MAXIMIN EFFECTS IN INHOMOGENEOUS LARGE-SCALE DATA

By Nicolai Meinshausen and Peter Bühlmann

ETH Zürich

Large-scale data are often characterized by some degree of inhomogeneity as data are either recorded in different time regimes or taken from multiple sources. We look at regression models and the effect of randomly changing coefficients, where the change is either smoothly in time or some other dimension or even without any such structure. Fitting varying-coefficient models or mixture models can be appropriate solutions but are computationally very demanding and often return more information than necessary. If we just ask for a model estimator that shows good predictive properties for all regimes of the data, then we are aiming for a simple linear model that is reliable for all possible subsets of the data. We propose the concept of "maximin effects" and a suitable estimator and look at its prediction accuracy from a theoretical point of view in a mixture model with known or unknown group structure. Under certain circumstances the estimator can be computed orders of magnitudes faster than standard penalized regression estimators, making computations on large-scale data feasible. Empirical examples complement the novel methodology and theory.

1. Introduction. "Big data" often refers to datasets that are large in different ways: there can be many observations, many variables or both, and the size can be measured against some historical standard or against available computational resources (e.g., the data might be too large to fit into memory). Data can also come from different sources, have inhomogeneities and might have to be processed in a streaming fashion. Here, we want to take a look at one specific aspect of "big data," the effect of inhomogeneities in the data in regression modeling. Specifically the question whether one is able to extract (in a computationally feasible way) a model that works for data that come from different time-regimes or that, more generally, have different underlying distributions.

From a perhaps slightly naive statistical point of view, a situation where we face computational challenges due to a large number of homogeneous observations in a database is not problematic. We can simply discard most of the observations and retain sufficiently many observations, chosen at random, to guarantee good predictive accuracy. The exact number of observations we have to retain will be a function of the desired predictive accuracy, the number of variables and the noise level. Keeping tens of thousands of observations will be sufficient for most practical purposes. Most estimators can easily deal with datasets of this size.

However, many large-scale datasets do not fit neatly into the standard framework of a single underlying model observed with independent and identically distributed errors. There are likely to be outliers in the data, and the truth might better be approximated with a mixture of models than a single one and underlying distributions of the variables might shift over time [Hand (2006)]. There has been a lot of work on various aspects of these issues. While we cannot provide an even approximately complete overview, some of the major themes can be found in work on robust estimation [Hampel et al. (1986), Huber (1964, 1973)], time-varying coefficients models [Cai, Fan and Li (2000), Fan and Zhang (1999), Hastie and Tibshirani (1993)], mixture models [Aitkin and Rubin (1985), Figueiredo and Jain (2002), McLachlan and Peel (2000)] and change-point estimation [Carlstein, Müller and Siegmund (1994)]. In a high-dimensional regression dataset, Städler, Bühlmann and van de Geer (2010) showed evidence for the presence of multiple components that can be exploited for variable selection. Mixed- and random-effects models [McCulloch (2006), Pinheiro and Bates (2000)] are related but do not have an observation-specific random effect. Varying-coefficient models seem particularly attractive to capture shifts in underlying distributions if the data are recorded chronologically, and the approach has been extended to cope with more general estimation problems, including estimation of time-varying graphs [Kolar et al. (2010)]. Mixture models, on the other hand, do not assume such a structure and try to infer the hidden states of the mixture class membership by using, for example, the EM-algorithm [Dempster, Laird and Rubin (1977)] or related approaches.

In some applications, trying to infer the full time-varying coefficients in a model or inferring the hidden states in a mixture model can be computationally challenging, and success is not always guaranteed from a statistical point of view. Moreover, we might not be interested in the hidden states or the exact time-evolution of the coefficients but rather in a simple model that can work reliably for all states or times. Our examples will mostly fit into a change-point model, where the underlying distribution can shift abruptly.

An example is given in Figure 1, which is based on price data of twelve financial future instruments (including foreign exchange, equity and commodity futures) on time-resolution of minutes over the course of ten years, after 2000. We use the past 5 minutes of log-returns of all instruments (i.e., 60 predictor variables) to forecast with a linear model the log-return of the Euro–Dollar exchange rate over the next minute (which is the response variable). Two-thirds of the data are used for training a least squares estimator, and the cumulative cross products of this model for the training and test data is shown in the first and second panel of Figure 1, respectively, where the cumulative gain up to time $1 \le t \le n$ is for response values $Y_i, i = 1, \ldots, n$ and predictions $\hat{Y}_i, i = 1, \ldots, n$ (both are assumed to mean-centered and predictions are normalized to have a second moment of 1)
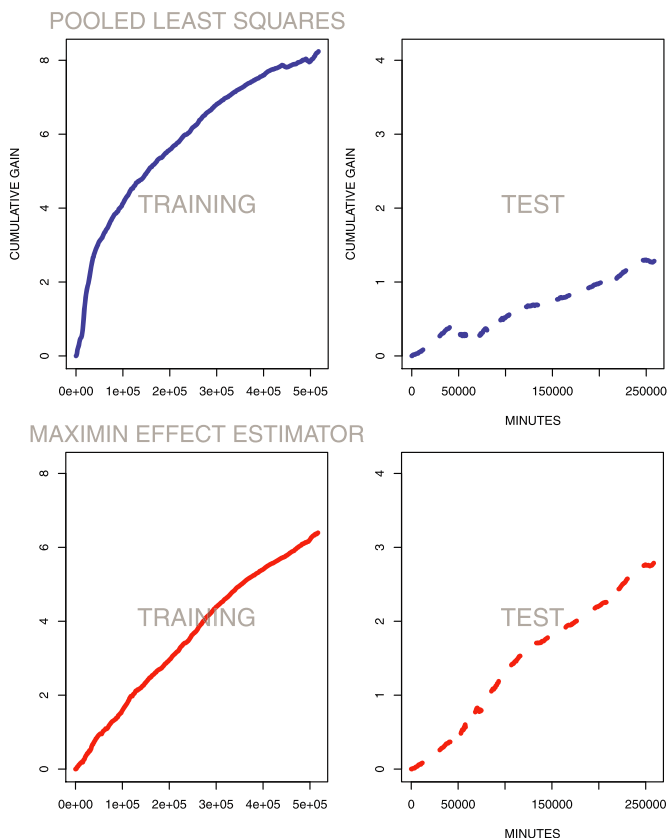
FIG. 1. *Forecasting minute log-returns of the Euro–Dollar exchange rate with a least squares linear model fit over the pooled training data and the proposed maximin effects estimator. The panels show the cumulative cross products as in* (1) *for the least squares estimator and the maximin effect estimator, respectively, where the aspect ratios are chosen such that the same effect strength will lead to the same slope in all four panels. There are more than half a million training observations for a model with just* 60 *free parameters, and yet the least squares estimator overfits, which leads to a degradation in performance on the test data. The performance of the maximin estimator is more consistent over the training data, which translates into a better performance on the test data.*

given by

$$\sum_{i=1}^{t}(Y_i \hat{Y}_i).\tag{1}$$

The training data show that the model works very well at the beginning of the train-ing period but then tails off, and performance on the test data is much worse than on the training period, even though there are more than half a million observations to fit a model with 60 parameters. In contrast, the third and fourth panel show the cumulative cross products of the "maximin effects" estimator that we propose. The

least squares estimator is maximizing the explained variance on the pooled dataset, leading (as in this example) to periods or groups of observations where the fit is very good, and others where barely any variance is explained by the fitted values. In contrast, the "maximin effects" estimator maximizes the explained variance for the worst group of observations, which have been divided in this example rather arbitrarily into 3 equally large blocks of consecutive observations; see Section 2 for more precise definitions regarding a group of observations. The estimator in the example is computed without a regularization penalty. The predictive accuracy is much more constant over time, and performance on the test data is in line with performance on the training data, as the estimator has not been as much influenced by the period at the beginning of the training set as the least squares estimator.

We will set notation and introduce the maximin effects estimator in Section 2, while showing some properties for known and unknown group structure in Section 3, discussing computational properties in Section 4 and concluding with an example in Section 5.

2. Maximin effects. We will first try to give a suitable and intuitive definition of maximin effects in mixture models or varying-coefficient models, while introducing the maximin effects estimator thereafter.

While we focus exclusively on regression here for ease of exposition, the same approach can be used, for example, for classification and graph estimation.

2.1. *Maximin effects for mixture models.* We will work with a mixture model, where for n observations of a real-valued response Y_i and a $1 \times p$ predictor variable $X_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$,

$$(2) \qquad\qquad Y_i = X_i B_i + \varepsilon_i \qquad \text{where } B_i \in \mathbb{R}^p \text{ and } B_i \sim F_B$$

for some unknown distribution F_B, either discrete or continuous. We also use the standard notation with the $n \times 1$ response vector Y, the $n \times p$ design matrix X and the $n \times 1$ error vector ε. The predictor variables X_i are random and independent, and the noise $\varepsilon_1, \ldots, \varepsilon_n$ fulfils $E(\varepsilon^t X) = 0$. Furthermore, the coefficients B_i are independent from the X_i, $i = 1, \ldots, n$. Independent noise is an example, but some dependencies between noise contributions are also possible in this framework, for example, if the observations have a time-ordering. The inhomogeneity of the data is thus solely caused by the variation of the regression coefficients among the sample points with indices $i = 1, \ldots, n$. We do not necessarily assume that the B_i, $i = 1, \ldots, n$ are independent. They can be organized in known or unknown groups. The following examples indicate the scope of the model: if F_B has point masses at a finite set of points, we are in the setting of classical finite mixture models, where B can take one of a finite number of values. In another scenario, realizations B_i are positively correlated over time if the observations are ordered in some chronological order, creating a smoothly varying effect over time. In the latter example, the model behaves more like a varying-coefficient model [Hastie

and Tibshirani (1993)]. A shift in the distribution of the predictor variables could conceivably be handled in a very similar manner. As a final example, the realizations B_i are most often the same, but a small fraction takes other values which can be viewed as outliers or contaminations.

We always assume that the random X_i are identically distributed from a distribution with population Gram matrix Σ. For a fixed regression coefficient $b \in \text{support}(F_B) \subseteq \mathbb{R}^p$, we can define two different optimality criteria: $R_{\beta;b}$ is the variance of the residuals in absence of additional errors on the observations, while $V_{\beta;b}$ is the explained variance of predictions with $\beta \in \mathbb{R}^p$:

$$(3) \qquad R_{\beta;b} = E\big(\|Xb - X\beta\|_2^2/n\big) = b^t \Sigma b - 2\beta^t \Sigma b + \beta^t \Sigma \beta,$$

$$(4) \qquad V_{\beta;b} = 2\beta^t \Sigma b - \beta^t \Sigma \beta.$$

Alternative expressions for (3) and (4) under the condition $E(\varepsilon^t X) = 0$ are

$$R_{\beta;b} = E_{Y,X}\big(\|Y - X\beta\|_2^2/n\big) - E\big(\|\varepsilon\|_2^2/n\big),$$

$$V_{\beta;b} = E\big(\|Y\|_2^2/n\big) - E_{Y,X}\big(\|Y - X\beta\|_2^2/n\big).$$

If we want to find a single p-dimensional regression coefficient that works optimally on average over $B \sim F_B$, the optimal choice are the pooled coefficients

$$(5) \qquad b_{\text{pool}} = \underset{\beta}{\arg\min}\, E_B(-V_{\beta;B}) = \underset{\beta}{\arg\min}\, E_B(R_{\beta;B}),$$

where the expectation is with respect to $B \sim F_B$. Note that in this case it is inconsequential for the pooled estimator, whether we minimize the residuals or maximize the explained variance.

If we are looking for effects that guarantee a good performance throughout all possible parameter values, in analogy to decision theoretic consideration [Wald (1945)], two possible definitions of effects are

$$(6) \qquad b_{\text{pred-maximin}} = \underset{\beta}{\arg\min}\, \underset{b \in F}{\max}\, R_{\beta;b},$$

$$b_{\text{maximin}} = \underset{\beta}{\arg\min}\, \underset{b \in F}{\max}\,(R_{\beta;b} - R_{0;b})$$

$$(7)$$

$$= \underset{\beta}{\arg\min}\, \underset{b \in F}{\max}\,(-V_{\beta;b}) = \underset{\beta}{\arg\max}\, \underset{b \in F}{\min}\,(V_{\beta;b}),$$

where $F = \text{support}(F_B)$. Alternatively, F could be the smallest region such that $P(B \in F) \geq 1 - \alpha$ for some $\alpha \in (0, 1]$, guaranteeing success for a large fraction of randomly chosen coefficient values.

Two comments are in order regarding the definition of maximin effects. First, the effects are optimizing for the worst-case scenario for $b \in F$. To be more precise, if we view future samples of B to be allowed to be chosen by an adversarial player, the *maximin effects* are then of a minimax regret form as they optimize the

objective function (explained variance) under the most adversarial scenario. Mini-max regret strategies have also been explored in game theoretical aspects of decision theory and machine learning, for example, in Cesa-Bianchi and Lugosi (2006) and Zinkevich et al. (2007), and bandit-type problems and on-line decision problems, [e.g., Audibert, Bubeck and Lugosi (2014), Auer, Cesa-Bianchi and Fischer (2002), Bartlett et al. (2008), Foster and Vohra (1999), Lai and Robbins (1985)]. We do not allow in our framework any choice about which distribution we sample from, contrary to bandit-type problems. Related to our setting is a paper by Eldar, Ben-Tal and Nemirovski (2004), who propose a linear minimax regret estimator which can be computed with convex optimization. Their estimator is optimizing a mean-squared error loss subject to various source uncertainties in the data. Our set-up is conceptually perhaps closest to the minimax framework in robust statistics [Huber (1964)]. However, we consider much more general situations than contaminated samples with a fraction of outliers: as discussed in Section 3.2.3, the latter fits into our framework as well.

Second, we have shown two different objectives (minimizing residual variance $R_{\beta;b}$ and maximizing explained variance $V_{\beta;b}$) that yield two different minimax-regret estimators ($b_{\text{pred-maximin}}$ and b_{maximin}). Using the first choice of minimizing residual variance has the main drawback that it is nonrobust when sampling regression coefficients: adding a small point mass to F_B can change the effects $b_{\text{pred-maximin}}$ drastically. The same is true for the pooled effects (5). Explained variance $V_{\beta;b}$ is expressed as residual variance if measured against the baseline of residual variance $R_{0,b}$ under a constant 0 prediction (we assume a mean-centered response throughout). This baseline is often appealing in practice as we would like to avoid doing worse than a constant prediction. Moreover, assume we choose instead a baseline of residual variance $R_{b_{\text{base}},b}$ for any vector b_{base} in the convex hull of the support of F (such as b_{pool} or $b_{\text{pred-maximin}}$). Whichever vector $\beta \neq b_{\text{base}}$ we choose in this scenario, we cannot avoid doing worse than the baseline for some values of $b \in F$ with the consequence that the problem will become trivial, as the optimal value under the most adversarial scenario can then always be achieved by a vanishing coefficient vector (thus keeping the baseline solution). Theorem 1 will provide a justification for this statement: once we shift the problem by the nonzero baseline b_{base}, the vector b_{base} will sit at the origin, and it will be a part of the convex hull of the equally shifted support of F, thus leading to a vanishing maximin-effect.

A simple two-dimensional example of "maximin effects" is shown in Figure 2. The coefficients are chosen as $B = (1, \eta)$, where η varies uniformly in $[-4, 6]$. The two random predictor variables are chosen independently with a standard normal distribution. The pooled estimator (5) is marked with a blue circle in the first panel of Figure 2, and the corresponding cumulative cross product in (1) is shown in the second panel if observations are ordered such that η decreases monotonically from 6 to -4. The pooled effects (5) maximizes the overall explained variance but suffers as η takes on negative values. Predictions in this range are even negatively
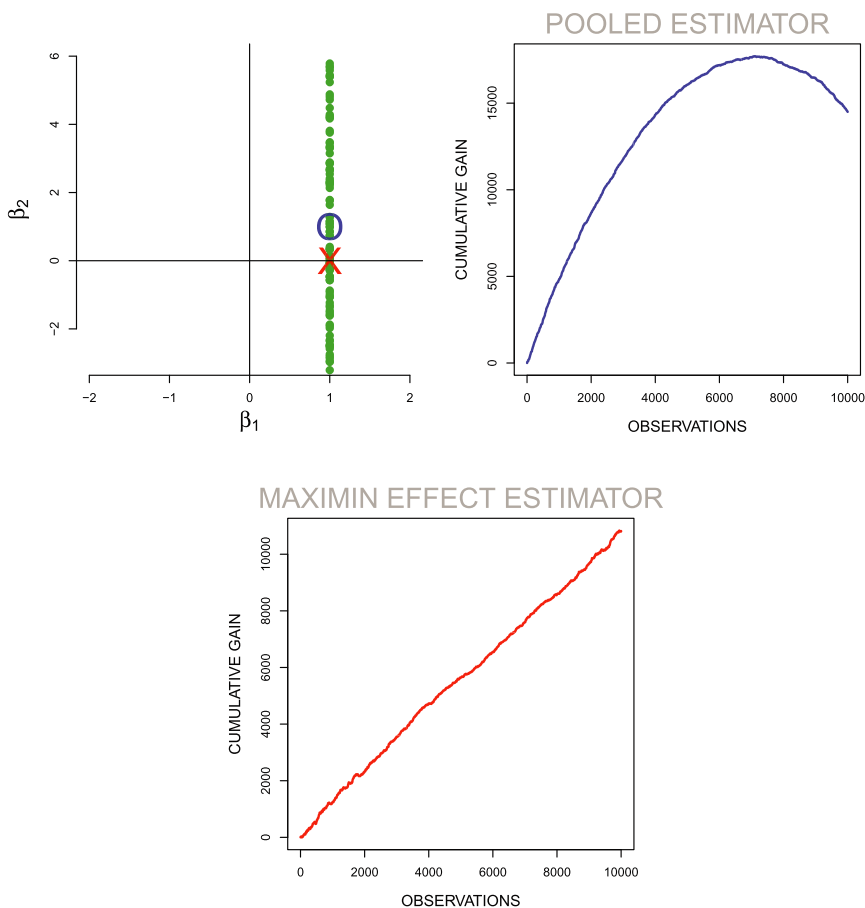
FIG. 2. *An illustration of the difference between the pooled effects b_{pool} in (5) and the maximin effects b_{maximin} in (7). First panel: the green dots indicate the values the random coefficient takes on with a maximin first component and a variable second component. The blue circle indicates the location of the pooled effects b_{pool}, while the red dot marks the location of the maximin effects b_{maximin}. Second and third panel: the cumulative cross product (1) for the pooled and maximin estimator, respectively. While the pooled estimator achieves a better overall fit, it does so at the cost of a highly variable performance.*

correlated with the responses, as can be seen by the negative slope of the cumulative cross product in the figure toward the right half of the observations. The effect is perhaps more drastic than in the real-data example in Figure 1 but of a similar nature. The maximin effects b_{maximin} in contrast take a nonzero value only for the first variables, where the effect is constant. This yields a constant explained variance throughout the whole parameter range, as shown in the third panel of Figure 2.

If we are in a classical regression model with a fixed regression coefficient vector, then F_B has just a point mass at some $b \in \mathbb{R}^p$, and (5), (6) and (7) will coincide.

The vector b_{maximin} is maximizing the explained variance under the most adverse realization of the random regression coefficient. The value 0 has a special status since we define the regret with respect to the 0 regression vector. Effects that can take opposite signs are set to 0 when using the maximin explained variance in b_{maximin} (similar to the value 0 getting special status when losing the rotational invariance in coordinate space when replacing a ridge penalty by a Lasso-type penalty).

The maximin effects have an interesting characterization.

THEOREM 1.   *Assume that the predictor variables are chosen randomly from a design with full-rank population Gram matrix Σ. Let H be the convex hull of the support F of F_B. The maximin-effect* (7) *is then given by*

$$b_{\mathrm{maximin}} = \operatorname*{argmin}_{\gamma \in H} \gamma^t \Sigma \gamma.$$

In particular, if $0 \in H$, then $b_{\mathrm{maximin}} \equiv 0$.

A proof is given in the Appendix. The maximin effects parameter is thus the one that is closest to the origin in the convex hull spanned by the support of F_B. In a classical regression setting with fixed regression coefficients β^*, F_B just has a point-mass at a β^*, and the maximin effect will, by Theorem 1, be identical to β^*. Figures 1 and 4 show examples of datasets with an interesting nonzero solution.

If the origin is included in the convex hull of all coefficients, the best lower bound that can be guaranteed is 0, and the maximin effects are consequently vanishing identically in this scenario. If the maximin effects vanish, a standard regression analysis will typically be misleading since the inner product between any estimated vector and true effects in F_B can take an arbitrary sign, depending on which effect in F_B is currently active. A vanishing maximin effect is thus a warning sign that standard regression analysis might not be appropriate.

Four examples of maximin effects are shown in Figure 3, comparing the pooled and the maximin effects. For $b_{\mathrm{pred\text{-}maximin}}$, as defined in (6), there is no equivalent characterization as in Theorem 1, as the value 0 has no special status.

It is noteworthy that the maximin effects are robust in the following sense: if we add new points to the support of F_B, we will always maintain or lower the distance to 0 as in Theorem 1. In the most extreme case, adding contaminations to the support of F_B will either leave the maximin effects unchanged or shrink the maximin effects toward 0. This is a direct consequence of Theorem 1.

We can also characterize the maximin effects in yet another way, using Theorem 1. Define the predictions and residuals as

$$\mathrm{Pred}_\beta := X\beta \quad \text{and} \quad \mathrm{Res}_{b,\beta} := Xb - X\beta.$$
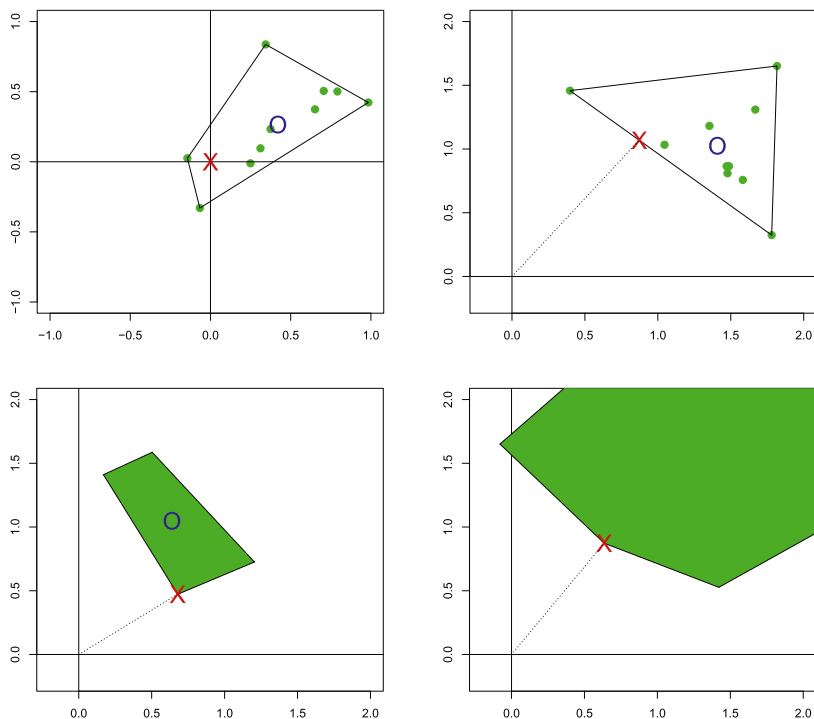
FIG. 3. *Four examples of the support of F_B (green dots or area), its convex hull (black lines), the pooled effects (blue circle) and the maximin effects (red cross). The maximin effects are the closest point to the origin in the convex hull of the support of F_B in the distance measure of Theorem 1. In the example on the first panel, the origin is contained in the convex hull, and the maximin effects thus vanish. In the second example the maximin effects rest on the convex hull of the support, but are not equal to zero. In both examples, the maximin effects are not part of the support itself. The third example shows a continuous support of F_B, while the last example has unbounded support of F_B. In both of these examples, the maximin effects are identical to a corner point of the support, but generalizations to the edge of the support are easily possible as well. In the last example, the pooled effects are thus infinite whereas the maximin effects have a robustness property by staying at the closest point to the origin.*

The maximin effects are then, using Theorem 1, characterized as the effects that maximize the norm of the predictions, subject to the constraint that the inner product between the predictions and the residuals is nonnegative for all possible $b \in F$,

$$(8) \qquad b_{\text{maximin}} = \underset{\beta}{\text{argmax}}\, E\big(\|\,\text{Pred}_\beta\,\|_2^2\big) \qquad \text{such that } \underset{b \in F}{\min}\, E(\text{Pred}_\beta, \text{Res}_{b,\beta}) \geq 0.$$

If B just takes a deterministic fixed value β^*, we recover of course $b_{\text{maximin}} = \beta^*$. The constraint in the optimization above requires that the predictions are never negatively correlated with the residuals if b can vary in F. The maximin predictions are in this sense the maximal predictions that are still "conservative" in the

sense that they can potentially "under-explain" a signal but can never be negatively correlated with the residuals.

In summary, if we want to maximize the explained variance if an adversary is allowed to pick a regression vector $b \in F$ or if test data are not expected to come from the same distribution as the training data with respect to the random coefficients, then estimating the maximin coefficients (7) seems a useful choice.

2.2. *Maximin effects estimator.* We introduce the maximin effects estimator first for data where we know a group-structure of the observations in the sense that within each group of observations the regression coefficient has a fixed (but unknown) value, which varies between groups.

To be more precise, suppose there are G groups of observations $g = 1, \ldots, G$, and each group has n_g samples. The indices belonging to a group are denoted by $\mathcal{I}_g \subset \{1, \ldots, n\}$ for all groups $g = 1, \ldots, G$. Let $X_{\mathcal{I}_g} = X_g$ denote the $n_g \times p$-dimensional submatrix of X that corresponds to choosing the rows in $\mathcal{I}_g$ and likewise for $Y_g = Y_{\mathcal{I}_g}$ and $\varepsilon_{\mathcal{I}_g} = \varepsilon_g$. If we are in situation where we know that the random coefficient is fixed at b_g within a group, then

$$Y_g = X_g b_g + \varepsilon_g, \qquad g = 1, \ldots, G.$$

Let $\hat{\Sigma}_g = n_g^{-1}(X_g^t X_g)$. The empirical counterpart to (4) is the explained variance in group g,

$$\hat{V}_\beta^g := \frac{2}{n_g} \beta^t X_g^t Y_g - \beta^t \hat{\Sigma}_g \beta.$$

A natural estimator for a sparse, consistent signal b_{maximin} is then a penalized version of the empirical minimizer. For $q \in [1, 2]$,

$$(9) \qquad \hat{\beta}^\lambda = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, L(\beta) + \lambda \|\beta\|_q \qquad \text{where } L(\beta) = \max_{g=1,\ldots,G} \left(-\hat{V}_\beta^g\right).$$

For $G = 1$, the loss function $L(\beta)$ is identical to quadratic loss up to a constant. For $G > 1$, however, the loss function will be different from quadratic loss. If $p \ll \min_g n_g$, one can use the unpenalized version ($\lambda = 0$). In the general case, the two most interesting choices for the penalty are $q = 1$, making the estimation lasso-like [Chen, Donoho and Saunders (2001), Tibshirani (1996)], and $q = 2$ for a ridge-type estimation [Hoerl and Kennard (1970)]. An equivalent version is given by the constrained optimization,

$$(10) \qquad \hat{\beta}^\kappa = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \, \max_{g=1,\ldots,G} \left(-\hat{V}_\beta^g\right) \qquad \text{such that } \|\beta\|_q \leq \kappa,$$

and we will mostly use the constrained version for the theoretical results. In practice the two versions can be used interchangeably, and the penalty parameter can be chosen by cross-validation, using hold-out samples for each unknown group and choosing the penalty parameter that maximizes the minimally explained variance on the hold-out samples from all groups.

The objective function in (9) or (10) is convex in its argument and can thus relatively easily be optimized; we will return to this issue later in Section 4.

2.3. *Maximin effects estimator for unknown groups.* In some applications there are no a priori known groups on which the realized value of the regression coefficients shows little or no variation. However, if the observations have, for example, a time ordering, and the effects are changing smoothly over time, we would suspect that taking blocks of consecutive observations would result in little variability of the realized coefficients within groups.

For some datasets, the groups are entirely unknown; see Städler, Bühlmann and van de Geer (2010) for an example. We propose in these cases to take G groups of m observations, where m is approximately of size n/G (modulo rounding to the next integer) if we sample without replacement. Alternatively, we can sample G groups with m observations each with replacement such that typically $Gm > n$.

For both cases mentioned above, once we have constructed the G groups, we use the same estimator as in (9) or (10). We discuss in Sections 3.2.2–3.2.3 the validity of the procedure based on such estimated groups.

3. Properties. The statistical properties of the lasso-type maximin effects estimator (10) with the ℓ_1-norm constraint will be examined first for the case of known groups in the observations and later be extended to the considerably more involved case (both from a theoretical and practical perspective) of unknown groups, either capturing smooth varying effects (over time) or more generally without such a (time) structure.

3.1. *Known groups.* Here we show a result for the lasso-type maximin effect estimator (10) for known groups of observations.

Specifically, there are G groups, and for simplicity, we assume that each group has $n_g \equiv n/G$ observations (without this assumption, we need to replace in the results below n/G by $\min_g n_g$). In each group, the explanatory variables are chosen randomly with Gram matrix Σ, yielding design matrices X_g, $g = 1, \ldots, G$. In each group,

$$Y_g = X_g b_g + \varepsilon_g \qquad \text{for } g = 1, \ldots, G,$$

for coefficients $b_g \in \mathbb{R}^p$ that are fixed in each group but can vary between groups. The maximin estimator is then the set of coefficients that work optimally across all groups in the sense of (7).

With estimator (10), we are now trying to maximize the explained variance in all groups.

THEOREM 2. *Let D be the maximal difference between the empirical Gram matrix $\hat{\Sigma}_g$ and population Gram matrix Σ, $D = \max_g \|\hat{\Sigma}_g - \Sigma\|_\infty$. If $\kappa \geq \max_g \|b_g\|_1$, then*

$$(11) \qquad \min_{b \in F} V_{\hat{\beta}^\kappa; b} \geq V^* - 6D\kappa^2 - \frac{4}{n/G} \max_g \|X_g^t \varepsilon_g\|_\infty \kappa,$$

where V^* is the optimal value that can be attained,

$$(12) \qquad V^* = \max_{\beta} \min_{b \in F} V_{\beta;b}.$$

A proof is given in the Appendix. For $D = 0$ (if the population and empirical versions of the Gram matrices are identical, as happens under fixed design; see a more detailed comment below), the estimator thus reaches the optimal value less a term

$$\max_g \frac{4}{n/G} \|X_g^t \varepsilon_g\|_\infty \max_g \|b_g\|_1,$$

which is a similar result to that of standard lasso estimation [see, e.g., Bühlmann and van de Geer (2011)], except that the first term $4(n/G)^{-1} \max_g \|X_g^t \varepsilon_g\|_\infty$ will increase when the number G of groups grows larger, which is the price we have to pay for estimating the maximin effects (7) instead of the pooled effect (5). On the other hand, the error is just a function of the noise ε and not influenced by the variability of b_g across groups, whereas standard lasso-type estimation of the pooled effect (5) would suffer if the variability of b_g is high across groups.

We note that one can also derive a similar bound if the Gram matrix of the predictors is allowed to depend on the group. In particular, we can have a fixed design in each group. In this case a corresponding result holds true with $D = 0$. If the design is random as in Theorem 2, we have an additional term in the bound that is proportional to D times the squared ℓ_1-norm of b_{maximin} and b_g, $g = 1, \ldots, G$. A more careful analysis for the special case of Gaussian random design [Raskutti, Wainwright and Yu (2010)] could render the bound again linear in κ, with more general design treated in Lecué and Mendelson (2015).

The two terms that are relevant for the rate are thus D and $\max_g \|X_g^t \varepsilon_g\|_\infty$. To give a simple bound for D if all predictor variables are drawn from the same population with Gram matrix Σ, we can, for example, get the following:

LEMMA 1. *If the predictor variables are chosen i.i.d. from a distribution with Gram matrix Σ and $\|X_i\|_\infty \leq M$ for $i = 1, \ldots, n$, then for any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*

$$(13) \qquad D^2 \leq \frac{2M^2}{n/G} \log\left(\frac{2p^2 G}{\alpha}\right).$$

The proof follows directly from Hoeffding's inequality, combined with a union bound over both the p^2 entries in each empirical Gram matrix and the number G of groups. Of course, a similar bound could be derived for a Gaussian or sub-Gaussian distribution of the explanatory variables.

If we additionally make a distributional assumption for the independent noise to control the term $\max_g \|X_g^t \varepsilon_g\|_\infty$, we get a simple bound for the estimator in (10).

COROLLARY 1. *Assume that the predictor variables are chosen i.i.d. from a distribution with Gram matrix Σ and $\|X\|_\infty \le 1$. If the errors $\varepsilon_i, i = 1, \ldots, n$ have a i.i.d. Gaussian distribution $\mathcal{N}(0, \sigma^2)$, then if $\kappa \ge \max_g \|b_g\|_1$, with probability at least $1 - \alpha$,*

$$\min_{b \in F} V_{\hat{\beta}^\kappa; b} \ge V^* - \frac{1}{\sqrt{n/G}} \left[6\sqrt{2\log\left(\frac{2p^2 G}{\alpha}\right)} \kappa^2 + 4\sigma \sqrt{2\log\left(\frac{2pG}{\sqrt{2\pi}\alpha}\right)} \kappa \right].$$

The error features the same $(n/G)^{-1}$ rate as lasso estimation on a single block of homogeneous data with n/G observations of a fixed signal. The success hinges obviously on the sparsity of the maximin solution. The bound becomes less tight when κ grows. Observe that κ is constrained from below by the sparsity of the regression coefficients. The problem thus becomes easier for sparse regression coefficients as one would expect from standard sparse regression [Bühlmann and van de Geer (2011)].

In summary, the maximin effects estimator (10) is able to estimate the maximin effects in a dataset with known groups. Note that an alternative would involve computing the Lasso-type estimator on each group and then constructing the estimator that yields the best minimally explained variance across all groups. In presence of a large number G of groups, the statistical properties of such a naive alternative procedure are unclear.

3.2. *Unknown groups*. The more difficult scenario is a mixture model, where there is no a priori known group structure for the observations, and each observation has its own realized value of the random coefficients. We assume that each coefficient $B \in F = \text{support}(F_B)$, where F is compact.

As mentioned previously, for the case of unknown groups, one solution is to apply estimator (10) to chosen groups, which can be chosen at random in the absence of any group information or in some way that reflects prior knowledge, for example, in the case coefficients varying over time.

3.2.1. *Pareto condition*. We will need to make one main assumption for recovery of the maximin coefficients for unknown groups, which will be discussed in a few examples below.

First, we define an essential subset of regression coefficients.

ASSUMPTION 1 (Essential subset). *A set $A = \{b_j; b_j \in F\}_{j=1,\ldots,d}$ is called an essential subset of $F = \text{support}(F_B)$ if the maximin effects for $B \sim \tilde{F}_B$, where the support of $\tilde{F}_B$ is A, are identical to the maximin effects as for the original problem with $B \sim F_B$.*

An essential subset is at most of cardinality $d \le p$ (if $d > p$, at least one b_j could be removed without changing the point that is closest to the origin in the convex hull of these points).

Two examples serve as simple illustrations: if $b_{\text{maximin}} \in F$, then the smallest essential subset is just b_{maximin} itself. If F is discrete, then an essential subset is always the support of F itself.

We now give the so-called Pareto condition which will be shown to be sufficient for recovery. For known groups, we do not need the condition, as it always fulfilled, as in Section 3.2.1. The condition is meant for cases where the groups are sampled randomly according to some mechanism, which we discuss with a few examples and settings in Section 3.2.1.

ASSUMPTION 2 (Pareto condition).   *Let $I_g \subset \{1, \ldots, n\}$ be the index sets for chosen groups $g = 1, \ldots, G$, and let B_i, $i = 1, \ldots, n$ be the regression coefficients at observation $i \in \{1, \ldots, n\}$. The assumption is that, for $\gamma \in (0, 1)$, with probability $1 - \gamma$ with respect to the random coefficients B_i, $i = 1, \ldots, n$ and a potentially random sampling of the groups, there exists an essential subset A of F such that for each $b \in A$ there exists a group $g \in \{1, \ldots, G\}$ for which $B_i = b$ for all $i \in I_g$.*

We call this the Pareto condition since it implies that the maximin vector is optimal in the sense that making the performance better in one group will make the performance worse in another group. The condition requires some of the groups to be "pure" in the sense that all observations in the group correspond to the same realization of the regression vector. We emphasize that the Pareto condition is formulated as the probability of a certain event: we find this construction simpler than requiring a random event condition.

The Pareto condition is fulfilled for a few examples which will be discussed further in Section 3.2.3, but the condition is not true in general. Without appropriate structure (of the type shown in the examples) cases exist where the condition is violated.

3.2.2. *Recovery assuming the Pareto condition.*   Using the Pareto condition, we get the following theorem for randomly sampled groups in the estimator (10):

THEOREM 3.   *Assume the Pareto condition is fulfilled, with corresponding probability $1 - \gamma$ for some $\gamma \geq 0$. If X_i, $i = 1, \ldots, n$ are i.i.d. from a distribution with Gram matrix Σ and $\|X\|_\infty \leq 1$ and $\kappa \geq \max_g \|b_g\|_1$, then with probability at least $1 - 3\alpha - \gamma$,*

$$\min_{b \in F} V_{\hat{\beta}^\kappa; b} \geq V^* - \frac{M}{\sqrt{m}} \quad and$$

$$(14) \qquad (\hat{\beta}^\kappa - b_{\text{maximin}})^t \Sigma (\hat{\beta}^\kappa - b_{\text{maximin}}) \leq \frac{M}{\sqrt{m}}$$

$$with \ M = 6\sqrt{2 \log\left(\frac{2p^2 G}{\alpha}\right)} \kappa^2 + 4 \max_g \frac{1}{\sqrt{m}} \|X_g^t \varepsilon_g\|_\infty \kappa.$$

A proof is given in the Appendix. If the smallest eigenvalue of the population covariance matrix Σ is bounded from below by some $\lambda_{\min} > 0$, then it follows further that with the same probability as above, $\|\hat{\beta}^\kappa - b_{\text{maximin}}\|_2^2 \leq M/(\sqrt{m}\lambda_{\min})$.

If the error distribution is Gaussian, we get the following:

COROLLARY 2. *If the assumptions of Theorem 3 are fulfilled, and additionally the errors ε_i, $i = 1, \ldots, n$ have a i.i.d. Gaussian distribution $\mathcal{N}(0, \sigma^2)$, then, with probability at least $1 - 3\alpha - \gamma$, the constant M in Theorem 3 can be chosen as*

$$M = 6\sqrt{2\log\left(\frac{2p^2 G}{\alpha}\right)}\kappa^2 + 4\sigma\sqrt{2\log\left(\frac{2pG}{\sqrt{2\pi}\alpha}\right)}\kappa.$$

This result is a generalization of Corollary 1. If we choose $m = n/G$, we obtain the results of Corollary 1. (Note, however, that we need the Pareto condition for Corollary 2 but not Corollary 1.) However, the results also show that we can choose m much larger than n/G by allowing an observation to appear in multiple groups, thus lowering the statistical error. Another point of view is that we keep the sample size in each group fixed but increase the number of groups G, thus increasing the chance that the Pareto condition will be fulfilled. We can thus infer the optimal maximin coefficients by randomly sampling groups and applying the maximin estimator (10) to these groups. The success hinges on the sparsity of the coefficients within the support of the distribution of the random coefficients.

We describe in Section 5 a cross-validation method for choosing the number of groups.

3.2.3. *Examples where the Pareto condition is fulfilled.* Theorem 3 rests on the Pareto condition. It is evident that an arbitrary random sampling scheme cannot lead to success in the sense of Theorem 3. The number G of groups, for example, plays a crucial part. Setting $G = 1$ leads just to pooled estimation, which yields in general a consistent estimator for the pooled coefficients and can thus not consistently estimate the maximin coefficients if they differ from the pooled coefficients.

Fixed groups with fixed design. The simplest example where the Pareto condition is fulfilled is the case of known groups, where B takes a single value b_g within each group $g = 1, \ldots, G$. By definition of the maximin coefficients, the Pareto condition is then fulfilled, and we are back to the setting of Section 3.1.

Chronological observations with a jump process. Assume the observations have a time-ordering, and we have a change-point model. Consider the case where the support of F_B is finite of cardinality J, that is, $F = \{b_1, \ldots, b_J\}$. Assume that

B for the first sample, namely B_1, is chosen uniformly at random among the J different possibilities. Thereafter, for $i = 2, \ldots, n$ and some $\delta \in (0, 1)$

$$B_i = \begin{cases} B_{i-1}, & \text{w.p. } 1 - \delta, \\ b_j, & \text{w.p. } \delta/J \text{ for all } j = 1, \ldots, J. \end{cases}$$

We build G groups of consecutive observations. Below, we will further show under which conditions on G and J the Pareto condition is fulfilled with high probability. The Pareto condition is fulfilled if we have for each possible value $b_1, \ldots, b_J$ a $g \in \{1, \ldots, G\}$ such that $B_i = b_j$ for all observations i in the gth set. Suppose we fix G and condition on $B_{i'} = b_j$ for some $j \in \{1, \ldots, J\}$ and some $i' \in \{1, \ldots, n - 2n/G\}$. Let L be the conditional length of the segment of observations $i \geq i'$ where $B_i = b_j$. Then

$$P\left(L \geq 2\frac{n}{G}\right) \geq (1 - \delta)^{2n/G} \geq 1 - \frac{2n\delta}{G}.$$

If indeed, $B_{i'} = b_j$ and $L \geq 2n/G$, then one block of observations of length n/G is guaranteed to have exclusively realizations of B equal to b_j. The probability that there is at least one i' for which $B_{i'} = b_j$ in $\{1, \ldots, n - 2n/G\}$ is greater than

$$1 - \left(1 - \frac{\delta}{J}\right)^{n - 2n/G} \geq 1 - \exp\left(-\frac{\delta}{J}(n - 2n/G)\right).$$

Using a union bound over all J distinct values the coefficients can take, the Pareto condition is fulfilled with corresponding probability at least $1 - \gamma$ for $\gamma \in (0, 1)$ if

$$G \geq 4\frac{n\delta J}{\gamma},$$

$$\frac{\delta(n - 1)}{J} \geq 1/\log\left(\frac{2J}{\gamma}\right).$$

The second condition states that the number of distinct classes J cannot be too large. More specifically, $\delta n/J$ is approximately the average number of contiguous blocks of observations that have a realization b_j of the random coefficient. The condition above implies that this average value needs to be larger than 1 for the scheme to work (as otherwise a value of the coefficients might not be sampled at all).

The first condition is a requirement on the number of groups G one has to pick. It yields an effective sample size n/G of order δ^{-1}, which is the order of the length of observations where the regression coefficient stays constant.

Contaminated samples and robustness. Assume that the regression coefficients come from a mixture distribution

(15) $$B = \begin{cases} b^*, & \text{with probability } 1 - \epsilon, \\ U, & \text{otherwise}, \end{cases}$$

where U follows a distribution F_U such that

$$(16) \qquad (u - b^*)^t \Sigma b \geq 0 \qquad \forall u \in \text{support}(F_U).$$

Note that the latter condition implies that $b_{\text{maximin}} = b^*$. A fraction ϵ of samples are contaminated in the sense that they have a different realized value of B.

We build G groups of observations by random sampling. Each group consists of m samples drawn at random without replacement from all n observations, and each group is sampled independently. We will argue under which circumstances the Pareto condition is fulfilled with high probability.

The Pareto condition is trivially fulfilled if we have a single group of observations where all realizations are identical to $b^* = b_{\text{maximin}}$. Suppose we divide the samples into G groups. Each group contains m observations, drawn at random without replacement from all n observations, independently for each group (and thus, the same re-sampled data point can occur in several groups). If for $\gamma \in (0, 1)$

$$(17) \qquad G \geq \frac{\log(1/\gamma)}{\log(1 - (1 - \epsilon)^m)},$$

then we guarantee the Pareto condition will be fulfilled with corresponding probability at least $1 - \gamma$ with respect to the random sampling of the coefficients and random sampling of the groups.

There is also a robustness inherent in the procedure. If sampling (15) holds without condition (16), then the samples U can come from an arbitrary distribution. If condition (17) is fulfilled, then we have again with probability at least $1 - \gamma$ that there is a group where B is equal to b^* for all observations in the group. We can then use Theorem 1 to show robustness properties of the estimate, as already discussed in the paragraphs after Theorem 1. Adding contaminated samples can only shrink the maximin effects parameter and the corresponding estimator toward the origin. The maximin effects estimator thus has robustness properties against outliers as long as at least one group does not contain outliers.

Some more examples are possible to derive, including for continuous distributions of B, but are beyond the scope of this manuscript. The basic intuition is that the convex hull of the effective coefficients in each groups needs to approximate the convex hull of the support of the random coefficients B in order for the Pareto condition to be fulfilled.

The outliers above are referred to as *b-outliers* in linear mixed models [McCulloch and Neuhaus (2001)]. An interesting question is whether the method is also robust to outliers in the noise, the so-called *e-outliers*. If we use a robust version of the explained variance $V_{\beta;b}$ in the maximin estimator definition, then the breakdown point of the maximin estimator will at least be identical to the breakdown point of the robust estimator for the explained variance. The reason is that the explained variance would have to be corrupted arbitrarily much in every of the G groups, requiring in each group $g = 1, \ldots, G$ at least $\lceil \rho n_g \rceil$ corrupted samples if $\rho \in (0, 1)$ is the breakdown point of the robust explained variance estimator.

Hence at least $\sum_g \rho n_g = \rho n$ samples would have to be corrupted for the modified maximin estimator to take on arbitrarily large values, and the breakdown point of a robust explained-variance estimator would thus be inherited by the maximin estimator.

Before presenting some numerical results, we first discuss now the computational aspects of the procedure.

4. Computation. The objective function of estimator (9) is convex, and the penalty is separable. Estimator (9) or the equivalent constrained formulation (10) could thus be computed using coordinate-wise updates, with a similar strategy as in the "glmnet" approach [Friedman, Hastie and Tibshirani (2009)] to fitting lasso- and ridge-penalized regression models. If p and n are large, this becomes computationally burdensome. We show two different possibilities.

4.1. *Iteratively reweighted estimation.* The estimation can be reduced to a series of weighted standard lasso or ridge estimation. The minimum in (10) can be approximated for positive terms by a sum

$$(18) \qquad \lim_{\zeta \to 0} \left(\sum_{g=1}^{G} (\hat{V}_{\beta}^{g})^{\zeta} \right)^{1/\zeta}.$$

This leads to a weighted estimation problem, where the weights are constant in each group, and weights are larger in groups where the explained variance is still small. For a fixed value of $\zeta > 0$, the solution of (18) is (setting $q = 1$ for Lasso-type estimation and $q = 2$ for ridge)

$$(19) \qquad \operatorname*{argmin}_{\beta \in \mathbb{R}^p} -2 \sum_{i=1}^{n} w_i Y_i (X\beta)_i + \sum_{i=1}^{n} w_i (X\beta)_i^2 + \lambda \|\beta\|_q,$$

where the weights w_i, $i = 1, \ldots, n$ are proportional to

$$(20) \qquad w_i \propto (V_{\hat{\beta}}^{g_i})^{\zeta - 1},$$

where g_i is the group that observation i belongs to. The strategy is now to alternate between updating the weights in (20), starting with uniform weights, and computing the solution to (19) for fixed weights. The solution in the first example in Figure 1 has been computed in this way, as a series of reweighted least squares estimators with $\zeta = 0.01$. A few rounds of the iteration are typically sufficient, and the computational burden is thus similar to standard least squares or lasso-type estimation.

4.2. *Computationally efficient solution for maximal penalties.* Computing estimator (9) by coordinate-wise updates or by iteratively reweighted penalized estimation requires, however, that either the design matrix X or the Gram matrices in all groups are kept in memory.

Another option is to look at the limit of $\hat{\beta}^\lambda$ as $\lambda \to \lambda_{\max}$, where $\lambda_{\max}$ is the supremum of the values for which the estimator does not vanish identically. In this limit,

$$\frac{\hat{\beta}^\lambda}{\|\hat{\beta}^\lambda\|_2} \to \frac{\hat{\beta}}{\|\hat{\beta}\|_2},$$

where $\hat{\beta}$ is the solution to

(21) $$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{such that} \quad \min_{g=1,\dots,G} (\beta^t X_g^t Y_g) \geq 1.$$

The quadratic term disappears in (21) as it will shrink like κ^2 if $\|\beta\|_1 = \kappa$, whereas the remaining two terms (penalty and objective) in the estimator scale linearly with κ, and the constant κ thus drops out of the equation or can be replaced with an arbitrary constant (modulo scaling of the solution) as $\kappa \to 0$. The constant 1 is chosen arbitrarily, and choosing a different constant would just rescale the solution. The estimator $\hat{\beta}$ in (21) can be computed with linear programming, and most importantly, the data matrix X_g enters only through its inner product with the response vector Y_g, achieving a great reduction in problem size. Estimator (21) still has to be re-scaled for optimal least squares prediction, but this is just a univariate optimization. Our only tuning parameter in this case is the number of groups G to choose (unless they are known), and we can optimize G by using cross-validation; see the section with numerical examples for details on how the cross-validation is implemented.

The solution $\hat{\beta}$ in (21) selects in general several variables, not just a single one as might be expected from the analogous situation for the standard lasso. For ridge estimation with $q = 2$, estimator (21) would correspond to marginal regression if we had only a single group, and this behavior has recently been examined in Genovese et al. (2012). However, the variability of the inner products in (21) across groups leads to sometimes appreciably different solutions and has a similar effect as the quadratic penalty that is written down explicitly in (9). We will use this latter estimation technique in (21) for the following high-dimensional example in Section 5, which will also demonstrate the computational advantages of this type of estimation.

The maximal penalty solution is suitable if either a fast initial estimator is desirable or if the data are very noisy. In the latter case the large penalty will be justified not only from a computational but also from a statistical point of view. The performance of the maximal penalty estimator should be seen as a lower bound for what is achievable with a more expensive estimation with a fine-tuned value of the penalty parameter.

While tight bounds on the worst-case and typical computational complexity are difficult to establish, the memory requirements are more immediate. If fitting a pooled estimation, the required memory is of order $O(\min\{p^2, np\})$, as either the whole matrix X or the Gram matrix has to be held in memory. For standard maximin estimation or estimation of mixture models with G groups, this is increased to $O(\min\{Gp^2, np\})$ since the Gram matrix has to be stored separately for all G groups. For the maximal penalty estimator (21) or its ridge counter-part, the memory complexity decreases to $O(pG)$, as one only has to store the G p-dimensional cross-products between the predictors and the responses in each group. The memory complexity of the maximal-penalty estimator is thus substantially reduced in the typical scenario where $G \ll \min\{n, p\}$, while the memory complexity is just marginally increased for the general case with arbitrary λ.

5. Numerical example. The example in Figure 1 illustrated that overfitting has to be a concern even if we have millions of observations at our disposal to fit quite low-dimensional models with less than one hundred parameters due to the shifts in the underlying distributions.

Next, we look at an example with millions of variables and thousands of observations. Kogan et al. (2009) collected a dataset of so-called 10-K reports from thousands of publicly traded U.S. companies in the years 1996–2000. For each report, unigrams and bigrams of word frequencies have been computed and used as predictors for the stock return volatility in the twelve-month period after the release of the report, which is here measured against the baseline of the volatility before the filing of the report. We use 3000 examples as a training set and the remaining just over 16,000 examples for testing. We compute both a cross-validated lasso and ridge estimator with the "glmnet" package [Friedman, Hastie and Tibshirani (2009)] and the estimator (21), once for a fixed number of 3 groups and once for a cross-validated number of groups, which is explained in the next paragraph. Below, we will further explain the procedure for cross-validation.

The histograms of

$$(22) \qquad \qquad \sum_{i \in I} Y_i \hat{Y}_i$$

are shown for the various methods, where both Y and $\hat{Y}$ are standardized to have mean 0 and variance 1. The groups in I are chosen randomly as 500 observations out of the training or test data. Form (22) avoids the choice of the scaling for the estimators of form (21) as the measure is invariant under rescaling of the predictions.

The results for lasso estimation ($q = 1$) are shown in Figure 4 and for ridge estimation ($q = 2$) in Figure 5 when selecting a varying number of predictor variables $p \in \{10^3, 10^4, 10^5, 10^6, 4{,}272{,}227\}$ as a consecutive block in the order given by the dataset in Kogan et al. (2009). Both the lasso and ridge estimators of the maximin effects are calculated under the maximal penalty (21), which has computational advantages and avoids having to chose a tuning parameter for the penalty.
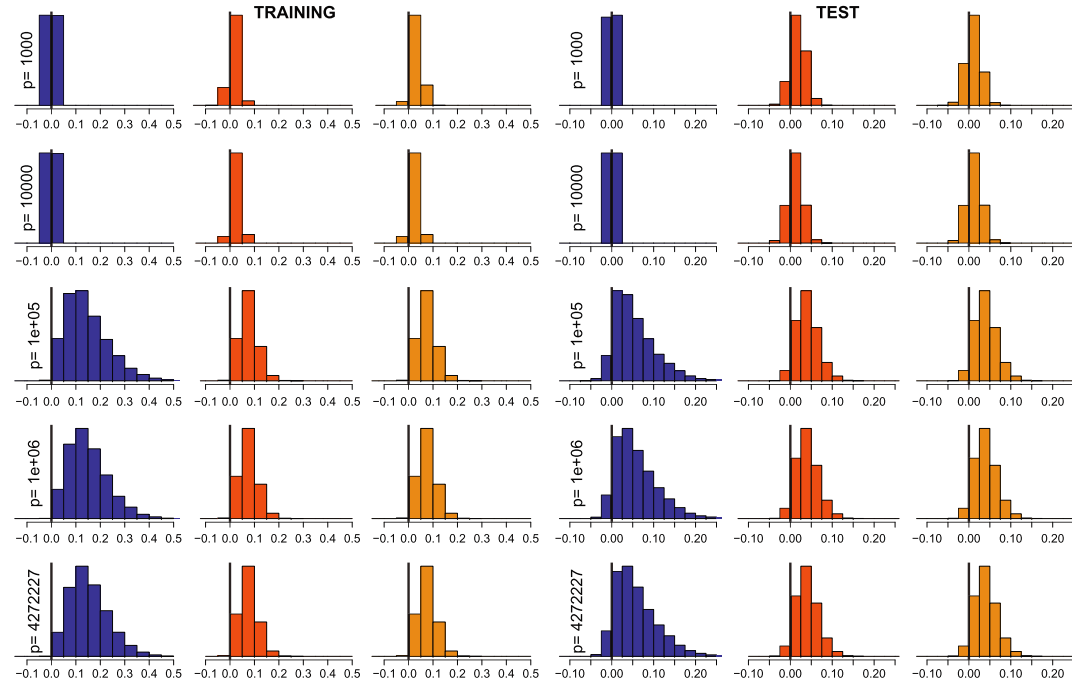
FIG. 4. *The histograms of the cross product* (22) *for lasso-type estimation on the training data* (*left three columns*) *and test data* (*right three columns*). *The predictions are standardized, and the measure is thus equivalent to explained variance after rescaling of the predictions. The number of predictor variables is increasing from the top to the bottom row, from* 1000 *to* 4,272,227. *The three columns in each panel correspond to standard cross-validated lasso estimation* (*blue*) *and the two maximin effects estimators with a fixed number of* $G = 3$ *groups* (*red*) *and a cross-validated choice of* G (*orange*). *In both training and test data, the explained variance is more variable under the pooled estimation than when estimating the maximin effects, while the average explained variance is similar. There is little difference between the estimator with a cross-validated choice of the number of groups and a fixed number of* $G = 3$ *groups. For* $p \leq 10^4$, *the cross-validated lasso estimator returns an empty model while the maximin estimation still finds some signal in the data, even if it is weaker than when using* $p \geq 10^5$ *predictor variables.*
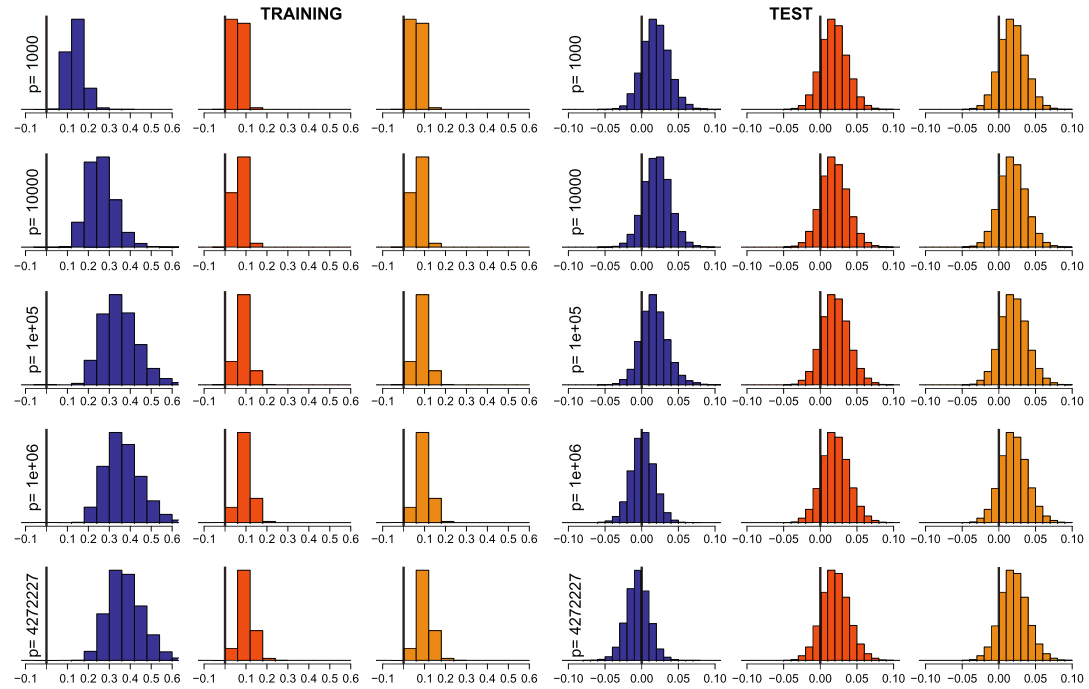
FIG. 5. *The results for ridge regression, analogous to Figure* 4. *The gap between training and test performance is much more pronounced for the pooled ridge estimate* (*blue*) *compared to the estimators of maximin effects* (*orange and red*). *Moreover, the probability of having a subset of observations with very small* (*or negative*) *explained variance is slightly higher for the pooled estimation. Estimation of the maximin effects was also three orders of magnitudes faster for the fixed number of groups* G (*red*) *than the pooled estimation* (*blue*).

As can be seen from the figures, the variability of the explained variance is indeed higher for pooled estimation, compared with the maximin effects estimators, especially for a lasso-type penalty. The difference in performance between training and test datasets is also larger for the pooled estimation as it is more prone to overfitting than the maximin effects estimators.

Cross-validation. As the dataset is not a priori grouped and the optimal group size is unknown, one needs to decide on the optimal number of groups G. One possibility indicated above is with cross-validation. To this end, we split the $n_{\text{train}} = 3000$ training data 100 times into two half-samples of $n_{\text{train}}/2$ observations, sampled uniformly at random. For each split, we then divide both half-samples of observations into smaller blocks of consecutive observations. (We would sample at random if the data did not have a time-ordering.) The first half-sample of $n_{\text{train}}/2$ observations is split into G blocks with a sample size $n_{\text{train}}/(2G)$ in each block. The second half-sample is split into g blocks, where $g = 5$ is chosen as large as possible while still leaving at least a few hundred observations in each block. For each split into two half-samples, we compute the maximin estimator on the blocks formed by the first half-sample and compute the explained variance in each of the g blocks of the second half-sample. (The result turns out to be rather insensitive to the precise choice of g; note that we want to choose g as high as possible but have to keep a minimal number of observations in each of the "test" groups in order not to be overwhelmed by noise in the estimation of the explained variances.) The worst-case explained variance over these g groups is then averaged for each value of G across the 100 random splits into two half-samples. We choose G to optimize this averaged worst-case explained variance. All groups are chosen here as consecutive blocks of equal size from the two half-samples, respectively, since the reports are ordered chronologically and it seems likely that there are shifts in the underlying distributions over time. If no such time-ordering applies, we would sample the groups at random within each half-sample for cross-validation.

Computational aspects. Estimation of the standard estimator (9) is in general slower than the pooled estimation over all data, at least as long as (9) is computed by iteratively reweighted pooled estimation. On the other hand, when going for the maximal penalty estimate as in (21), the solution can be computed using quadratic or linear programming for ridge and lasso penalties, respectively, and the design matrix enters only through the inner products on the right-hand side of (21). Figure 6 shows the necessary computational time as a function of the dimensionality p of the data and the number n of samples. The advantage of the maximin effects estimator with a cross-validated choice of the number of groups is visible across the entire range of the dimensionality. The relative speed advantage of the maximin estimation is more than a factor 10 for ridge estimation. Choosing just a fixed number of groups can get the relative advantage to three orders of magnitude
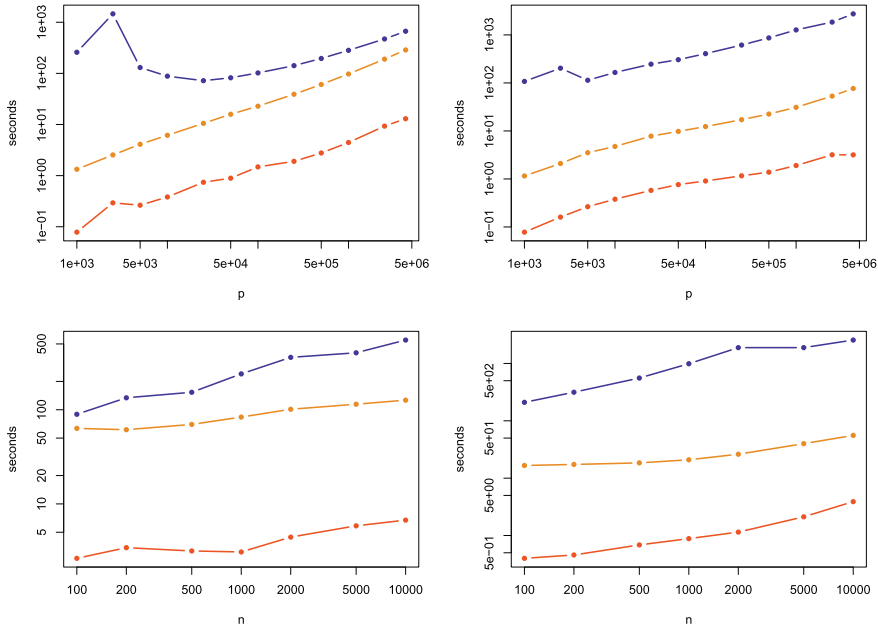
FIG. 6. *The timings in seconds of the three estimation methods as a function of the number of predictor variables for n = 3000 observations for Lasso-type estimation* (*left*) *and ridge estimation* (*second from left*): *the cross-validated pooled estimate* (*blue*), *and the two estimators of maximin effects with a fixed number of groups G = 3* (*red*) *and a cross-validated choice of the number of groups* (*orange*). *Analogous plots for the timings in seconds as a function of the number of samples for p = 10⁶ variables in the two right panels. Estimation of maximin effects was often orders of magnitudes faster than the pooled estimation.*

for ridge estimation, which can be useful in its own right or as an initial check as to whether there is any signal in the data at all. The computational complexity is roughly similar as a function of p for the methods whereas the maximin effects have a better scaling as a function of n, as expected since the dimension n drops out of the memory requirements for estimation and is replaced by the much smaller number G of groups.

6. Discussion. One characteristic of large-sale datasets is the mix of a large number of observations from different sources or different regimes. Due to the inhomogeneity of the data, estimating regressions or classifications or graphs over the pool of all available data is likely to estimate effects that might be strong for one part of the data but very weak or even of opposite sign for another part. Here, we proposed to estimate effects which are present for all possible groupings of the data (even if they might be masked by noise if we make the groups unreasonably small). The improvement in predictive accuracy can be seen empirically.

We have introduced the notion of maximin effects and proposed an estimator for these effects, using either a lasso or ridge-type penalty. If we have known groups

of observations with a different parameter setting in each group, the estimator is guaranteed to do as well in estimating the maximin effects as standard Lasso estimation would in estimating the average effect in a single group of these observations. For datasets with unknown groups, we have proposed to sample groups at random from the available observations. This has a similar flavor to "stability selection" [Meinshausen and Bühlmann (2010)] and the "bolasso" [Bach (2008)] or [Bradic (2013)], where models are fitted repeatedly over random (bootstrap) samples of the data. In contrast to these approaches, though, the estimator is trying to infer the "maximin effects" if the underlying regression coefficients change randomly, which is a novel concept. We have presented theoretical guarantees for the statistical accuracy, an efficient computational algorithm which is feasible for large-scale problems, as well as empirical results on real data demonstrating improved performance for prediction.

We expect that the notion of maximin effects is useful beyond regression and classification for 'big data" applications, both from a statistical as well as computational point of view, potentially helping to avoid detecting too many spurious effects that are not replicable.

APPENDIX

A.1. Proof of Theorem 1. Let H be its convex hull of the support of F. The maxim effects are given by definition as

$$b_{\text{maximin}} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \min_{b \in F} 2\beta^t \Sigma b - \beta^t \Sigma \beta$$

$$= \underset{\beta \in \mathbb{R}^p}{\arg\max} \min_{b \in H} 2\beta^t \Sigma b - \beta^t \Sigma \beta,$$

where the second equality follows by linearity of the objective function in b.

Let $C^t C = \Sigma$ be the Cholesky decomposition of Σ. Since we assumed Σ to be full-rank, C is invertible, and we define

$$\tilde{H} := \{Cz : z \in H\} \subseteq \mathbb{R}^p.$$

Then

$$(23) \qquad b_{\text{maximin}} = C^{-1} \underset{\xi \in \mathbb{R}^p}{\arg\max} \Big(\min_{u \in \tilde{H}} 2\xi^t u - \xi^t \xi \Big).$$

Define ξ^* to be the choice of weights in the simplex that is minimizing the ℓ_2-norm of the corresponding vector in the convex hull $\tilde{H}$,

$$(24) \qquad \xi^* := \underset{u \in \tilde{H}}{\arg\min} \|u\|_2^2.$$

The proof is complete if we can show that

$$(25) \qquad \xi^* = \underset{\xi \in \mathbb{R}^p}{\arg\max} \Big(\min_{u \in \tilde{H}} 2\xi^t u - \xi^t \xi \Big),$$

since this implies the result, using (23), the invertibility of C and

$$C^{-1} \underset{u \in \tilde{H}}{\operatorname{argmin}} \|u\|_2^2 = \underset{u \in H}{\operatorname{argmin}} u^t \Sigma u.$$

By definition of ξ^* in (24), it holds true that for every $\mu \in \tilde{H}$,

$$\forall 0 \le \nu \le 1: \qquad \|\xi^* + \nu(\mu - \xi^*)\|_2 \ge \|\xi^*\|_2,$$

since ξ^* and μ are from a convex set (namely $\tilde{H}$), and ξ^* minimizes the ℓ_2-norm over this convex set. Since equality holds for $\nu = 0$, the derivative of the left-hand side with respect to ν at $\nu = 0$ has to be positive, which is equivalent to

$$(26) \qquad 2(\xi^*)^t \mu - 2(\xi^*)^t \xi^* \ge 0 \qquad \text{for all } \mu \in \tilde{H}.$$

Hence

$$(27) \qquad 2(\xi^*)^t \mu - (\xi^*)^t \xi^* \ge (\xi^*)^t \xi^* \qquad \text{for all } \mu \in \tilde{H}.$$

Choosing $\xi = \xi^*$ yields thus a value of the objective function of at least $(\xi^*)^t \xi^*$ in (25).

On the other hand, choosing $u = \xi^*$ in (25), for all $\xi \in \mathbb{R}^p$,

$$(28) \qquad 2\xi^t \xi^* - \xi^t \xi \le (\xi^*)^t \xi^*,$$

with equality only if $\xi \equiv \xi^*$. The value of the objective function in (25) can hence not exceed $(\xi^*)^t \xi^*$. Choosing $\xi = \xi^*$ in (25) yields thus the optimal value of the objective function and is indeed a solution to (25), which completes the proof.

A.2. Proof of Theorem 2. We write $\hat{\beta}$ instead of $\hat{\beta}^\kappa$ to simplify notation, and use the constrained version of the estimator. Note that the theorem is for known groups, and thus $\{b : b \in F\} = \{b : b = b_g \text{ for some } g \in \{1, \ldots, G\}\}$. Let $\delta_g = (G/n) X_g^t \varepsilon_g$ for all $g \in \{1, \ldots, G\}$. Using the basic inequality, for any fixed vector ξ in the feasible region, that is, for all ξ with $\|\xi\|_1 \le \kappa$,

$$(29) \qquad \min_g \{2\hat{\beta}^t \hat{\Sigma}_g b_g - \hat{\beta}^t \hat{\Sigma}_g \hat{\beta} + 2\hat{\beta}^t \delta_g\} \ge \min_g \{2\xi^t \hat{\Sigma}_g b_g - \xi^t \hat{\Sigma}_g \xi + 2\xi^t \delta_g\}.$$

Using the definition of $D = \max_g \|\hat{\Sigma}_g - \Sigma\|_\infty$,

$$(30) \qquad \min_g \{2\hat{\beta}^t \Sigma b_g - \hat{\beta}^t \Sigma \hat{\beta}\} + 2D\|\hat{\beta}\|_1 \max_g \|b_g\|_1 + D\|\hat{\beta}\|_1^2 + 2\|\hat{\beta}\|_1 \|\delta_g\|_\infty$$

$$\ge \min_g \{2\xi^t \Sigma b_g - \xi^t \Sigma \xi\}$$

$$(31)$$

$$\qquad - 2D\|\xi\|_1 \max_g \|b_g\|_1 - D\|\xi\|_1^2 - 2\|\xi\|_1 \|\delta_g\|_\infty.$$

Hence, using $\xi = b_{\text{maximin}}$, and using that by definition of $\hat{\beta}$, $\kappa \ge \max\{\max_g \|b_g\|_1, \|\hat{\beta}\|_1\}$ (and hence, when using Theorem 1 that b_{maximin} is in the convex hull of F_B, also $\kappa \ge \|b_{\text{maximin}}\|_1$), it follows that

$$(32) \qquad \min_g \{2\hat{\beta}^t \Sigma b_g - \hat{\beta}^t \Sigma \hat{\beta}\} \ge V^* - 6D\kappa^2 - 4 \max_g \|\delta_g\|_\infty \kappa,$$

where

$$V^* = \min_g \{2b_{\text{maximin}}^t \Sigma b_g - b_{\text{maximin}}^t \Sigma b_{\text{maximin}}\},$$

which completes the proof.

A.3. Proof of Theorem 3. Starting as in the proof of Theorem 2, let $I_g \subseteq \{1, \ldots, n\}$ be the index set of the gth group. The explained variance in group g when using a regression vector $\xi \in \mathbb{R}^p$ can then be written as

$$\sum_{i \in I_g} 2(X_i \xi)(X_i B_i) - \sum_{i \in I_g} (X_i \xi)^2.$$

Analogous to (29), we have the basic inequality for all ξ with $\|\xi\|_1 \leq \kappa$,

$$\min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i \hat{\beta})(X_i B_i) - (X_i \hat{\beta})^2 + 2(X_i \hat{\beta})\varepsilon_i\big)$$

(33)

$$\geq \min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i \xi)(X_i B_i) - (X_i \xi)^2 + 2(X_i \xi)\varepsilon_i\big).$$

As the Pareto condition is fulfilled (with corresponding probability $1 - \gamma$), there exists a subset $\tilde{G} \subseteq \{1, \ldots, G\}$ such that $B_i = b_g$ for all $i \in I_g, g \in \tilde{G}$, and an essential subset is formed by $A = \{b_g; g \in \tilde{G}\}$. Restricting the minimum on the left-hand side of (33) over all groups in $\tilde{G}$, we have for all ξ with $\|\xi\|_1 \leq \kappa$,

$$\min_{g \in \tilde{G}} (2\hat{\beta}^t \hat{\Sigma}_g b_g - \hat{\beta}^t \hat{\Sigma}_g \hat{\beta} + 2\hat{\beta}^t \delta_g)$$

$$\geq \min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i \xi)(X_i B_i) - (X_i \xi)^2 + 2(X_i \xi)\varepsilon_i\big),$$

where $\delta_g = (1/m)X_g^t \varepsilon_g$. Hence, using $\|\xi\|_1 \leq \kappa$,

$$\min_{g \in \tilde{G}} \{2\hat{\beta}^t \hat{\Sigma}_g b_g - \hat{\beta}^t \hat{\Sigma}_g \hat{\beta}\} + 4\kappa \|\delta_g\|_\infty \geq \min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i \xi)(X_i B_i) - (X_i \xi)^2\big).$$

Analogous to (30), the first term on the left-hand side is bounded with probability at least $1 - \alpha$ by

$$\min_{g \in \tilde{G}} \{2\hat{\beta}^t \hat{\Sigma}_g b_g - \hat{\beta}^t \hat{\Sigma}_g \hat{\beta}\} \leq \min_{g \in \tilde{G}} \{2\hat{\beta}^t \Sigma b_g - \hat{\beta}^t \Sigma \hat{\beta}\} + 3D\kappa^2.$$

Since A is by assumption an essential subset, we have that the first term on the left-hand side is bounded with probability at least $1 - \alpha$ by

$$\min_{g \in \tilde{G}} \{2\hat{\beta}^t \hat{\Sigma}_g b_g - \hat{\beta}^t \hat{\Sigma}_g \hat{\beta}\} \leq \min_{b \in F} V_{\hat{\beta};b} + 3D\kappa^2.$$

Thus, for all ξ with $\|\xi\|_1 \leq \kappa$,

$$\min_{b \in F} V_{\hat{\beta};b} + 4\kappa \|\delta_g\|_\infty + 3D\kappa^2 \geq \min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i \xi)(X_i B_i) - (X_i \xi)^2\big).$$

Since b_{maximin} is in the feasible region, we can use it for ξ to get

(34)
$$\min_{b \in F} V_{\hat{\beta};b} + 4\kappa \|\delta_g\|_\infty + 3D\kappa^2$$
$$\geq \min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i b_{\text{maximin}})(X_i B_i) - (X_i b_{\text{maximin}})^2\big).$$

Now, by definition of b_{maximin}, when letting H be the convex hull of F (where F is again the support of F_B),

$$\min_{b \in F} E\big(2(X_i b_{\text{maximin}})(X_i b) - (X_i b_{\text{maximin}})^2\big)$$
$$= \min_{b \in H} E\big(2(X_i b_{\text{maximin}})(X_i b) - (X_i b_{\text{maximin}})^2\big)$$
$$= E\big((X_i b_{\text{maximin}})^2\big) = b_{\text{maximin}}^t \Sigma b_{\text{maximin}} = V^*,$$

where we have used in the first equality linearity with respect to the argument $b \in F$ and in the second the definition of b_{maximin} and the fact that b_{maximin} is in the convex hull of the support F of F_B.

Now bounding the fluctuations on the right-hand side of (34), we use that $|(X_i \xi)(X_i b - X_i \xi)| \leq \|X\|_\infty^2 (\max_{b \in F} \|b\|_1 + \|\xi\|_1)\|\xi\|_1 \leq 2\kappa^2$. Using Hoeffding's inequality and a union bound over all groups, for any $\alpha \in (0, 1)$, if $\kappa \geq \max_g \|b_g\|_1$ and $\kappa \geq \|\xi\|_1$,

$$P\bigg(\min_g \frac{1}{m} \sum_{i \in I_g} \big(2(X_i \xi)(X_i B_i) - (X_i \xi)^2\big) \geq b_{\text{maximin}}^t \Sigma b_{\text{maximin}} - \frac{\sqrt{2\log(G/\alpha)}}{\sqrt{m}} \kappa^2\bigg)$$
$$\geq 1 - \alpha.$$

Plugging this into (34), it holds with probability $1 - 2\alpha$ for all ξ with $\|\xi\|_1 \leq \kappa$ that

(35)
$$\min_{b \in F} V_{\hat{\beta};b} \geq V^* - \frac{\sqrt{2\log(G/\alpha)}}{\sqrt{m}} \kappa^2 - 4\kappa \|\delta_g\|_\infty - 3D\kappa^2,$$

which shows the first part of the claim in (14) if we use a union bound to exclude the event which does not correspond to the Pareto condition, and hence this excluded event has corresponding probability at most γ. The value of D can be bounded with the help of (13) with probability $1 - \alpha$ to yield

$$\min_{b \in F} V_{\hat{\beta};b} \geq V^* - 6\frac{\sqrt{2\log(2p^2 G/\alpha)}}{\sqrt{m}} \kappa^2 - 4\|\delta_g\|_\infty \kappa,$$

and the latter bound will then hold true with probability at least $1 - 3\alpha - \gamma$. The second part in (14) follows as (35) implies

$$(36) \qquad 2\hat{\beta}^t \Sigma b_{\text{maximin}} - \hat{\beta}^t \Sigma \hat{\beta} \geq V^* - 6 \frac{\sqrt{2 \log(2p^2 G/\alpha)}}{\sqrt{m}} \kappa^2 - 4 \|\delta_g\|_\infty \kappa.$$

The claim follows by $V^* = b_{\text{maximin}}^t \Sigma b_{\text{maximin}}$ and rearranging terms.

REFERENCES

AITKIN, M. and RUBIN, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **47** 67–75.

AUDIBERT, J.-Y., BUBECK, S. and LUGOSI, G. (2014). Regret in online combinatorial optimization. *Math. Oper. Res.* **39** 31–45. MR3173002

AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.

BACH, F. R. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In *Proceedings of the* 25*th International Conference on Machine Learning* 33–40. ACM, New York.

BARTLETT, P. L., DANI, V., HAYES, T., KAKADE, S., RAKHLIN, A. and TEWARI, A. (2008). High-probability regret bounds for bandit online linear optimization. Technical report, Univ. California, Berkeley, CA.

BRADIC, J. (2013). Efficient support recovery via weighted maximum-contrast subagging. Preprint. Available at arXiv:1306.3494.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

CAI, Z., FAN, J. and LI, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446

CARLSTEIN, E. G., MÜLLER, H. G. and SIEGMUND, D. (1994). *Change-Point Problems*. IMS, Hayward, CA. MR1477909

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction*, *Learning*, *and Games*. Cambridge Univ. Press, Cambridge. MR2409394

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43** 129–159. MR1854649

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

ELDAR, Y. C., BEN-TAL, A. and NEMIROVSKI, A. (2004). Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *IEEE Trans. Signal Process.* **52** 2177–2188. MR2085579

FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

FIGUEIREDO, M. A. T. and JAIN, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 381–396.

FOSTER, D. P. and VOHRA, R. (1999). Regret in the on-line decision problem. *Games Econom. Behav.* **29** 7–35. MR1729308

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R Package Version 1.

GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143. MR2956354

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. Wiley, New York. MR0829458

HAND, D. J. (2006). Classifier technology and the illusion of progress. *Statist*. *Sci*. **21** 1–34. MR2275965

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **55** 757–796. MR1229881

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann*. *Math*. *Stat*. **35** 73–101. MR0161415

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann*. *Statist*. **1** 799–821. MR0356373

KOGAN, S., LEVIN, D., ROUTLEDGE, B. R., SAGI, J. S. and SMITH, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies*: *The* 2009 *Annual Conference of the North American Chapter of the Association for Computational Linguistics* 272–280. ACM, New York.

KOLAR, M., SONG, L., AHMED, A. and XING, E. P. (2010). Estimating time-varying networks. *Ann*. *Appl*. *Stat*. **4** 94–123. MR2758086

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math*. **6** 4–22. MR0776826

LECUÉ, G. and MENDELSON, S. (2015). Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.* To appear.

MCCULLOCH, C. E. (2006). *Generalized Linear Mixed Models*. Wiley, New York.

MCCULLOCH, C. E. and NEUHAUS, J. (2001). *Generalized Linear Mixed Models*. Wiley, New York.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **72** 417–473. MR2758523

PINHEIRO, J. C. and BATES, D. M. (2000). *Linear Mixed-Effects Models*: *Basic Concepts and Examples*. Springer, Berlin.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res*. **11** 2241–2259. MR2719855

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **58** 267–288. MR1379242

WALD, A. (1945). Statistical decision functions which minimize the maximum risk. *Ann*. *of Math*. (2) **46** 265–280. MR0012402

ZINKEVICH, M., JOHANSON, M., BOWLING, M. and PICCIONE, C. (2007). Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems* 1729–1736. NIPS.

SEMINAR FÜR STATISTIK
ETH ZÜRICH
RÄMISTRASSE 101
CH-8092 ZÜRICH
SWITZERLAND
E-MAIL: meinshausen@stat.math.ethz.ch
        buhlmann@stat.math.ethz.ch