EMS Surv. Math. Sci. x (201x), xxx–xxx
DOI 10.4171/EMSS/x

EMS Surveys in
Mathematical Sciences
© European Mathematical Society

High-dimensional statistics, with applications to genome-wide association studies

Peter Bühlmann

We present a selective review on high-dimensional statistics where the dimensionality of the unknown parameter in a model can be much larger than the sample size in a dataset (e.g. the number of people in a study). Particular attention is given to recent developments for quantifying uncertainty in high-dimensional scenarios. Assessing statistical uncertainties enables to describe some degree of replicability of scientific findings, an ingredient of key importance for many applications. We also show here how modern high-dimensional statistics offers new perspectives in an important area in genetics: novel ways of analyzing genome-wide association studies, towards inferring more causal-oriented conclusions.

Mathematics Subject Classification (*2010*). 62J07; 62F25, 62P10.

Keywords. De-sparsified Lasso, Hierarchical multiple testing, Lasso, ℓ_1-norm regularization, Sparsity.

1. Introduction

High-dimensional statistics is an emerging area where statistical inference is pursued for models with very many parameters. The wording "very many" is here always relative to the sample size of a dataset: denoting by d the number of unknown parameters and by n the sample size of a dataset, "very many" or "high-dimensional" means that $d \gg n$. Such situations arise very often in nowadays applications. For example in genetics or genomics, thousands or 100-thousands of variables are measured for only $n \approx 10 - 1000$ individuals: many models have at least one unknown parameter for every measured variable (as e.g. in a linear model introduced in (2)), and thus we have that $d \gg n$. When making statistical

Peter Bühlmann, Seminar for Statistics, Department of Mathematics, Rämistrasse 101, ETH Zürich, CH-8092 Zürich, Switzerland
E-mail: buhlmann@stat.math.ethz.ch

inference for the d parameters based on n data points, the problem is ill-posed in general: a key assumption is sparsity which ensures that the inference problem has a well-posed solution and is (nearly) optimal in terms of accuracy: sparsity means that many of the d unknown parameters are (nearly) irrelevant, and such an assumption seems to be well supported by empirical findings in practice. This article will emphasize the importance of high-dimensional statistics for an important problem in genetics and genomics with so-called genome-wide association studies (GWAS): there, the number of genetic variables (single nucleotide polymorphisms) is about $0.5 - 1 \cdot 10^6$ and one or a few outcomes of say a disease status are measured as well, while the number of individuals in a single study is in the range of $n \approx 1'000 - 5'000$. As of 2011, 1'200 human GWAS over 200 diseases have been completed (Wikipedia on "Genome-wide association study"), and GWAS is nowadays a major approach with fair amount of replicability [15] to understand genetic-disease status associations. We will show in Section 5 that novel high-dimensional statistics leads to new methodology and tools for GWAS. Of course, there are other major application areas where high-dimensional statistics plays a key role: compressed sensing [22],[21],[13],[12] which can be seen as a "noise-free version" of estimation in high-dimensional statistics has become crucial in modern signal processing with high impact on imaging and fast MRI [40],[27],[36].

Much has happened in the last 2 decades in high-dimensional statistics with an almost exponential growth of contributions. We refer to the monographs [33],[8],[30],[34] which contain many references. This article reviews some of the important but nowadays rather standard methodology and theory (Section 2), includes more recent results on uncertainty quantification and statistical confidence statements (Sections 3 and 4) and then discusses the impact of the methods for GWAS (Section 5).

2. Data and high-dimensional models

We consider here the standard set-up in statistics. We have observed data $z_1, \ldots z_n$ (with values in a sub-space of $\mathbb{R}^q$) and these values are assumed to be realizations of i.i.d. (i.e., independent, identically distributed) random variables (vectors) $Z_1, \ldots, Z_n$, each having a probability distribution P_{θ^0} which is known up to an unknown parameter $\theta^0 \in \mathbb{R}^d$. That is,

$$Z_1, \ldots, Z_n \text{ i.i.d. } \sim P_{\theta^0}. \tag{1}$$

The superscript "^{0}" indicates that θ^0 is the true underlying parameter, and we implicitly think that the probability distribution P_θ comes from a certain model. Of course, if the model is not correct for the observed data, the notion of a true parameter θ^0 does not make sense. We will discuss the issue of model misspecification in Section 6. As mentioned already, the scenario is called high-dimensional if $\dim(\theta) = d$ is (much) larger than sample size n.

The i.i.d formulation with a known probability distribution up to an unknown parameter θ^0 in (1) is sometimes too stringent. We relax it for the case of a fixed design regression model in (2).

2.1. High-dimensional linear model.

Consider the following set-up. The data points $z_i = (y_i, x_i)$ $(i = 1, \ldots, n)$ where $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is a univariate response variable and $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ are p-dimensional covariates. The corresponding random variables as in (1) are denoted by $Z_i = (Y_i, X_i)$.

A linear model is relating Y and X as follows:

$$Y_i = \sum_{j=1}^{p} \beta_j^0 X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n), \tag{2}$$

where the random noise terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2 < \infty$ and ε_i independent of X_i (for all i). The covariates X_i can be random or deterministic, and the responses Y_i are random. We often use the short-hand notation

$$Y = X\beta^0 + \varepsilon,$$

where $Y_{n\times 1} = (Y_1, \ldots, Y_n)^T, \varepsilon_{n\times 1} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and $X_{n\times p} = [X^{(1)}, \ldots, X^{(p)}]$ with $X^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})^T$ is the $n \times 1$ vector of the observed jth covariate. When assuming fixed (deterministic) X_i, the assumption in (1) does not hold: the Y_i's are independent but they have different distributions depending on the value of the X_i's whereas the noise terms are i.i.d. Furthermore, the distribution of the noise does not need to be known. From a methodological point of view, random or fixed covariates do not need a different treatment, and for mathematical theory we can write instead of (1):

$$Y_1, \ldots, Y_n \text{ independent}, \quad Y_i | X_i \sim P_{\theta^0 | X_i},$$

with $\theta^0 = \beta^0$ and where $P_{\theta^0|X_i}$ is given by the distribution of the noise with mean zero which might be known up to unknown parameters (e.g. $\mathcal{N}(0, \sigma^2)$ Gaussian mean zero with unknown variance σ^2) or which might be entirely unknown.

2.1.1. Importance of covariates and interpretation of the regression coefficients.

Like in classical regression one would typically measure the importance of a single covariate $X^{(j)}$ by the magnitude of the absolute value of the regression coefficient $|\beta_j^0|$. This makes only sense after normalization of the covariates to the same (empirical) variance, e.g., the empirical variance of $X^{(j)}$ is equal to one. We implicitly assume such a standardization in the sequel of the article. Note that the importance can be seen as the absolute value of the partial derivative w.r.t. $x^{(j)}$ of the regression function $m(x) = \sum_{j=1}^{p} \beta_j^0 x^{(j)}$ and it can also be interpreted as the change in the regression function $m(x)$ when changing $x^{(j)}$ by one unit.

A stochastic interpretation of the regression coefficient with random X is in terms of partial correlation (which is equal to the conditional correlation in random design Gaussian models). Consider the partial correlation $\mathrm{Parcor}(Y, X^{(j)}|\{X^{(k)};\ k \neq j\})$. Then

$$\beta_j^0 = \mathrm{Parcor}(Y, X^{(j)}|\{X^{(k)};\ k \neq j\}) \frac{\sqrt{K_{j+1,j+1}}}{\sqrt{K_{1,1}}} = -\frac{K_{1,j+1}}{K_{1,1}}, \tag{3}$$

where $K = \Sigma(Y, X)^{-1}$ and $\Sigma(Y, X) = \mathrm{Cov}((Y, X^{(1)}, \ldots, X^{(p)}))$ is the $(p+1) \times (p+1)$ covariance matrix of the variables $Y, X^{(1)}, \ldots, X^{(p)}$. Since the diagonal elements of $K = \Sigma(Y, X)^{-1}$ are non-zero we see that:

$$\beta_j^0 = 0 \iff \mathrm{Parcor}(Y, X^{(j)}|\{X^{(k)};\ k \neq j\}) = 0 \iff K_{1,j+1} = 0.$$

Thus, by the first equivalence above, the regression coefficient measures the strength of linear association of $X^{(j)}$ on Y which has not already been explained by all the other variables $\{X^{(k)};\ k \neq j\}$; and by the second equivalence, whether a regression coefficient is zero is encoded in the *inverse* covariance matrix $K = \Sigma(Y, X)^{-1}$.

The regression coefficient β_j^0 is also linked to direct causal effects, a property which we will exploit in Section 5.1.

2.1.2. Marginal correlations. A simple-minded approach consists of considering marginal correlations

$$\rho_j = \mathrm{Cor}(Y, X^{(j)})\ (j = 1, \ldots, p)$$

and sort the importance of a covariate $X^{(j)}$ for the response Y according to the magnitude of the absolute values $|\rho_j|$.

In terms of the covariance matrix $\Sigma(Y, X)$ in formula (3) we have that

$$\rho_j = \Sigma(Y, X)_{1,j+1} / \sqrt{\Sigma(Y, X)_{1,1} \Sigma(Y, X)_{j+1,j+1}}$$

and the value of ρ_j does not depend on how many and which other covariates $\{X_k;\ k \neq j\}$ are in the model; in particular, the marginal correlations are scaled entries of the covariance matrix $\Sigma(Y, X)$, in contrast to involving its inverse as for the regression coefficients, see formula (3).

Marginal correlations ρ_j may exhibit substantial magnitude due to high correlations among the covariates: for example, if β_1^0 is large and $\beta_2^0 = 0$ but the correlation between $X^{(1)}$ and $X^{(2)}$ is large, then ρ_2 will be large (and ρ_1 as well) despite that $\beta_2^0 = 0$. Under a restrictive assumption on the covariance matrix of the covariates, a sparsity condition on β^0 and a condition requiring that the non-zero elements of β^0 are sufficiently large, it is shown that capturing all covariates with the largest marginal correlations will include the set of all non-zero regression coefficients [25]: the wording "the largest" is depending on a tuning parameter. Such a marginal correlation screening procedure is simple and its corresponding empirical counterpart is easy to be estimated from finite data by using standard empirical correlations between pairs of variables.

One can also perform other variable screening methods: very popular is the one based on ℓ_1-norm regularization, as discussed in Section 2.3.1.

2.2. Estimation in high-dimensional linear models. Consider for simplicity the high-dimensional linear model as in (2). If the unknown parameter of interest has dimensionality $\dim(\beta^0) = p \gg n$, we need to employ some regularization for achieving accurate statistical estimation. A very popular choice is regularized least squares estimation procedures such as the Lasso (**L**east **a**bsolute **s**hrinkage and **s**election **o**perator) based on ℓ_1-norm regularization [52]:

$$\hat{\beta} = \hat{\beta}(\lambda) = \text{argmin}_\beta \left(\|Y - X\beta\|_2^2/n + \lambda\|\beta\|_1 \right), \tag{4}$$

where $\lambda > 0$ is a tuning parameter for choosing the amount of regularization. The superscript "$\hat{}$" denotes throughout this article an estimator (or its realized value based on the data). Historically older is Tikhonov or Ridge regularized least squares:

$$\hat{\beta} = \hat{\beta}(\lambda) = \text{argmin}_\beta \left(\|Y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2 \right). \tag{5}$$

Due to the geometry of the ℓ_1-norm, the Lasso in (4) leads to a sparse estimator where some (or many) components equal exactly zero, i.e., $\hat{\beta}_j(\lambda) = 0$ for some or many indices j. The amount of sparsity depends on the choice of the regularization parameter and the data. This is in contrast to ℓ_2-norm regularization in (5) where all the estimated regression coefficients are different from zero, i.e., $\hat{\beta}_j \neq 0$ for all j.

2.3. Statistical properties of the Lasso. As mentioned above, the Lasso in (4) leads to a sparse estimator. Thus, we would expect that the Lasso performs well if the unknown underlying parameter β^0 is sparse with many entries being exactly zero. And indeed, sparsity of β^0 is a key condition for the developed theory. Consider the following assumptions.

(A1) Sparsity: Denote by $S_0 = \text{supp}(\beta^0) = \{j;\ \beta_j^0 \neq 0\}$ the support of β^0, sometimes also called the active set of the regression parameter β^0, with cardinality $s_0 = |S_0|$. We will often assume an upper bound for s_0, as appearing in e.g. the discussion after Proposition 2.2.

Furthermore, there is a general identifiability problem: for $p > n$, the design matrix has $\text{rank}(X) \leq n < p$ and the null-space of X is not empty. Thus, if we want to infer β^0 from data we need an additional identifiability assumption:

(A2) Compatibility condition [53]: For some $\phi_0 > 0$ and for all β satisfying $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$ it holds that

$$\|\beta_{S_0}\|_1^2 \leq (\beta^T \hat{\Sigma} \beta) s_0/\phi_0^2,$$

where $\hat{\Sigma} = X^T X/n$ and β_S, for an index set $S \subseteq \{1, \ldots, p\}$, has elements set to zero outside the set S, i.e., $(\beta_S)_j = 0$ $(j \notin S)$ and $(\beta_S)_j = \beta_j$ $(j \in S)$. The value $\phi_0 > 0$ is called the compatibility constant.

The factor 3 in the definition of (A2) is not crucial and $\hat{\Sigma}$ can be thought of as an empirical covariance matrix of the p covariates. Assumption (A2) is sometimes called ℓ_1 restricted eigenvalue condition for $\hat{\Sigma}$: note the analogy to the classical concept of the smallest eigenvalue $\lambda^2_{\mathrm{min}}(\hat{\Sigma})$ which satisfies for all β:

$$\|\beta\|_2^2 \le \beta^T \hat{\Sigma}\beta/\lambda^2_{\mathrm{min}}(\hat{\Sigma}).$$

(The additional factor s_0 in (A2) occurs due to using the ℓ_1- instead of the ℓ_2-norm). Since $\hat{\Sigma}$ is singular (due to $p > n$) we have that $\lambda^2_{\mathrm{min}}(\hat{\Sigma}) = 0$ and the bound above is uninformative. That is why one introduces the concept of *restricted* eigenvalues where lower-bounding the quadratic form $\beta\hat{\Sigma}\beta$ has only to hold for a restricted set of β satisfying the cone condition $\|\beta_{S_0^c}\|_1 \le 3\|\beta_{S_0}\|_1$. We will give in Proposition 4.1 a sufficient condition to ensure that the compatibility condition holds. We also note that the compatibility condition is weaker than requiring the celebrated restricted isometry property from compressed sensing [13, cf.] or than the restricted ℓ_2-eigenvalue [1]: a comparison of these conditions has been given in [55].

Asymptotics. Throughout the paper, all asymptotic statements are for a scenario where both p and n tend to infinity, allowing for $p \gg n$. Thereby, the model e.g. as in (2) changes, and we thus always adopt a "changing model" (sometimes called "triangular array") asymptotics.

The following oracle inequality is derived in [8, Th.6.1]:

Theorem 2.1. *Consider a linear model as in* (2) *with fixed design X and consider the sparsity s_0 in (A1) and the compatibility constant $\phi_0 > 0$ in (A2). Then, on the event $\mathcal{F} = \{\max_{j=1,\ldots,p} 2|\varepsilon^T X^{(j)}/n| \le \lambda_0\}$ (see Proposition* 2.2 *below) and when using the Lasso in* (4) *with $\lambda \ge 2\lambda_0$ it holds :*

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \le 4\lambda^2 s_0/\phi_0^2.$$

Furthermore, the probability for the event $\mathcal{F}$ can be lower-bounded for i.i.d. Gaussian errors [8, Lem.6.2].

Proposition 2.2. *Consider a linear model as in* (2) *with fixed design and Gaussian errors $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Assume that the covariates are scaled such that $\hat{\Sigma}_{j,j} = 1$ for all j, where $\hat{\Sigma} = X^T X/n$. Then, for $\lambda_0 \ge 2\sigma\sqrt{\frac{t^2+2\log(p)}{n}}$, $t > 0$,*

$$\mathbb{P}[\mathcal{F}] \ge 1 - 2\exp(-t^2/2).$$

Proposition 2.2 says that for Gaussian errors, the regularization parameter should be chosen as $\lambda = 2\lambda_0 \asymp \sqrt{\log(p)/n}$ so that $\mathbb{P}[\mathcal{F}]$ becomes overwhelmingly large. With that rate we obtain from Theorem 2.1, assuming ϕ_0 in (A2) is bounded away from zero:

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P(s_0 \log(p)/n),$$
$$\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n}).$$

The first bound establishes near optimality for estimating the regression surface $X\beta^0$ and for predicting new observations: would we know the set of true relevant variables S_0, the optimal rate is $O_P(s_0/n)$ and thus, we pay a price of $\log(p)$ by not knowing which of the variables have non-zero regression coefficients. The second bound establishes a bound for the estimation error for the unknown high-dimensional parameter β^0 in terms of the ℓ_1-norm. For this (and the former) to converge to zero, we need that the sparsity satisfies $s_0 = o(\sqrt{n/\log(p)})$. This can be improved to $s_0 = o(n/\log(p))$ when looking at the estimation error $\|\hat{\beta} - \beta^0\|_2$ and when slightly strengthening assumption (A2) to an ℓ_2 restricted eigenvalue condition [1].

Proposition 2.2 can be generalized in various directions, dealing e.g. with correlated error terms, and with non-Gaussian errors. Regarding the latter, one can assume independent errors ε_i with mean zero and finite second moments, and bounded covariates $\max_{i=1,\ldots,n;j=1,\ldots,p} |X_i^{(j)}| \le M < \infty$ for some constant M. Then, the same qualitative statement as in Proposition 2.2 holds [8, Ex.14.3].

We now discuss a scenario where the compatibility assumption (A2) holds. Consider the following setting:

(A3) The rows of X are arising from i.i.d. sampling from a probability distribution with covariance matrix Σ. Once they are sampled, they are considered as fixed so that X is a fixed design matrix as assumed in Theorem 2.1.

The following proposition is given in asymptotic form where $p \ge n \to \infty$ and is a consequence of the non-asymptotic result in [8, Cor.6.8].

Proposition 2.3. *Consider the setting in (A3) with sub-Gaussian random variables $X_i^{(j)}$ with bounded sub-Gaussian norms and with smallest eigenvalue for the covariance matrix $\lambda_{\min}(\Sigma) \ge L > 0$. Assume that $s_0 = o(\sqrt{n/\log(p)})$. Then, with probability (w.r.t. sampling the rows of X) tending to one as $p \ge n \to \infty$, the matrix $\hat{\Sigma}$ satisfies (A2) with $\phi_0^2 \ge L/2$.*

Proposition 4.1 implies the somewhat “obscure” compatibility condition (A2), under the assumptions of sparsity in (A1), sufficiently nice distribution and sufficiently nice behavior of the population covariance matrix Σ (which has no relation to the sample size of observed data) in the setting (A3).

2.3.1. Variable screening and selection.

Theorem 2.1 is about estimation of the regression surface $X\beta^0$ (w.r.t. ℓ_2-norm) and of the regression parameter vector β^0 with respect to the ℓ_1-norm. Since the Lasso is a sparse estimator one would hope for good variable selection properties: denote by $\hat{S} = \{j; \hat{\beta}_j \neq 0\}$ the support of the estimated parameter vector $\hat{\beta}$. We would hope that

$$\hat{S} = S^0 \text{ with high probability,} \tag{6}$$

or being at least approximately equal. Such a property of correct variable selection or support recovery in (6) is often too ambitious in practice. The problem is mathematically well understood as there are two sufficient and essentially necessary conditions: the first one concerns that the non-zero coefficients should be sufficiently large, namely a so-called beta-min condition

$$\min\{|\beta_j^0|;\ \beta_j^0 \neq 0\} = \min_{j \in S_0^c} |\beta_j^0| \geq C(s_0, p, n), \quad C(s_0, p, n) \asymp \sqrt{s_0 \log(p)/n}, \tag{7}$$

and the second one is a condition on the design matrix X, a so-called irrepresentable condition which has been given first in [44, 63, 61]. While a beta-min condition as in (7) is unavoidable, at least requiring a lower bound of the order $\sqrt{\log(p)/n}$ [39, cf.], the second irrepresentable condition on the design is rather strong. When relaxing such an irrepresentable condition, we loose the exact variable selection property in (6) but still obtain a very useful screening property, saying that

$$\hat{S} \supseteq S_0 \text{ with high probability.} \tag{8}$$

The variable screening property holds on the set $\mathcal{F}$ in Theorem 2.1, assuming the compatibility condition (A2) on the design X and a beta-min condition with $C(s_0, p, n) > 4\lambda s_0/\phi_0^2 \asymp s_0\sqrt{\log(p)/n}$ where for the latter asymptotic relation we implicitly require that $\phi_0^2 \geq L > 0$ is bounded away from zero. In fact, such a result is an immediate consequence of Theorem 2.1 using the bound for $\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda s_0/\phi_0^2$. When assuming a slightly stronger restricted eigenvalue condition on X, a beta-min condition with $C(s_0, p, n) \asymp \sqrt{s_0 \log(p)/n}$ is sufficient for the screening property in (8) [1, cf.].

The variable screening property in (8) is a highly interesting dimension reduction result: assuming a compatibility condition like (A2), the cardinality of the set of Lasso-estimated variables is $|\hat{S}| = |\hat{S}(\lambda)| \leq \min(n, p)$ for all parameters λ. Thus, thanks to (8), we have reduced the number of variables to at most $min(n, p)$ without having dropped an active variable from S_0 which is a massive dimensionality reduction if $p \gg n$ (e.g. a reduction from $p \approx 10^6$ to $n \approx O(10^3)$ variables in datasets as discussed in Section 5); that is, the Lasso would keep all variables from S_0 and would discard many noise variables form S_0^c.

While correct variable selection (6) with the Lasso necessarily requires strong conditions, variable screening (8) can be established under weaker assumptions on the design X (namely a compatibility instead of an irrepresentable condition). The name Lasso is a shortcut for Least Absolute Shrinkage and Selection Operator [52], but in view of the mentioned results above, it should rather be translated as Least Absolute Shrinkage and **S**creening Operator [3].

Sufficient conditions for (8) have been mentioned above: they might still be rather strong involving a beta-min assumption. Empirical evidence suggests that the Lasso and other sparse estimators are often rather "unstable" indicating that variable screening is a too ambitious goal. To quantify stability and the relevance of selected variables, stability selection [45] has been proposed, and the technique in connection with the Lasso is nowadays used in a wide range of applications. Other methods to quantify uncertainty and reliability are discussed in Section 3.

2.3.2. Main ideas to prove Theorem 2.1 and Proposition 2.2. The key step is to separate the stochastic and deterministic part: in the linear model with fixed design, the only stochastic element is the noise term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ which arises in the event $\mathcal{F}$.

Sketch of the proof for Theorem 2.1. The algebraic manipulations go as follows. Using the fact the $\hat{\beta}$ minimizes the ℓ_1-norm regularized squared error we get

$$\|Y - X\hat{\beta}\|_2^2/n + \lambda\|\hat{\beta}\|_1 \leq \|Y - X\beta^0\|_2^2/n + \lambda\|\beta^0\|_1$$

and from this the so-called Basic Inequality [8, Lem.6.1]:

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta}\|_1 \leq 2\varepsilon^T X(\hat{\beta} - \beta^0)/n + \lambda\|\beta^0\|_1.$$

One the right-hand side, the stochastic part can be easily bounded as follows:

$$|2\varepsilon^T X(\hat{\beta} - \beta^0)/n| \leq \max_{j=1,\ldots,p} |2\varepsilon^T X^{(j)}/n| \|\hat{\beta} - \beta^0\|_1,$$

and thus on $\mathcal{F}$:

$$|2\varepsilon^T X(\hat{\beta} - \beta^0)/n| \leq \lambda_0\|\hat{\beta} - \beta^0\|_1.$$

It is at this first step where the stochastic part is separated by using the event $\mathcal{F}$.

With some further elementary operations involving the triangle inequality for the ℓ_1-norm we obtain [8, Lem.6.3]: on the event $\mathcal{F}$ and for $\lambda \geq 2\lambda_0$,

$$2\|X(\hat{\beta} - \beta^0)\|_2^2/n + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta^0_{S^0}\|_1.$$

We note that $\|X(\hat{\beta} - \beta^0)\|_2^2/n = (\hat{\beta} - \beta^0)^T\hat{\Sigma}(\hat{\beta} - \beta^0)$. Thanks to the compatibility condition (A2), we can thus relate $\|\hat{\beta}_{S_0} - \beta^0_{S^0}\|_1$ (from the right-hand side) to $\|X(\hat{\beta} - \beta^0)\|_2^2/n$ (from the left-hand side), namely:

$$\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 \leq \left(\|X(\hat{\beta} - \beta^0)\|_2^2/n\frac{s_0}{\phi_0^2}\right)^{1/2}, \tag{9}$$

by showing that $\hat{\beta} - \beta^0$ satisfies the cone condition on $\mathcal{F}$). Simple algebra then leads to the statement in Theorem 2.1. The sparsity s_0 from assumption (A1) and the compatibility condition (A2) come only once into play, namely in the bound in (9).

Sketch of the proof for Proposition 2.2. Due to the Gaussian assumption $V_j = \varepsilon^T X^{(j)}/(\sigma\sqrt{n}) \sim \mathcal{N}(0, 1)$. Although the V_j's are dependent, we can use the union bound to obtain

$$\mathbb{P}[\max_{j=1,\ldots,p} |V_j| > c] \leq 2p\exp(-c^2/2).$$

This then leads to the statement of Proposition 2.2.

3. Assigning uncertainty

It is often of interest to assign uncertainties for inferring the true underlying regression coefficients β^0_j $(j = 1,\dots,p)$ in the linear model (2). The classical concept of confidence intervals for single coefficients β^0_j or confidence regions for a set $G \subseteq \{1,\dots,p\}$ of coefficients $\{\beta^0_j;\ j \in G\}$ can capture this. Somewhat easier is the case with statistical hypothesis testing for single or group hypotheses:

$$\text{single: } H_{0,j} : \beta^0_j = 0 \text{ versus } H_{A,j} :\ \beta^0_j \neq 0, \tag{10}$$

$$\text{group: } H_{0,G} :\ \beta^0_j = 0\ \forall\ j \in G \text{ versus } H_{A,G} :\ \beta^0_j \neq 0 \text{ for at least one } j \in G. \tag{11}$$

These hypotheses are rather natural and often of major interest, namely for inferring whether a regression coefficient is zero (i.e., the corresponding covariate has no effect) or not.

A result as in Theorem 2.1 does not quantify uncertainty in a refined way by considering a properly normalized limiting distribution of $\hat{\beta} - \beta^0$. Such a limiting distribution for the Lasso is very difficult to derive in the high-dimensional setting: with a fixed dimension asymptotics where p remains fixed and $n \to \infty$, the problem has been solved with a non-continuous limiting distribution (due to the sparsity of the estimator) [38]. Owing to this simple case and due to the non-continuity of the limit one has to accept that an undesirable super-efficiency phenomenon will arise: the Lasso would show good asymptotic behavior to estimate the regression parameters whose values equal zero, and it will be rather poor for the non-zero parameters. This is in analogy to Hodges example of a super-efficient estimator. To circumvent this problem, we will look at a "de-biased" or "de-sparsified" Lasso estimator in Section 3.1 having a Gaussian (and hence smooth) limiting distribution.

3.1. The de-sparsified Lasso. The de-sparsified Lasso [56], originally introduced under the name de-biased Lasso [60], is an estimator which exhibits optimal performance for estimating low-dimensional components of β^0, as described in Theorem 3.1 and (14) or (15). A pragmatic motivation is given by least squares estimation in the low-dimensional case $p < n$ with $\mathrm{rank}(X) = p \le n$:

$$\hat{\beta}_{\mathrm{LS}} = \mathrm{argmin}_\beta \|Y - X\beta\|_2^2/n.$$

Then, it holds that for the jth component,

$$\hat{\beta}_{\mathrm{LS},j} = \frac{Z^T_{\mathrm{LS},j} Y}{Z^T_{\mathrm{LS},j} X^{(j)}}\ (j = 1,\dots,p),$$

where $Z_{\mathrm{LS},j}$ is the residual vector of least squares regression of $X^{(j)}$ versus $\{X^{(k)};\ k \neq j\}$. If $p > n$, and thus $\mathrm{rank}(X) < p$, this construction is not possible since $Z_{\mathrm{LS},j}$ would be the zero-vector. The idea is then to replace the least squares regression in the construction of $Z_{\mathrm{LS},j}$ by a Lasso regression:

$$Z_j = X^{(j)} - X_{-j}\hat{\gamma}_j,$$
$$\hat{\gamma}_j = \mathrm{argmin}_{\gamma \in \mathbb{R}^{p-1}}(\|X^{(j)} - X_{-j}\gamma\|_2^2/n + \lambda_X\|\gamma\|_1),$$

where X_{-j} is the $n \times (p-1)$ sub-matrix of X without its jth column. Analogous to the least squares representation above, we consider

$$\frac{Z_j^T Y}{Z_j^T X^{(j)}} = \beta_j^0 + \underbrace{\sum_{k \neq j} \frac{Z_j^T X^{(k)}}{Z_j^T X^{(j)}}\beta_k^0}_{\text{bias}} + \underbrace{\frac{Z_j^T \varepsilon}{Z_j^T X^{(j)}}}_{\text{noise}}. \tag{12}$$

Since there is now a bias (in contrast to least squares in $p < n$ settings where $Z_{\mathrm{LS},j}^T X^{(k)} = 0$ $(k \neq j)$ due to orthogonality of the projection), we correct this bias term by plugging in the Lasso estimator $\hat{\beta}$:

$$\hat{b}_j = \frac{Z_j^T Y}{Z_j^T X^{(j)}} - \sum_{k \neq j} \frac{Z_j^T X^{(k)}}{Z_j^T X^{(j)}}\hat{\beta}_k \ (j = 1, \ldots, p). \tag{13}$$

The estimator $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_p)^T$ is called the de-sparsified Lasso, owing its name to the fact that $\hat{b}_j \neq 0$ for all j. This estimator involves two tuning parameters, namely λ_X in the construction of Z_j and λ for the Lasso estimator $\hat{\beta}$.

One can derive a result as follows.

Theorem 3.1. *Consider a linear model as in* (2) *with fixed design X and Gaussian errors $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Then, the de-sparsified Lasso estimator in* (13) *satisfies*

$$\sqrt{n}(\hat{b}_j - \beta_j^0) = W_j + \Delta_j \ (j = 1, \ldots, p),$$
$$(W_1, \ldots, W_p)^T \sim \mathcal{N}_p(0, \sigma^2\Omega), \ \Omega_{j,k} = \frac{n^{-1}Z_j^T Z^{(k)}}{n^{-1}Z_j^T X^{(j)} n^{-1} Z_k^T X^{(k)}}$$
$$\max_{j=1,\ldots,p} |\Delta_j| \leq \frac{\sqrt{n}\lambda_X}{\min_{j=1,\ldots,p}|n^{-1}Z_j^T X^{(j)}|}\|\hat{\beta} - \beta^0\|_1.$$

The issue is now to show that $\max_{j=1,\ldots,p}|\Delta_j| = o_P(1)$ which would then imply a Gaussian limit or functions thereof, see (17). We know from Theorem 2.1 using $\lambda \asymp \sqrt{\log(p)/n}$ that for sparse settings as formulated in assumption (A1) and assuming the compatibility condition (A2) we have that

$$\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n}).$$

Furthermore, one can argue that $\min_{j=1,\ldots,p}|n^{-1}Z_j^T X^{(j)}| \geq L > 0$ is bounded from below when adopting the setting as in (A3), using $\lambda_x \asymp \sqrt{\log(p)/n}$ and assuming regularity conditions. This then leads to

$$\max_{j=1,\ldots,p}|\Delta_j| = O_P(\sqrt{n}\lambda_X s_0\sqrt{\log(p)/n}) = O_P(s_0\log(p)/\sqrt{n}),$$

and the right-hand side converges to zero if $s_0 = o(\sqrt{n}/\log(p))$. Summarizing, assuming (A1) with $s_0 = o(\sqrt{n}/\log(p))$, (A2), (A3) and Gaussian errors (and some additional minor conditions) we obtain

$$\sqrt{n}(\hat{b} - \beta^0) = W + \Delta,\ W \sim \mathcal{N}_p(0, \sigma^2 \Omega),\ \max_{j=1,\dots,p} |\Delta_j| = o_P(1). \tag{14}$$

Furthermore, when assuming sparsity of Σ^{-1} in (A3) one can establish optimality saying that Ω is asymptotically the smallest covariance matrix among all regular estimators [60, 56], that is, it reaches the so-called semiparametric information bound [2, cf.].

We note that the scaled version is sometimes a better representation and preferentially used for simultaneous inference (see later (18)):

$$\sqrt{n}(\hat{b}_j - \beta_j^0)/\sqrt{\Omega_{j,j}} = \frac{|Z_j^T X_j/n|}{\|Z_j\|_2/\sqrt{n}} \sqrt{n}(\hat{b}_j - \beta_j^0) = \frac{|Z_j^T X_j/n|}{\|Z_j\|_2/\sqrt{n}} W_j + \tilde{\Delta}_j,$$

where

$$\tilde{\Delta}_j = \frac{|Z_j^T X_j/n|}{\|Z_j\|_2/\sqrt{n}} \Delta_j,\quad \max_{j=1,\dots,p} |\tilde{\Delta}_j| \le \lambda_X \frac{1}{\|Z_j\|_2/\sqrt{n}} \|\hat{\beta} - \beta^0\|_1 = o_P(1), \tag{15}$$

assuming the conditions above for the last bound in the second line. In the following, we often use the scaled version.

Formula (14) and (15) establish an asymptotic pivotal property of the desparsified Lasso estimator $\hat{b}$. It can be directly used to construct confidence intervals or statistical hypothesis tests for single parameters β_j^0 in the usual fashion, observing that σ^2 can be estimated as

$$\hat{\sigma}^2 = \|Y - X\hat{\beta}\|_2^2/n,$$

or using the factor $(n - \hat{s})^{-1}$ instead of n^{-1}, where $\hat{s} = |\mathrm{supp}(\hat{\beta})|$ denotes the number of non-zero estimated regression coefficients. A two-sided confidence interval with coverage level $1 - \alpha$ for a single parameter β_j^0 is

$$\hat{b}_j \pm n^{-1/2} \Phi^{-1}(1 - \alpha/2) \hat{\sigma} \sqrt{\Omega_{j,j}}. \tag{16}$$

3.1.1. Simultaneous inference and multiple testing adjustment.

Formula (15) implies that we can do simultaneous inference with respect to the sup-norm. That is, for *any* group $G \subseteq \{1, \dots, p\}$, regardless of its cardinality,

$$\max_{j \in G} |\sqrt{n}(\hat{b}_j - \beta_j^0)/\sqrt{\Omega_{j,j}}| = \max_{j \in G} |W_j/\sqrt{\Omega_{j,j}}| + o_P(1). \tag{17}$$

Simultaneous hypothesis tests. As a concrete example, consider a hypothesis $H_{0,G}$ for a group $G \subseteq \{1, \dots, p\}$ as described in (11). Formula (15) then implies: under $H_{0,G}$,

$$\max_{j \in G} |\hat{b}_j/\sqrt{\Omega_{j,j}}| = \max_{j \in G} |W_j/\sqrt{\Omega_{j,j}}| + o_P(1), \tag{18}$$

and this can be used to construct a statistical test for $H_{0,G}$. The distribution of $\max_{j \in G} |W_j/\sqrt{\Omega_{j,j}}|$ is difficult to derive analytically but it can be easily simulated and estimated:

$$\mathbb{P}[\max_{j \in G} |W_j/\sqrt{\Omega_{j,j}}| \le c] = G(c/\sigma) = G(c/\hat{\sigma}) + o_P(1)$$

where $G(c) = \mathbb{P}[\max_{j \in G} |V_j/\sqrt{\Omega_{j,j}}| \le c]$ with $V_j = W_j/\sigma$ is straightforward to simulate since the covariance structure of $\{V_j;\ j \in G\}$ is known. We would expect good power of the test statistics $\max_{j \in G} |\hat{b}_j/\sqrt{\Omega_{j,j}}|$ if the alternative is sparse with only a few $\beta_j^0 \ne 0$ for $j \in G$.

Multiple testing adjustment. Formula (18) also allows for multiple testing adjustment among m tests, with respect to the familywise error rate (FWER) defined as

$$\text{FWER} = \mathbb{P}[\text{at least one false positive among } m \text{ hypothesis tests}].$$

When performing m hypothesis tests, it is important to adjust for multiplicity. If each of these m tests is performed at significance level α, it holds that FWER $\le m\alpha$. This is due to the union bound but might be rather conservative in general. However, if the tests were independent, then FWER $= (1 - (1 - \alpha)^m) \approx m\alpha$ which essentially achieves the upper bound. We can simply adjust the statistical tests by dividing the significance level by m or multiplying the corresponding p-values by m: this is known as the Bonferroni correction. However, for dependent tests, such a Bonferroni correction is overly conservative. When using the exact distribution in (18) one obtains much more powerful multiple testing adjustment for controlling the familywise error rate: this is essentially the Westfall-Young procedure [58] which has been shown to be optimal in certain settings [47]. Thus, multiple testing adjustment based on the simultaneous pivot as in (18) leads to a procedure which controls the familywise error rate and yet it has good power, particularly for settings with dependent tests and in comparison to a Bonferroni (or Bonferroni-Holm) multiple testing adjustment.

As we have seen above, simultaneous inference with respect to the sup-norm is a straightforward implication of Theorem 2.1 and (14). When looking at other norms such as the ℓ_2-norm, i.e.,

$$\|\sqrt{n}(\hat{b}_G - \beta_G^0)\|_2 = \sqrt{\sum_{j \in G} (\sqrt{n}(\hat{b}_j - \beta_j^0))^2},$$

the cardinality $|G|$ plays an important role. (Of course, we may preferentially consider the ℓ_2-norm of the scaled version which is $\sqrt{\sum_{j \in G} (\sqrt{n}(\hat{b}_j - \beta_j^0)/\sqrt{\Omega_{j,j}})^2}$). The remainder term Δ in Theorem 2.1 is not controlled for the ℓ_2-norm and this implies that a χ^2-approximation for $\|\sqrt{n}(\hat{b}_G - \beta_G^0)\|_2^2$ is only valid for G with cardinality $|G|$ sufficiently slowly growing with n, and in particular that $|G| = o(n)$ is necessary; we refer the reader to [54, Th.5.3].

3.1.2. Non-Gaussian and heteroscedastic errors. Analogues of Theorem 3.1 and its consequences in (14) and (15) can also be established for non-Gaussian errors in the model (2). The details are given in [20]. We invoke a Lindeberg central limit theorem for the noise term in (12)

$$n^{-1/2} Z_j^T \varepsilon \ (j = 1, \ldots, p) \tag{19}$$

by assuming $2 + \delta$ ($\delta > 0$) moments for ε_i. If $G \subseteq \{1, \ldots, p\}$ is a fixed subset we obtain as an analogue to (14) and (15), with W as specified there:

$$\{\sqrt{n}(\hat{b}_j - \beta_j^0)/\sqrt{\Omega_{j,j}};\ j \in G\} \Longrightarrow \{W_j/\sqrt{\Omega_{j,j}};\ j \in G\}.$$

The difficulty is to derive a high-dimensional limiting distribution when $|G|$ is large. A result for the sup-norm is possible: for essentially any $G \subseteq \{1, \ldots, p\}$,

$$\max_{j \in G} |\sqrt{n}(\hat{b}_j - \beta_j^0)/\sqrt{\Omega_{j,j}}| \Longrightarrow \max_{j \in G} |W_j/\sqrt{\Omega_{j,j}}|, \tag{20}$$

with W_j as in (14). This builds on work about Gaussian approximations of maxima of sums of random vectors [18] and requires a slight restriction on the cardinality of G, namely $\log(|G|) = o(n^{1/7})$ (besides sparsity in (A1) and the compatibility condition in (A2)). We note that (20) coincides with (17) and hence, simultaneous inference as discussed in Section 3.1.1 is also possible and guaranteed for non-Gaussian errors.

3.2. The bootstrap and heteroscedastic errors. The bootstrap is a simulation technique for approximating the distribution of an estimator or the output of an algorithm. If we knew the data-generating distribution of the data, we could derive the distribution of the estimator of interest by simulation. Since we do not have access to the data generating distribution, one can plug-in some form of the empirical distribution based on the data, and this is called the bootstrap [23]. Often, this amounts to resampling the data (n times with replacement) and this in turn is the reason for the name "bootstrap": it looks paradoxical in analogy to the story about Barron Münchhausen who tries to escape from sinking into the swamp by pulling himself out using his own bootstraps. Unfortunately, the simple bootstrap idea is not always consistent: essentially, the bootstrap leads to a consistent (asymptotically correct) approximation if the corresponding limiting distribution of the estimator is Gaussian [28, 29]. Thus, applying the bootstrap directly to the Lasso or another sparse estimator would lead to inconsistent results; or under strong (unrealistic assumptions) it is shown to lead to a valid asymptotic approximation [16] but exhibiting poor performance due to the super-efficiency phenomenon mentioned already in Section 3, see [19].

Bootstrapping the de-sparsified estimator is a natural method for estimating the limiting distribution in (14), (15), (17) or (18). From a practical point of view, this means to use a computational algorithm only without the need to rely on some analytic form for the maximum of many dependent (limiting) Gaussian variables: the method is not only convenient and easy to use, but it often also leads to better results than using Gaussian approximations since (i) a linearization with the remainder term Δ in Theorem 3.1 is circumvented and (ii) a Gaussian approximation for (19) might be too crude for finite sample size in presence of strongly non-Gaussian errors. Furthermore, much in the spirit of multiple testing adjustment mentioned in Section 3.1.1, the bootstrap provides a powerful way for simultaneous inference where the dependence structure of $\sqrt{n}(\hat{b}_j - \beta_j^0)/\sqrt{\Omega_{j,j}}$ across j is automatically taken into account.

The bootstrap technique can also cope well with heteroscedastic errors, i.e., $\varepsilon_1, \ldots, \varepsilon_n$ in (2) are independent but with different variances $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 \neq$ constant. This extension to a more general structure of the error terms is important for many practical applications.

Overall, bootstrapping the de-sparsified estimator leads to a reliable and automatic procedure for possibly simultaneous inference of β^0 in the model (2), and its advantages are particularly visible for non-Gaussian or heteroscedastic errors. We refer the reader for more details to [20].

3.3. Other approaches. Other methods have been proposed for statistical inference in high-dimensional settings. Related to the de-sparsified Lasso is a method from [35], exhibiting some optimality without requiring sparsity of Σ^{-1} in (A3). Earlier work in [4] proposes a projection with ℓ_2-norm regularization instead of using the projection onto Z_j as in (13): empirical results suggest that this method is conservative but often more reliable for controlling the probability of false positives [19].

An earlier idea is based on sample splitting [57] which has been improved to multi sample splitting in [48]. The approach is based on a two-stage approach: in a first stage, variable screening is performed with e.g. the Lasso, see Section 2.3.1, and statistical inference is then done in a second stage with the selected variables only. The latter step is low-dimensional by requiring that the number of selected variables in the first stage is smaller than sample size (which automatically holds for the Lasso). To avoid overoptimistic p-values and confidence intervals one should not perform the statistical inference using the same data-points which have been used for screening the important variables: sample splitting or a more sophisticated multi sample splitting technique avoid such a pitfall. From a computational view-point, the (multi) sample splitting method is very attractive since the statistical inference task has to be computed for a low-dimensional problem only while the screening in stage two can be computed rather efficiently. This advantage is important for applications like genome-wide association studies where the number of covariates is in the order of $O(10^6)$, see Section 5. The developed theory for (multi) sample splitting techniques essentially requires that all relevant variables from S^0 are selected in the first stage: this property necessitates a beta-min assumption as in (7), in contrast to e.g. the de-sparsified Lasso.

4. Hierarchical inference

The methodology presented in Section 3 is the building block for powerful statistical inference in presence of strongly correlated covariates as is often the case in very high-dimensional problems. When testing the effects of single covariates and considering the hypotheses $H_{0,j}$ in (10), it happens rather frequently that none or only very few statistically significant variables are found, see for example [5] for applications in biology. This is mainly due to near non-identifiability of a single covariate since its effect can be essentially explained by a few others (due to high correlation or near linear dependence), see also Section 2.1.1. In contrast, a group G of covariates is often easier to be detected as significant with respect to the null-hypothesis $H_{0,G}$ in (11).

Testing of such groups can be done in a hierarchical manner enabling computationally and statistically efficient multiple testing adjustment. A hierarchy is given in terms of a tree-structure, often but not necessarily a binary tree. The nodes in the tree correspond to groups of variables: a child G' (by going from top to bottom in the hierarchical order of the tree) of a node or group G has the property that $G' \subset G$, and the children of a node or group G build a partition of G. Usually, the top node contains all the variables $\{1, \ldots, p\}$ and the p nodes at the very bottom of the tree correspond to single variables, i.e., $\{1\}, \{2\}, \ldots, \{p\}$, see Figure 1.

Together with the identifiability issue mentioned at the beginning of the section, we aim to construct a tree such that highly correlated variables are in the same groups: this can be achieved by a standard hierarchical clustering algorithm [32, cf.], for example using average linkage and the dissimilarity matrix given by $1-$ (empirical correlation)^2. Other clustering algorithms can be used, for example based on canonical correlation [7], or one can rely on any pre-specified hierarchical tree structure.

The key idea is to pursue testing of the groups in a sequential fashion, starting with the top node and then successively moving down the hierarchy until a group doesn't exhibit a significant effect. Figure 2 illustrates this point, showing that we might proceed rather deep in the hierarchy at some parts of the tree whereas at other parts the testing procedure stops due to a group which is not found to exhibit a significant effect. We need some multiple testing adjustment of the p-values: interestingly, due to the hierarchical nature, it is not overly severe at the upper parts of the hierarchy as we will describe next.

Denote the nodes in the tree by G and the corresponding group null-hypotheses by $H_{0,G}$ (which do not necessarily have to be of the form as in (11)). Denote by $d(G)$ the level of the tree of the node (or group) G and by $n(G)$ the number of nodes at level $d(G)$: for example, when $G = \{1, \ldots, p\}$ corresponds to the top node in a cluster tree containing all variables, we have that $d(G) = 1$ and $n(G) = 1$. We only correct for multiplicity in a depth-wise manner in the tree:

$$P_{G;\mathrm{adjusted}} = P_G \cdot p/|G|, \tag{21}$$

see Figure 1 for an illustration why this is a depth-wise Bonferroni correction if the groups are balanced. More generally one can use

$$P_{G;\text{adjusted}} = P_G \cdot c(G), \quad \sum_{G' \in \mathcal{P}} \frac{1}{c(G')} = 1, \tag{22}$$

where $\mathcal{P}$ denotes any partition of $\{1, \dots, p\}$ with sets from the hierarchical tree. The proposal in (21) is a special case of this rule with $c(G) = p/|G|$. If the tree has the same number of offspring (e.g. a binary tree with two offspring throughout the entire tree) we could also use the unweighted version,

$$\text{depth-wise Bonferroni correction: } P_{G;\text{adjusted}} = P_G \cdot n(G).$$

The sequential nature with stopping can be formulated in terms of p-values by adding a hierarchical constraint:

$$P_{G;\text{hierarchically}-\text{adjusted}} = \max_{G' \supset G} P_{G',\text{adjusted}}, \tag{23}$$

implying that once we stop rejecting a node, we cannot reject further down in the tree hierarchy and thus, we can simply stop the procedure when a node is not found as being significant. The following then holds.

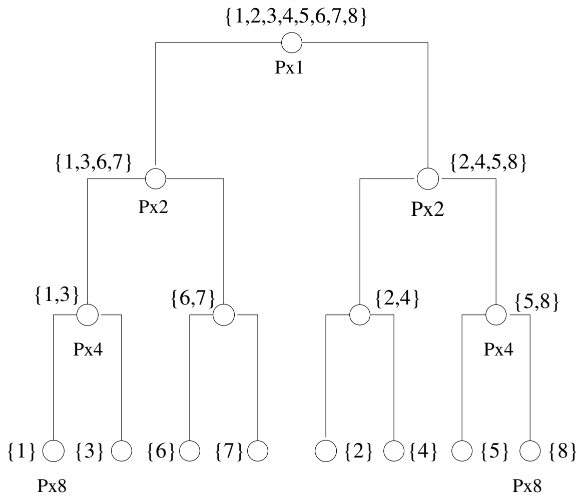

Figure 1. Hierarchical grouping of 8 variables where different groups are denoted by $\{\dots\}$. The capital letter "P" is a generic notation for the raw p-value corresponding to a group hypothesis $H_{0,G}$ of a group G, which is then adjusted as in (21). Since the hierarchy has the same number of offspring throughout the tree, the adjustment is the depth-wise Bonferroni correction which amounts to multiply the p-values in every depth of the tree by the number of nodes in the corresponding depth; no multiplicity adjustment at the top node, then multiplication by the factor 2 (depth 2), 4 (depth 3), and 8 (depth 4).

Proposition 4.1. *Consider an arbitrary hierarchy of hypotheses tests in terms of a tree structure. Consider the procedure described above with depth-wise adjustment as in* (21) *or* (22) *and with hierarchy constraint as in* (23)*. Then, the familywise error rate (FWER) is controlled: that is, for* $0 < \alpha < 1$*, when rejecting a hypothesis* $H_{0,G}$ *if and only if* $P_{G;\mathrm{hierarchically-adjusted}} \le \alpha$*, we have that* $\mathrm{FWER} = \mathbb{P}[$*at least one false rejection*$] \le \alpha$*.*

Proof. Consider first a slightly different setting of hierarchical testing of m null-hypotheses $H_1 \prec H_2 \prec H_m$, where "$H_i \prec H_j$" denotes that H_i is tested first before H_j. (These hypotheses correspond to the m different depths of the tree considered in Proposition 4.1). The hierarchical testing rule stops considering further tests if some hypothesis cannot be rejected: that is, if H_j is not rejected by the test (i.e., the corresponding $P_j > \alpha$ for some significance level α), all subsequent $H_{j+1} \prec H_{j+2} \prec \ldots \prec H_m$ will not be considered further (i.e., will not be rejected).

Consider any constellation of the m hypotheses being true or false, e.g. for $m = 5$,

$$H_1 = F, H_2 = F, H_3 = F, H_4 = T, H_5 = F, H_6 = T.$$

A false positive decision can only arise when falsely rejecting the *first* true hypothesis. This holds because: (i) obviously, we only get a false positive rejection if a hypothesis is true; (ii) if the test does not reject the first true hypothesis, the procedure will stop (due to the hierarchical constraint) and would not produce a false positive. In the example above, this means that we only need to control the probability of falsely rejecting H_4. In general, denote by

$$j^* = \mathrm{argmin}_j\{H_j = T \text{ and } H_i = F \text{ for all } i = 1, \ldots, j-1\},$$

with the meaning that $j^* = 1$ if $H_1 = T$ and $j^* = m + 1$ if all $H_i = F$ for $i = 1, \ldots, m$. Thus, by the argument above, the familywise error rate can be bounded as follows: if $j^* \le m$,

$$\begin{aligned}
\mathrm{FWER} &= \mathbb{P}[\text{at least one false rejection}] \\
&= \mathbb{P}[H_i \text{ correctly rejected for } i = 1, \ldots j^* - 1 \text{ and } H_{j^*} \text{ falsely rejected}] \\
&\le \mathbb{P}[H_{j^*} \text{ falsely rejected}] \le \alpha,
\end{aligned}$$

and $\mathrm{FWER} = 0$ if $j^* = m + 1$. Here, the decision to reject is given by the rule that the corresponding p-value is smaller or equal to the significance level α. We see explicitly that due to the hierarchical constraint there is no need for multiple testing adjustment.

For the hierarchical procedure with a tree, we can proceed analogously. We consider the *first* (group) hypotheses which are true when moving downwards along the different branches of the tree, and denote them by $H_{j_1^*}, \ldots, H_{j_r^*}$. The familywise error rate can then be bounded as follows: if $r \ge 1$ saying that there is at least one true hypothesis,

$$\begin{aligned} \text{FWER} &= \mathbb{P}[\text{at least one false rejection}] \\ &\leq \mathbb{P}[\cup_{k=1}^{r}\{H_{j_k^*} \text{ falsely rejected}] \leq \sum_{k=1}^{r} \mathbb{P}[H_{j_k^*} \text{ falsely rejected}] \leq \alpha, \end{aligned}$$

where the last inequality holds due to (21) or (22), saying that the adjustment factors for the significance levels for any partition with sets from the hierarchical tree sum up to one. Note that the FWER = 0 if all hypotheses are false. □

The procedure described above and justified in Proposition 4.1 shares a few features to be pointed out. First, it relies on the premise that large groups should be easier to detect and found to be significant, due to the fact that the identifiability is much better posed. We address this issue at the end of this section. In fact, the method has indeed built in the hierarchical constraint (23) that once we cannot reject $H_{0,G}$ for some group G, we do not consider any other sub-groups of G which arise as descendants further down in the tree hierarchy. Due to the sequential nature of the testing procedure, multiple testing adjustment for controlling the familywise error rate is rather mild (for upper parts in the tree) as we only correct for multiplicity at each depth of the tree, i.e., the root node does not need any adjustment, and if it were found to be significant, the next children nodes only need a correction according to the number of nodes at depth 2 of the tree, and so on; see Figure 1.

Further refinements with respect to hierarchical multiple testing adjustment are possible, as described in [42]. But the essential gain in computational and statistical power is in terms of the sequential and hierarchical nature of the procedure as illustrated in Figures 1 and 2. In particular, the method automatically adapts to the resolution level: if the regression parameter of a single variable is very large in absolute value, the procedure might detect such a single variable as being significant; if the signal is not so strong or if there is substantial correlation (or near linear dependence) within a large group of variables, the method might only identify such a large group as being significant; Figure 2 illustrates this point. Naturally, finding a large group to be significant (coarse resolution) is much less informative than detecting a small group or even a single variable.

The power of the hierarchical method is mainly hinging on the assumption that null-hypotheses further up in the tree are easier to reject, that is the p-values are typically getting larger when moving downwards the tree. In low-dimensional regression problems this is typically true when using partial F-tests for testing $H_{0,G}:\ \beta^0_j = 0\ \forall j \in G$. In the high-dimensional case and when using a test-statistics as in (18) for the groups G based on the de-sparsified Lasso, then testing the top node with $G = \{1, \ldots, p\}$ amounts to be equivalent to familywise error rate adjusted testing of single variables: thus, in this case, testing the top node in the tree is not easier than testing all individual components (the bottom nodes in the tree). In the high-dimensional setting, the hierarchical method has been advocated when testing the group null-hypotheses $H_{0,G}$ with a different method than

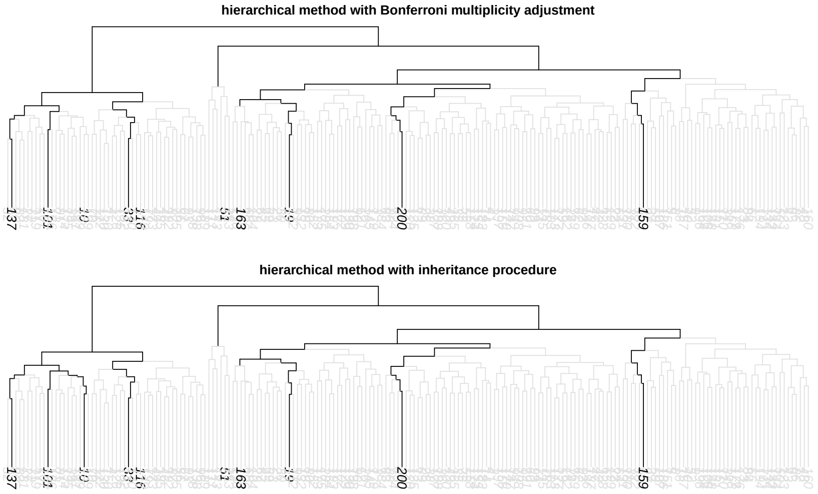

Figure 2. Simulated example with $p = 500$ and $n = 100$. The numbers in black (bold) denote the active variables with $\beta^0_j \neq 0$ (and corresponding $H_{0,j}$ being false). Top panel: hierarchical procedure with weighted Bonferroni adjustment as described in (21). Bottom panel: A refined procedure (including so-called inheritance) which detects in addition the single variable 10. For details see [42], and the figure is also taken from [42].

the de-sparsified Lasso and the corresponding maximum test-statistics, namely a multi-sample splitting technique introduced in [48] and extended for hierarchical testing in [41, 42]. The drawback of such a method is that the currently available theoretical justification requires that essentially all the non-zero regression coefficients are sufficiently large (whereas the de-sparsified Lasso does not require such a gap condition), see also Section 3.3.

5. Genome-wide association studies

A major problem in genetics is the association between genetic markers and the status of a disease. As of 2011, 1'200 human Genome-Wide Association Studies (GWAS) over 200 diseases have been completed (Wikipedia on "Genome-wide association study"), and GWAS are among the most important approaches for further understanding of genetic influences to disease status. In each study, the sample size is usually around $1'000 - 5'000$ whereas the number of genetic markers in terms of single nucleotide polymorphisms (SNPs) is in the order of 10^6.

We consider a linear logistic regression model. The disease status is denoted by $Y \in \{0, 1\}$ (where $Y = 1$ denotes "diseased", and $Y = 0$ "healthy") and the genetic SNP j by $X^{(j)}$ ($j = 1, \ldots, p \approx 10^6$) with $X^{(j)} \in \{0, 1, 2\}$ corresponding to the number of minor alleles at genomic position j. The linear logistic regression model is as follows:

$$
\begin{aligned}
&Y_1, \ldots Y_n \text{ independent,}\\
&\mathbb{P}[Y_i = 1|X_i] = \pi(X_i),\ \log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = \sum_{j=1}^{p} \beta_j^0 X_i^{(j)}.
\end{aligned}
\qquad (24)
$$

Thereby, the discrete structure of the variables $X_i^{(j)} \in \{0, 1, 2\}$ is not given some special attention (and instead is modeled with a scale as if the variables were continuous; alternatively, one could treat these variables as ordinal). We see that the logistic transform $\log(\pi(X_i)/(1-\pi(X_i)))$ is linear in the parameter β^0, analogously to a linear model in (2) which we would use if the "disease status" or so-called phenotype would be measured with a continuous variable, as in the application outlined just before Section 5.1. Of main interest are significance statements for testing null-hypotheses $H_{0,G}$ as in (11) or $H_{0,j}$ as in (10). That is, if a group G is significant, it means that there is at least one SNP j in the group G whose corresponding regression coefficient β_j^0 would be significant: we would then call SNP j a significant and even to a certain extent causal (see Section 5.1) SNP for the disease status Y.

We summarize here some results from [11] where some data from the Wellcome Trust Case Control Consortium (WTCCC) is analyzed [51], see also `https://www.wtccc.org.uk/`. Seven major diseases are studied, for each of them measuring a binary disease status and genome-wide SNPs. After missing data handling, the number of controls (having healthy disease status $Y = 0$) is 2934 whereas the number of cases (with disease status $Y = 1$) is about 1'700-1'800, depending on the disease; the number of SNPs is approximately 380'000. We use the hierarchical inference procedure described in Section 4 to infer significance statements for testing $H_{0,G}$ in (11) in the logistic model (24).

The tree hierarchy is constructed as follows. The top node comprises all the $\approx$ 380'000 SNPs, the next level at depth 2 is partitioned into groups of SNPs corresponding to the 22 chromosomes. Then, further groups or nodes down in the tree hierarchy are constructed from hierarchical clustering with average linkage and based on the dissimilarity measure between two SNP variables $X^{(j)}$ and $X^{(k)}$ as $1 - \hat{\rho}_{jk}^2$ with $\hat{\rho}_{jk}$ equal to the empirical correlation between $X^{(j)}$ and $X^{(k)}$. Although hierarchical clustering is computationally rather demanding, one can trivially distribute the computation according to the 22 chromosomes (i.e., 22 groups which are pre-specified from the biological context).

For the statistical inference and corresponding p-values of groups we use the hierarchical method described in Section 4 with a depth-wise weighted Bonferroni adjustment described in (21); the statistical tests are based on a multi-sample splitting approach and likelihood ratio testing (in analogy to the partial F-test in linear models). The details are described in [11].

We structure the obtained results according to "small" and large detected groups of SNPs, as illustrated in Tables 1 and 2. In the former case, for some diseases, we are able to detect some single SNPs. This is a spectacular finding:

to appreciate, there are 16 single SNP markers which have a significant effect on Y *even when adjusting for the effects of all* $\approx 380'000$ *(many thousands!) other SNPs*. We mainly see these findings for diseases which are "known" to be more genetically driven. In contrast, there are diseases where we cannot identify small groups pf SNPs, namely for Bipolar disease and hypertension (out of the 7 diseases considered here). The results for large groups of significant SNPs for Bipolar disease are summarized in Table 2. We cannot identify a smaller region in the genome which would exhibit a significant association: however, we are able to identify chromosomes which show a significant effect.

disease	group of SNPs	chromosome	p-value
CAD	rs1333049	9	$1.7 \cdot 10^{-3}$
CD	rs10210302	2	$4.6 \cdot 10^{-5}$
CD	rs2076756	16	$1.3 \cdot 10^{-3}$
CD	rs10761659	10	$1.5 \cdot 10^{-2}$
CD	rs10883371	10	$2.4 \cdot 10^{-2}$
CD	rs2542151	18	$1.5 \cdot 10^{-2}$
CD	cardinality = 5	1	$4.5 \cdot 10^{-2}$
CD	cardinality = 9	5	$2.7 \cdot 10^{-3}$
RA	rs6679677	1	$5.9 \cdot 10^{-11}$
RA	rs9272346	6	$1.4 \cdot 10^{-6}$
T1D	rs6679677	1	$3.6 \cdot 10^{-11}$
T1D	rs2523691	6	$6.0 \cdot 10^{-5}$
T1D	rs9272723	6	$2.2 \cdot 10^{-4}$
T1D	rs17696736	12	$6.5 \cdot 10^{-4}$
T1D	rs9272346	6	$2.4 \cdot 10^{-3}$
T1D	rs11171739	12	$1.3 \cdot 10^{-2}$
T1D	rs17388568	4	$2.7 \cdot 10^{-2}$
T1D	rs12924729	16	$3.4 \cdot 10^{-2}$
T2D	cardinality = 4	16	$4.7 \cdot 10^{-2}$
T2D	cardinality = 10	10	$1.7 \cdot 10^{-5}$

Table 1. Significance of small groups of SNPs. Left column: CAD = coronary artery disease, CD = Crohn's disease; RA = rheumatoid arthritis, T1D = type 1 diabetes, T2D = type 2 diabetes. Second column: single SNPs are given in terms of an ID, larger groups of SNPs in terms of the cardinality of the group. Right column: P-values are ordered within a disease. Results are taken from [11].

The validation of the findings, part of them being summarized in Tables 1 and 2 is not straightforward. One can consider "consensus" with other results in the literature and with follow up studies which have been performed by the WTCCC consortium: we refer for more details to [11]. From a statistical modeling perspective, however, we have a formally valid way to interpret the results as discussed in Section 5.1.

cardinality of group	chromosome	p-value
6695	1	0.027
12134	1	0.047
14451	2	0.016
7338	2	0.036
1649	3	0.021
24832	4	0.008
14040	5	0.030
24193	6	0.041
20643	7	0.013
21594	8	0.027
11929	9	0.009
22517	10	0.021
15269	12	0.038
4389	14	0.048
11055	15	0.032
10382	16	0.047

Table 2. Significance of large groups of SNPs for Bipolar disease. Ordered by chromosome number and p-values within chromosomes. Results are taken from [11].

Continuous response: GWAS for Arabidopsis Thaliana. The hierarchical inference method described in Section 4 has also been applied to GWAS for the model plant Arabidopsis Thaliana. In contrast to humans, one can perform validation experiments. We summarize here the results from [37]. The response variable (or phenotype) Y is measuring the root meristem zone-length (root size) of the plant, the covariates correspond to $p = 241'051$ SNPs and the sample size is $n = 201$. We consider a linear model as in (2) and perform hierarchical inference as described in Section 4. The hierarchy is inferred from hierarchical clustering of the SNPs with dissimilarity given by $1 - (\text{empirical correlation})^2$. Four new significant small groups of SNPs are found, besides nearly all "previously known" associations: these 4 groups are within and neighboring to the so-called PEPR2 gene. A validation experiment is then performed by growing wild-type (non-manipulated) plants and mutant plants with "loss-of-function" of the PEPR2 gene. This experiment exhibited a significant difference with respect to root meristem (root size), providing evidence that the method has indeed identified a relevant group of SNPs. The details are given in [37].

5.1. A causal interpretation. Causal inference deals with "directional associations": instead of just observing that two random variables X and Y are dependent, causal inference would allow to quantify to what extent say X is a cause for Y. Causal inference is intrinsically related to analyzing the effect of an intervention, and it typically tries to predict such an intervention effect from observational data, without performing a randomized intervention experiment. We refer the reader for an extensive treatment of the topic to the book by Pearl [49].

The intention here is to only make a brief connection to structural equation and graphical models which will then allow for some causal interpretation when keeping some caution in mind that some additional assumptions are needed. As in our setting discussed in the previous sections, we consider $p+1$ random variables $X^{(1)}, \ldots, X^{(p)}, X^{(p+1)} = Y$ (and the data is interpreted as being n i.i.d. realizations thereof). We represent the random variables as nodes in a graph, and we assume a directed acyclic graph (DAG) D which encodes the causal influence diagram among these $p+1$ random variables. We denote by pa_j the parents of node j in the DAG D. A corresponding structural equation model is specified as follows:

$$\begin{aligned}&\text{the conditional distributions } \mathcal{L}(X^{(j)}|X^{(\mathrm{pa}_j)})\ (j = 1, \ldots, p+1)\\ &\qquad\qquad\qquad\qquad\text{are of a certain form or model class,}\\ &X^{(j)}|X^{(\mathrm{pa}_j)}\ (j = 1, \ldots, p+1) \text{ are jointly conditionally independent.}\end{aligned} \tag{25}$$

The joint distribution then obeys the Markov property with respect to the DAG D [49, cf.]. The following result holds.

Proposition 5.1. *Assume that the variables $X^{(1)}, \ldots, X^{(p)}, X^{(p+1)} = Y$ satisfy a structural equation model in* (25) *with underlying DAG D. We also assume that the structural equation of Y is according to a linear or logistic regression model:*

$$\textit{linear:} \qquad Y = \sum_{j \in \mathrm{pa}_Y} B_{Y,j} X^{(j)} + \varepsilon^{(Y)},\ \varepsilon^{(Y)} \textit{ independent of } X^{(\mathrm{pa}_Y)},$$

$$\textit{logistic:} \qquad \mathbb{P}[Y = 1|X] = \pi(X),\ \log(\pi(X)/(1-\pi(X))) = \sum_{j \in \mathrm{pa}_Y} B_{Y,j} X^{(j)}.$$

Consider the linear or logistic regression coefficients β^0 in the regression of Y versus all $X^{(1)}, \ldots, X^{(p)}$ and assume that it is identifiable from the distribution. Then, if $\beta^0_j \neq 0$, it holds that $X^{(j)} \in \mathrm{pa}_Y$ and there is a directed edge $X^{(j)} \to Y$ (i.e., a direct causal effect from $X^{(j)}$ to Y).

Proof. The slightly non-trivial part is to show that when running a regression of Y versus all $X^{(1)}, \ldots, X^{(p)}$, we actually obtain the coefficients $B_{Y,j}$ from the structural equation for Y, i.e., $\beta^0_j = B_{Y,j}$. The argument is as follows. The DAG D induces an ordering among the variables such that $\mathrm{pa}_j \subseteq \{j-1, \ldots, 1\}$, assuming for notational simplicity that the variables have already been ordered (according to such an order). Since Y is childless we can choose an ordering where Y is the last element. The conditional distribution then satisfies thanks to the Markov property:

$$\mathcal{L}(Y|X^{(p)}, \ldots, X^{(1)}) = \mathcal{L}(Y|X^{(\mathrm{pa}_Y)}).$$

This completes the proof. □

Causal interpretation. As a consequence, under the assumptions in Proposition 5.1, the inference techniques for regression lead to a causal interpretation. The main assumptions for such a substantially more sharpened interpretation are: (i) the underlying true model is a structural equation model with a DAG structure and a linear or logistic form for the structural equation of Y; (ii) there are no hidden confounder variables between Y and some of the $X^{(j)}$'s; (iii) the response variable Y is childless. The last assumption (iii) is rather plausible for GWAS since one believes that the genetic factors are the causes for the disease and ruling out that the disease would cause a certain constellation of genetic factors. The second assumption (ii) is rather strong and perhaps the main additional assumption: however, in view of measuring thousands of genetic markers, the premise of having measured all the relevant factors is somewhat less unrealistic; the first assumption (i) about the acyclicity of the causal influence diagram is not important as long as there is no feedback from the response Y to the X variables (which is plausible for GWAS), while the requirement for a linear or logistic form might be problematic in view of possible interactions among the X-variables and/or nonlinear regression functions. The latter is a misspecification and of the same nature as when having misspecified the functional form in a regression model, a topic which we will discuss in Section 6.

One should always be careful when adopting a causal interpretation. However, and this is a main point, the regression model taking all the variables into account is much more appropriate than a marginal approach where the response Y is marginally regressed or correlated to one SNP variable at a time. This has been the standard approach over many years in GWAS, including extensions with mixed models and adjusting for a few other covariates [62]. The approach based on modern high-dimensional statistics presented in Sections 2, 3 and 4 comes much closer to a causal interpretation as described in Proposition 5.1. And that is among the main reasons why we believe that such kind of approaches should lead to more reliable results for GWAS in comparison to the older marginal techniques.

6. Misspecification of the linear model

The results in the previous sections for statistical confidence or testing of linear model parameters rely on the correctness of a linear model as in (2). If the model is not correct, we have to distinguish more carefully between random and fixed design matrix X (and the latter case may also arise when conditioning on X). A detailed treatment is given in [9].

6.1. Random design. Consider a nonlinear model with random design:

$$Y = f^0(X) + \varepsilon, \tag{26}$$

where ε is independent of the $p \times 1$ random vector X with $\mathbb{E}[\varepsilon] = 0$. For simplicity, we also assume that Y and X are centered, i.e., $\mathbb{E}[f^0(X)] = \mathbb{E}[X^{(j)}] = 0$ for all j.

The observed data is assumed as n i.i.d. realizations of (Y, X). The best projected parameter in a linear model w.r.t. L_2-norm is

$$\beta^* = \text{argmin}_\beta \mathbb{E}[|f^0(X) - X^T\beta|^2],$$

and it is unique if $\Sigma = \mathbb{E}[X^TX] = \text{Cov}(X)$ is positive definite. Thus, when fitting a (high-dimensional) linear model to data being i.i.d. realizations from the nonlinear model (26), we will estimate the projected parameter β^*. The inference machinery for obtaining p-values and confidence intervals is also valid under the following modification and assumption.

First, due to random design, the variance of say the de-sparsified estimator in (13) is

$$v_j^2 = \text{Var}((Z_j^T X^{(j)}/n)\sqrt{n}\hat{b}_j) \asymp \text{Var}(n^{-1/2} Z_j^T \varepsilon) = n^{-1} \sum_{i=1}^n \text{Var}(Z_{j;i}\varepsilon_i),$$

where $Z_{j;i}$ denotes the ith component of Z_j. The quantity above can be estimated by the empirical analogue

$$\hat{v}_j^2 = n^{-1} \sum_{i=1}^n (Z_{j;i}\hat{\varepsilon}_i - \overline{Z_j\hat{\varepsilon}})^2, \tag{27}$$

where $\overline{u} = n^{-1}\sum_{i=1}^n u_i$ denotes the arithmetic mean of the entries of an $n \times 1$ vector u. One can then show under similar conditions as for (15) that

$$\frac{(Z_j^T X^{(j)}/n)\sqrt{n}(\hat{b}_j - \beta_j^0)}{\hat{v}_j} \Longrightarrow \mathcal{N}(0, 1).$$

The difference is that for random design under model-misspecification, we have to use a different estimator for the variance (rather than $\hat{\sigma}^2\|Z_j\|_2^2/n$ in the fixed design case as appearing in e.g. (16)), namely the one in (27) which is equivalent to the "sandwich estimator" from White [59].

The other issue concerns the assumption of sparsity for β^*. In general, β^* might be much less sparse or even dense even if β^0 is sparse. When assuming a block-structure for the covariance matrix Σ, some bounds for the sparsity of β^* can be obtained, see [9]; another bound on the sparsity is implied by Proposition 6.1 below. Thus, the de-sparsified Lasso with the variance formula in (27) can be justified.

Finally, the question is whether there is any relation between β^* and β^0. Denote by $S(f^0)$ the support of f^0, i.e., the indices of the variables $X^{(1)}, \ldots, X^{(p)}$ which are having an influence in f^0; and by $S^* = \text{supp}(\beta^*) = \{j;\ \beta_j^8 \neq 0\}$. The following result is given in [9].

Proposition 6.1. *Assume that* $X \sim \mathcal{N}_p(0, \Sigma)$. *Then,*

$$S^* \subseteq S(f^0).$$

The proposition says that if e.g. the de-sparsified Lasso finds a significant variable in the misspecified linear model with Gaussian design, then it must be a true active variable in $S(f^0)$, and the error control against false positive statements is controlled with the de-sparsified Lasso procedure.

Conditioning on $\boldsymbol{X}$. It is worthwhile to consider the argument of conditioning. If the model would be correctly specified with $f^0(X) = X^T\beta^0$, we would have

$$Y = X^T\beta^0 + \varepsilon.$$

When conditioning on X, since $\mathbb{E}[\varepsilon|X] = \mathbb{E}[\varepsilon] = 0$, we would have a fixed design linear model with mean zero error term. And having a procedure which constructs confidence intervals for fixed design then also works for random design (since confidence statements hold for every fixed realization of X). On the other hand, if the model is misspecified with $f^0(X) \neq X^T\beta^*$ we get the representation

$$Y = X^T\beta^* + \underbrace{(f^0(X) - X^T\beta^* + \varepsilon)}_{=:\eta}.$$

We have for the error term that $\mathbb{E}[\eta] = 0$ but when conditioning on X we obtain that $\mathbb{E}[\eta|X] = f^0(X) - X^T\beta^* \neq 0$. Thus, this explains that for the misspecified random design case, the inference for the projected parameter β^* should be done and interpreted as unconditional while the conditional inference for β^* is not valid.

Another viewpoint when conditioning on X or when having a fixed design case will be treated next.

6.2. Fixed design. Consider a nonlinear model as in (26) but with fixed design corresponding to n observations. We can then always represent the $n \times 1$ vector of the nonlinear function at the data points as

$$\mathbf{f}^0 = (f^0(X_1), \ldots, f^0(X_n))^T = X\beta^0,$$

for many solutions β^0, and implicitly assuming that $\mathrm{rank}(X) = n$. For example, compressed sensing is solving the convex optimization problem

$$\beta^0_{\mathrm{comprsens}} = \mathrm{argmin}_\beta \|\beta\|_1 \text{ such that } X\beta = \mathbf{f}^0.$$

Under an restricted isometry property for the design X it is known that $\beta^0_{\mathrm{comprsens}}$ equals the ℓ_0-sparsest representation [13, cf.] (and such results can also be obtained assuming the weaker compatibility condition as in (A2)). Thus, high-dimensionality is a "kind of a blessing" since we can represent any nonlinear additive error model as

$$Y = X\beta^0 + \varepsilon,$$

and model-misspecification with respect to nonlinearity is not an issue. The only question is whether there are solutions β^0 which are sufficiently sparse, fulfilling our required ℓ_0-sparsity assumption in (A1); we do not discuss here the issue of using other bases (other design matrices X) and dictionary learning. Regarding the latter ℓ_0-sparsity, the theory for the de-sparsified Lasso can actually be extended to require weak sparsity only w.r.t $\|\beta^0\|_q$ with $0 \le q < 1$ which allows for a greater generality of sparsity and somewhat weakens the ℓ_0-assumption in (A1) [54].

The inference machinery from Section 3 is to be interpreted as follows: a confidence interval for the jth variable covers with high probability β_j^0 for all sufficiently sparse solutions β^0; but we have to "bet" that there is at least one representation with a sufficiently sparse β^0. The latter assumption or "bet" can actually be investigated with a statistical goodness of fit test for deciding whether a linear model representation with a sparse β^0 fits the data adequately [50].

7. Software

Open source software for high-dimensional statistics is available as `R`- and `Bioconductor`-packages.

Fitting ℓ_1-regularized generalized regression models is efficiently implemented in the `glmnet` R-package [26].

Statistical significance testing and confidence intervals are implemented in the `hdi` R-package [43], and some corresponding background and comparative overview is given in [19]. Hierarchical inference, especially in the context of GWAS is available from the `hierGWAS` `Bioconductor`-package [10].

8. Conclusions

High-dimensional statistics deals with estimation and quantifying uncertainty in models where the dimension of the unknown parameter is much larger than the sample size. A key assumption is sparsity, for mathematically deriving near optimality in various forms as well as for accuracy in practice.

In this article, we give a review of some of the important concepts and results and illustrate their potential for applications to genome-wide association studies, one of the most active research fields in genetics. For simplicity, our exposition of theory and most of the methodology is given for linear models as in (2): extensions to case of generalized linear or nonparametric models are treated in e.g. [8, 54]. The recently developed theory and machinery for quantifying uncertainty in terms of confidence intervals and hierarchical multiple statistical hypotheses testing (Sections 3 and 4) is opening the door for many applications where assigning of statistical uncertainty has a long tradition to quantify replicability and scientific relevance of findings, notably in medical research. Our application to genome-wide association studies (Section 5) illustrates that the new techniques offer something which has not been possible before: namely to obtain statistical p-values of regression coefficients or groups in a multiple linear model with $O(10^6)$ variables, and thus enabling a much more causal-oriented interpretation. In fact, as illustrated here, we obtain interesting results for studies from the Wellcome Trust Case Control Consortium.

We have not considered here the setting of large-scale or "Big" data [24, cf.]. Recent developments in this field include scalability of algorithms and computational–statistical trade-offs [14, 17], or addressing the issue of heterogeneity [46, 6, 31]. Some of the mathematical techniques and methodology from high-dimensional statistics remain as important key elements in this new field.

Acknowledgements. The author would like to thank Marta Sanz-Solé for some useful comments.

References

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

[2] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, 1998.

[3] P. Bühlmann. Invited discussion on "Regression shrinkage and selection via the Lasso: a retrospective (R. Tibshirani)". *Journal of the Royal Statistical Society, Series B*, 73:277–279, 2011.

[4] P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242, 2013.

[5] P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications*, 1:255–278, 2014.

[6] P. Bühlmann and N. Meinshausen. Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104:126–135, 2016.

[7] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1858, 2013.

[8] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

[9] P. Bühlmann and S. van de Geer. High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9:1449–1473, 2015.

[10] L. Buzdugan. *hierGWAS: Asessing statistical significance in predictive GWA studies*, 2016. R package version 1.3.0.

[11] L. Buzdugan, M. Kalisch, A. Navarro, D. Schunk, E. Fehr, and P. Bühlmann. Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics*, 32:1990–2000, 2016.

[12] E.J. Candès, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2006.

[13] E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52:5406–5425, 2006.

[14] V. Chandrasekaran and M.I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110:E1181–E1190, 2013.

[15] S.J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D.J. Hunter, G. Thomas, J.N. Hirschhorn, G. Abecasis, D. Altshuler, J.E. Bailey-Wilson, et al. Replicating genotype–phenotype associations. *Nature*, 447:655–660, 2007.

[16] A. Chatterjee and S.N. Lahiri. Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics*, 41:1232–1259, 2013.

[17] Y. Chen and M.J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees, 2015. Preprint arXiv:1509.03025.

[18] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41:2786–2819, 2013.

[19] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical Science*, 30:533–558, 2015.

[20] R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, to appear (with discussion), 2016. Preprint arXiv:1606.03940.

[21] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

[22] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001.

[23] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[24] J. Fan, F. Han, and H. Liu. Challenges of big data analysis. *National Science Review*, 1:293–314, 2014.

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

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

[27] U. Gamper, P. Boesiger, and S. Kozerke. Compressed sensing in dynamic MRI. *Magnetic Resonance in Medicine*, 59:365–373, 2008.

[28] E. Giné and J. Zinn. Necessary conditions for the bootstrap of the mean. *Annals of Statistics*, 17:684–691, 06 1989.

[29] E. Giné and J. Zinn. Bootstrapping general empirical measures. *Annals of Probability*, 18:851–869, 04 1990.

[30] C. Giraud. *Introduction to High-Dimensional Statistics.* CRC Press, 2014.

[31] S.M. Gross and R. Tibshirani. Data shared Lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235, 2016.

[32] J. Hartigan. *Clustering Algorithms.* Wiley, 1975.

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

[34] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations.* CRC Press, 2015.

[35] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[36] H. Jung, K. Sung, K.S. Nayak, E.Y. Kim, and J.C. Ye. k-t focuss: A general compressed sensing framework for high resolution dynamic MRI. *Magnetic Resonance in Medicine*, 61:103–116, 2009.

[37] J.R. Klasen, E. Barbez, L. Meier, N. Meinshausen, P. Bühlmann, M. Koornneef, W. Busch, and K. Schneeberger. A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nature Communications*, 7:Article number 13299 (doi:10.1038/ncomms13299), 2016.

[38] K. Knight and W. Fu. Asymptotics of Lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

[39] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

[40] M. Lustig, D.L. Donoho, and J.M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58:1182–1195, 2007.

[41] J. Mandozzi and P. Bühlmann. Hierarchical testing in the high-dimensional setting with correlated variables. *Journal of the American Statistical Association*, 111:331–343, 2016.

[42] J. Mandozzi and P. Bühlmann. A sequential rejection testing method for high-dimensional regression with correlated variables. *International Journal of Biostatistics*, 12:79–95, 2016.

[43] L. Meier, N. Meinshausen, and R. Dezeure. *hdi: High-Dimensional Inference*, 2014. R package version 0.1-2.

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

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

[46] N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *Annals of Statistics*, 43:1801–1830, 2015.

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

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

[49] J. Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, 2000.

[50] R.D. Shah and P. Bühlmann. Goodness of fit tests for high-dimensional models, 2015. Preprint arXiv:1511.03334.

[51] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

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

[53] S. van de Geer. The deterministic Lasso. In *JSM proceedings, 2007, 140.* American Statistical Association, 2007.

[54] S. van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités des Saint-Flour XLV – 2015.* Lecture Notes in Mathematics 2159. Springer, 2016.

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

[56] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

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

[58] P.H. Westfall and S.S. Young. *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment.* John Wiley & Sons, 1993.

[59] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

[60] C.-H. Zhang and S.S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, 76:217–242, 2014.

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

[62] X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11:407–409, 2014.

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

Received submission date; revised revision date