter, $\mathbf{b} \sim \mathcal{N}_q(0, V)$. 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 in which p is large but 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: As in Equation 13, we consider[11]

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

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

As for generalized linear models, we can use the multisample-splitting method from Section 3.2.1 or the stability selection method described in Section 3.3 for quantifying uncertainties. For the former, we can use the screening method from the penalized estimator in Equation 15; a valid procedure for low-dimensional generalized linear mixed models must replace the t-test in step 3 of Algorithm 1, and we can then proceed with Algorithm 2. For stability selection for the fixed effects variables, we can use $\hat{S}$ from Equation 15.

> **Example 2 (grouped data about riboflavin production with *B. subtilis*):** One data set of riboflavin production with *B. subtilis* (see also Section 1) consists of measurements ($p = 4{,}088$ gene expressions and the riboflavin production rate) at different time points (longitudinal data) with $N = 28$ groups each having two to six observations at different times, and the total number of samples is $n = 111$. We refer to Example 1 in Section 1 for further description of the data set, which is denoted as `riboflavinGrouped` and made available in the **Supplemental Material**.
>
> We can fit a linear mixed model to this data with the 28 different groups. After some preliminary analysis, a reasonable model consists of two independent random effects and $p = 4{,}088$ fixed effects. The estimator in Equation 15 with a Gaussian distribution can be computed with the R package `lmmlasso` (Schelldorfer 2011). The results are presented in Schelldorfer et al. (2011).

The estimator in Equation 15 for generalized mixed effects models can be computed with the R package `glmmixedlasso` (Schelldorfer et al. 2013), which is available from R-Forge (Theußl & Zeisleis 2009).

4.3. Gaussian Graphical Models

For a further extension of linear models, we mention the Gaussian graphical model

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \text{ i.i.d.} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

---

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

Assuming that $\mathbf{\Sigma}^{-1}$ 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 follows:

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

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 $\{X^{(\ell)}; \quad \ell \neq j, k\}$.

Estimation of such a graph in the high-dimensional scenario can be done with a nodewise Lasso approach (Meinshausen & 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). Uncertainties could be assigned using the multisample-splitting method from Section 3.2.1. In view of the many edges and the multivariate nature of the model, Meinshausen & Bühlmann (2010) advocate stability selection.

Liu et al. (2012) and Xue & Zou (2012) give extensions for non-Gaussian continuous distributions, based on copula models. We consider such an extension in Section 4.3.1 using the nonparanormal transformation. Undirected graphical model estimation for cases 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). Because huge seems to be more elaborate, we report using only this package. For the riboflavin production data, we estimate an undirected graph with the Meinshausen–Bühlmann method (Meinshausen & Bühlmann 2006) and select the regularization parameter using a variant of stability selection in Section 3.3 termed StARS (Liu et al. 2010). For illustration and simplicity (without deeper biological implications), we estimate the undirected graph for the 100 genes with the largest empirical variance and riboflavin production and denote this reduced data set by riboflavinV100 (the fitting and selection process on the complete data set takes approximately 2 CPU hours). **Figure 3** shows the resulting graph. We refer the interested reader to the vignette contained in 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)
```
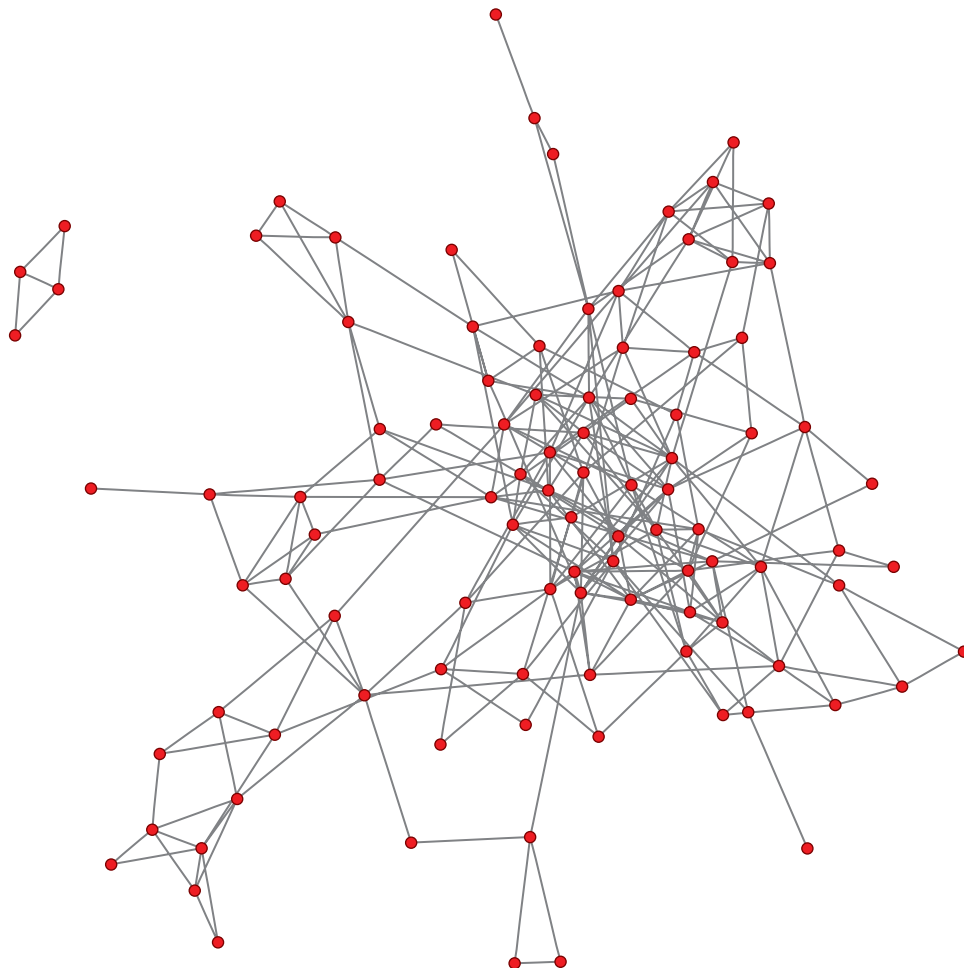
Figure 3

Estimated undirected graph for the `riboflavinV100` data set. For ease of reproduction, only the 100 genes with the largest empirical variance and the response variable measuring 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 one has response (or grouping or class) variables, Y_i, and p-dimensional (co)variables, $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(p)})^T$, with $(Y_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$ i.i.d., 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 $Y = \mu + \alpha_j X^{(j)} + \text{noise}$. Such marginal association parameters are very different from the parameters in, e.g., a linear model as in Equation 1 or more general regression models; the latter parameters measure the strength of association, which no other variables explain. Some recent attempts exist for variable screening in a linear model (Equation 1) as in Equation 6, based on marginal correlations. Under some rather strong conditions on the design matrix, the proposed methods provide a superset of S as

in Equation 6 (Fan & 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 $\mathbf{X}_i$ is not really problematic when estimating marginal association parameters, even if $p \gg n$.[12] The only drawback occurs when adjusting tests (and confidence intervals) with respect to multiplicity, especially when considering all $p \gg n$ marginal associations.

GWAS are examples in which, oftentimes, only marginal associations are considered. For example, if Y_i is binary, encoding the healthy or diseased status of an individual, and $X_i^{(j)}$ is a categorical variable with three levels describing a single-nucleotide polymorphism at position j in the genome, then 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 $\approx 10^6$, whereas the sample size is in the hundreds or low thousands.

5.1. Multiple Testing Adjustment

In the GWAS example above, we have p-values $P_1, \ldots, P_p$, in which p is very large and adjusting for multiplicity is crucial (Roeder & Wasserman 2009). Common type I error measures for multiple testing are the FWER (the probability of at least one false positive selection) and the 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 owing to such generality, the method is often overly conservative; the Westfall–Young method (Westfall & Young 1989) offers an alternative, at least for some cases, and often has better power (Meinshausen et al. 2011). The Benjamini–Hochberg procedure (Benjamini & Hochberg 1995) leads to FDR control for independent hypotheses, and a modification allowing for arbitrary dependence among the tests (Benjamini & Yekutieli 2001) is again conservative. If p is very large, then detecting a single significant marginal association is often hard because of the large multiple testing adjustment factor. A hierarchical approach, in which 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 way 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 data set how to select genes, controlling the FWER at 0.05 and using simple linear regression as a 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
```

---

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

```
  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]
```

When controlling the FWER at 0.05, 53 genes are selected. Finally, we show how to select genes, controlling the FDR at 0.1 and, again, using simple linear regression as a 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]
```

When controlling the FDR at 0.1, 375 genes are selected.

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 (Section 3). This difference is expected because the marginal approach measures total association, which may be explained away by taking 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 number of selected variables.

6. CAUSAL INFERENCE BASED ON DIRECTED ACYCLIC GRAPHS

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 (Equation 1), then we assign a high importance to covariable $X^{(j)}$ for explaining 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 variable $X^{(j)}$ and measuring the size of its effect on response Y. This description can be formalized, for example, using Pearl's do-operator calculus (Pearl 2000).
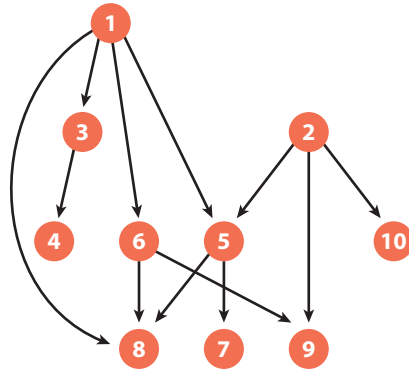
Figure 4

Example of a causal directed acyclic graph 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 (because node 1 is the only parent node of node 3; variable 1 is an adjustment variable).

To illustrate the difference from regression, we consider the situation in which the response Y and the covariables $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})^T$ have a $(p + 1)$-dimensional Gaussian distribution. We can then always relate Y to $\mathbf{X}$ with a linear model as in Equation 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, e.g., 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 variable $X^{(j)}$, which measures, e.g., its expression, then the expression 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 in many other $X^{(k)}$'s ($k \neq j$), which in turn influences response Y. A directed acyclic graph (DAG), which has no directed cycles, is a common way to formalize the causal structure. Such a total effect of an intervention at $X^{(j)}$ on the response Y, denoted by γ_j, can then be quantified using 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 $\{X^{(k)}; k \in \text{pa}(j)\}$, where $\text{pa}(j)$ denotes the parental set of nodes of vertex j, i.e., $\text{pa}(j) = \{k;$ there is a directed edge from k to $j\}$ [$\text{pa}(j)$ are sometimes called the adjustment variables]. **Figure 4** presents 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, $\text{pa}(j)$. 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 nonintervention data), and we can infer only a Markov equivalence class of DAGs (Pearl 2000, Spirtes et al. 2000). The latter can be estimated, for example, by the PC algorithm (Spirtes et al. 2000) or the ℓ_0-penalized maximum likelihood estimation (Chickering 2002). In the high-dimensional setting, in which $p \gg n$ but the

true underlying DAG is sparse, consistency of the estimation has been established for both the PC algorithm (Kalisch & Bühlmann 2007) and the ℓ_0-penalized maximum likelihood estimator (van de Geer & Bühlmann 2013).

Because we can identify only a Markov equivalence class from observational data, we cannot infer a causal or intervention effect, γ_j, from observational data. However, we can still identify lower bounds for $|\gamma_j|$ with the 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 the absolute value of their 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, in which the selection algorithm, $\hat{S}$, is given by the (top q) highest lower-bound estimates (see Section 6.2). Stekhoven et al. (2012) advocate a related procedure.

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 a feedback mechanism, and that all relevant variables in the causal system are observed. Some relaxations of these conditions have been worked out: The fast causal inference (FCI) algorithm allows for hidden variables (Spirtes et al. 2000, Colombo et al. 2012), whereas 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 riboflavin production. For ease of reproduction, only the 100 genes with the 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, $\gamma_{\text{YCIC\_at}}$ (owing to uniqueness within an estimated Markov equivalence class). This finding suggests that gene YCIC_at has a causal effect on riboflavin production; in particular, if one increases the expression of YCIC_at by one unit, then 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 $p = 4{,}088$ genes. Then, the causal effect, $\gamma_{\text{YCIC\_at}}$, is estimated as nonidentifiable (because the estimated Markov equivalence class leads to different causal effects). However, obtaining an estimated lower bound for the absolute value of the causal effect $|\gamma_{\text{YCIC\_at}}|$ is possible. 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 $\pi_{\text{thres}} = 0.54$ results in $\mathbb{E}[V] \leq 3$, and based on $B = 100$ random splits, we find XHLA_at as a stable gene that has (potentially) a causal effect on the riboflavin production rate.

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 Development Core Team 2012).

3. Two additional main assumptions are usually required to guarantee reasonable performance, besides the standard conditions for low-dimensional settings: (*a*) sparsity for the underlying structure and (*b*) identifiability of the model. An exception is the marginal approach, which does not necessarily require such conditions.

4. Regarding Summary Point 3, sparsity is a fundamental and basic assumption regarding 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, a much more ambitious task than regression or classification, further assumptions are required.

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

6. In view of Summary 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 derived from 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.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

LITERATURE CITED

Banerjee O, El Ghaoui L, d'Aspremont A. 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* 9:485–516

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

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29:1165–88

Bühlmann P. 2013. Statistical significance in high-dimensional linear models. *Bernoulli J.* 19:1212–42

Bühlmann P, Kalisch M, Maathuis M. 2010. Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* 97:261–78

Bühlmann P, Mandozzi J. 2013. High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Stat.* In press. doi: 10.1007/s00180-013-0436-3

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

Candès E, Tao T. 2007. The Dantzig selector: statistical estimation when *p* is much larger than *n*. *Ann. Stat.* 35:2313–51

Chickering D. 2002. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3:507–54

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

Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96:1348–60

Fan J, Lv J. 2008. Sure independence screening for ultra-high dimensional feature space. *J. R. Stat. Soc. Ser. B* 70:849–911

Fellinghauer B, Bühlmann P, Ryffel M, von Rhein M, Reinhardt J. 2013. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comput. Stat. Data Anal.* 64:132–52

Friedman J, Hastie T, Tibshirani R. 2007. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–41

Friedman J, Hastie T, Tibshirani R. 2010. Regularized paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33:1–22

Friedman J, Hastie T, Tibshirani R. 2011. Glasso: graphical lasso—estimation of Gaussian graphical models. *R Package Version 1.*7

Gasser T, Kneip A, Köhler W. 1991. A flexible and fast method for automatic smoothing. *J. Am. Stat. Assoc.* 86:643–52

Gautier L, Cope L, Bolstad B, Irizarry R. 2004. Affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307–15

Genovese C, Jin J, Wasserman L, Yao Z. 2012. A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* 13:2107–43

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer. 2nd ed.

He Q, 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. *Scand. J. Stat.* 6:65–70

Kalisch M, Bühlmann P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* 8:613–36

Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. 2012. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* 47(11):1–26

Lauritzen S. 1996. *Graphical Models*. Oxford: Oxford Univ. Press

Lee J-M, Zhang S, Saha S, Anna SS, Jiang C, Perkins J. 2001. RNA expression analysis using an antisense *Bacillus subtilis* genome array. *J. Bacteriol.* 183:7371–80

Liu H, Han F, Yuan M, Lafferty J, Wasserman L. 2012. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Stat.* 40:2293–326

Liu H, Roeder K, Wasserman L. 2010. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, Vol. 23, ed. J Lafferty, CKI Williams, J Shawe-Taylor, RS Zemel, A Culotta, pp. 1432–40. Red Hook, NY: Curran Assoc.

Maathuis M, Colombo D, Kalisch M, Bühlmann P. 2010. Predicting causal effects in large-scale systems from observational data. *Nat. Methods* 7:247–48

Maathuis M, Kalisch M, Bühlmann P. 2009. Estimating high-dimensional intervention effects from observational data. *Ann. Stat.* 37:3133–64

McCullagh P, Nelder J. 1989. *Generalized Linear Models*. London: Chapman & Hall. 2nd ed.

Meier L. 2013. Hdi: high-dimensional inference. *R Package Version 0.0-1/r2*. **http://hdi.r-forge.r-project.org**

Meinshausen N. 2007. Relaxed Lasso. *Comput. Stat. Data Anal.* 52:374–93

Meinshausen N. 2008. Hierarchical testing of variable importance. *Biometrika* 95:265–78

Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34:1436–62

Meinshausen N, Bühlmann P. 2010. Stability selection. *J. R. Stat. Soc. Ser. B* 72:417–73

Meinshausen N, Maathuis M, Bühlmann P. 2011. Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *Ann. Stat.* 39:3369–91

Meinshausen N, Meier L, Bühlmann P. 2009. P-values for high-dimensional regression. *J. Am. Stat. Assoc.* 104:1671–81

Mooij J, Janzing D, Heskes T, Schölkopf B. 2011. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems*, Vol. 24, ed. J Shawe-Taylor, RS Zemel, P Bartlett, F Pereira, KQ Weinberger, pp. 639–47. Red Hook, NY: Curran Assoc.

Pearl J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press

Pinheiro J, Bates D. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer

Pollard KS, Gilbert HN, Ge Y, Taylor S, Dudoit S. 2012. Multtest: resampling-based multiple hypothesis testing. *R Package Version 2.14.0*

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vienna: R Found. Stat. Comput.

Richardson T. 1996. A discovery algorithm for directed cyclic graphs. *Proc. 12th Conf. Uncertain. Artif. Intell. (1996)*, ed. E Horvitz, F Jensen, pp. 454–61. San Francisco: Morgan Kaufmann

Roeder K, Wasserman L. 2009. Genome-wide significance levels and weighted hypothesis testing. *Stat. Sci.* 24:398–413

Schelldorfer J. 2011. Lmmlasso: linear mixed-effects models with Lasso. *R Package Version 0.1-2*

Schelldorfer J, Bühlmann P, van de Geer S. 2011. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.* 38:197–214

Schelldorfer J, Meier L, Bühlmann P. 2013. GLMMLasso: an algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *J. Comput. Graph. Stat.* In press. doi: 10.1080/10618600.2013.773239

Shah R, Samworth R. 2013. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Ser. B* 75:55–80

Spirtes P. 1995. Directed cyclic graphical representations of feedback models. *Proc. 11th Conf. Uncertain. Artif. Intell. (1995)*, ed. P Besnard, S Hanks, pp. 491–99. San Francisco: Morgan Kaufmann

Spirtes P, Glymour C, Scheines R. 2000. *Causation, Prediction, and Search*. Cambridge: MIT Press. 2nd ed.

Stekhoven D, Moraes I, Sveinbjörnsson G, Hennig L, Maathuis M, Bühlmann P. 2012. Causal stability ranking. *Bioinformatics* 28:2819–23

Sun T, Zhang C-H. 2012. Scaled sparse linear regression. *Biometrika* 99:879–98

Theußl S, Zeisleis A. 2009. Collaborative software development using R-Forge. *R J.* 1:9–14

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58:267–88

van de Geer S. 2008. High-dimensional generalized linear models and the Lasso. *Ann. Stat.* 36:614–45

van de Geer S, Bühlmann P. 2009. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* 3:1360–92

van de Geer S, Bühlmann P. 2013. ℓ_0-Penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Stat.* 41:536–67

van de Geer S, Bühlmann P, Ritov Y. 2013. On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv: 1303.0518 [math.ST]

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

Wasserman L, Roeder K. 2009. High-dimensional variable selection. *Ann. Stat.* 37:2178–201

Westfall P, Young S. 1989. *p* value adjustments for multiple tests in multivariate binomial models. *J. Am. Stat. Assoc.* 84:780–86

Xue L, Zou H. 2012. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Stat.* 40:2541–71

Zamboni N, Fischer E, Muffler A, Wyss M, Hohmann H-P, Sauer U. 2005. Transient expression and flux changes during a shift from high to low riboflavin production in continuous cultures of *Bacillus subtilis*. *Biotechnol. Bioeng.* 89:219–32

Zhang C-H. 2010. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 38:894–942

Zhang C-H, Zhang S. 2013. Confidence intervals for low dimensional parameters with high dimensional data. *J. R. Stat. Soc. B*. In press. doi: 10.1111/rssb.12026

Zhao P, Yu B. 2006. On model selection consistency of Lasso. *J. Mach. Learn. Res.* 7:2541–63

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. 2012. The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* 13:1059–62

Zou H. 2006. The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.* 101:1418–29

Zou H, Li R. 2008. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* 36:1509–33. Discussion. 1534–66

Ⱥ

Contents