INVITED PAPER

Magging: Maximin Aggregation for Inhomogeneous Large-Scale Data

In this paper, the authors show how maximum aggregation can address certain challenges in large-scale data analysis of inhomogeneous data.

By Peter Bühlmann and Nicolai Meinshausen

ABSTRACT | Large-scale data analysis poses both statistical and computational problems which need to be addressed simultaneously. A solution is often straightforward if the data are homogeneous: one can use classical ideas of subsampling and mean aggregation to get a computationally efficient solution with acceptable statistical accuracy, where the aggregation step simply averages the results obtained on distinct subsets of the data. However, if the data exhibit inhomogeneities (and typically they do), the same approach will be inadequate, as it will be unduly influenced by effects that are not persistent across all the data due to, for example, outliers or time-varying effects. We show that a tweak to the aggregation step can produce an estimator of effects which are common to all data, and hence interesting for interpretation and often leading to better prediction than pooled effects.

KEYWORDS | Big data; distributed computing; heterogeneity; high-dimensional statistics; linear model; mixture model; statistical robustness

I. INTRODUCTION

"Big data" often refers to a large collection of observations and the associated computational issues in processing the data. Some of the new challenges from a statistical perspective include:

1) the analysis has to be computationally efficient while retaining statistical efficiency [6];
2) the data are "dirty": they contain outliers, shifting distributions, unbalanced designs, to mention a few.

There is also often the problem of dealing with data in real time, which we add to the (broadly interpreted) first challenge of computational efficiency [8].

We believe that many large-scale data are inherently inhomogeneous: that is, they are neither independent identically distributed (i.i.d.) nor stationary observations from a distribution. Standard statistical models (e.g., linear or generalized linear models for regression or classification, Gaussian graphical models) fail to capture the inhomogeneity structure in the data. By ignoring it, prediction performance can become very poor and interpretation of model parameters might be completely wrong. Statistical approaches for dealing with inhomogeneous data include mixed effect models [11], mixture models [9], and clusterwise regression models [7]: while they are certainly valuable in their own right, they are typically computationally very cumbersome for large-scale data. We present here a framework and methodology which addresses the issue of inhomogeneous data while still being vastly more efficient to compute than fitting much more complicated models such as the ones mentioned above.

A. Subsampling and Aggregation

If we ignore the inhomogeneous part of the data for a moment, a simple approach to address the computational burden with large-scale data is based on (random) subsampling: construct groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$ with $\mathcal{G}_g \subset \{1, \ldots, n\}$, where n denotes the sample size and $\{1, \ldots, n\}$ is the index set for the samples. The groups might be overlapping (i.e., $\mathcal{G}_g \cap \mathcal{G}_{g'} \neq \emptyset$ for $g \neq g'$) and do not necessarily cover the index space of samples $\{1, \ldots, n\}$. For every group $\mathcal{G}_g$, we compute an estimator (the output of an algorithm) $\hat{\theta}_g$ and these estimates are then aggregated to a single "overall" estimate $\hat{\theta}_{\text{aggr}}$, which can be achieved in different ways.

If we divide the data into G groups of approximately equal size and the computational complexity of the estimator scales

Manuscript received September 10, 2014; revised April 23, 2015; accepted July 18, 2015. Date of publication December 7, 2015; date of current version December 18, 2015.

The authors are with the Seminar für Statistik, ETH Zürich, Zurich 8092, Switzerland (e-mail: peter.buehlmann@stat.math.ethz.ch).

Digital Object Identifier: 10.1109/JPROC.2015.2494161

for n samples like n^{α} for some $\alpha > 1$, then the subsampling-based approach above will typically yield a computational complexity which is a factor $G^{\alpha-1}$ faster than computing the estimator on all data, while often just incurring an insubstantial increase in statistical error. In addition, and importantly, effective parallel distributed computing is very easy to do and such subsampling-based algorithms are well suited for computation with large-scale data.

Subsampling and aggregation can thus partially address the first challenge about feasible computation but fails for the second challenge about proper estimation and inference in presence of inhomogeneous data. We will show that a tweak to the aggregation step, which we call "maximin aggregation," can often deal also with the second challenge by focusing on effects that are common to all data (and not just mere outliers or time-varying effects).

B. Bagging: Aggregation by Averaging

In the context of homogeneous data, Breiman [1] showed good prediction performance in connection with mean or majority voting aggregation and tree algorithms for regression or classification, respectively. Bagging simply averages the individual estimators or predictions.

C. Stacking and Convex Aggregation

Again in the context of homogeneous data, the following approaches have been advocated. Instead of assigning a uniform weight to each individual estimator as in bagging, Wolpert [14] and Breiman [2] proposed to learn the optimal weights by optimizing on a new set of data. Convex aggregation for regression has been studied in [5] and has been proved to lead to approximately equally good performance as the best member of the initial ensemble of estimators. But in fact, in practice, bagging and stacking can exceed the best single estimator in the ensemble if the data are homogeneous.

D. Magging: Convex Maximin Aggregation

With inhomogeneous data, and in contrast to data being i.i.d. or stationary realizations from a distribution, the above schemes can be misleading as they give all data points equal weight and can easily be misled by strong effects which are present in only small parts of the data and absent for all other data. We show that a different type of aggregation can still lead to consistent estimation of the effects which are common in all heterogeneous data, the so-called maximin effects [10]. The maximin aggregation, which we call magging, is very simple and general and can easily be implemented for large-scale data.

II. AGGREGATION FOR REGRESSION ESTIMATORS

We now give some more details for the various aggregation schemes in the context of linear regression models with an $n \times p$ predictor (design) matrix X, whose rows correspond to n samples of the p-dimensional predictor variable, and with the n-dimensional response vector $Y \in \mathbb{R}^n$; at this point, we do not assume a true p-dimensional regression parameter; see also the model in (2). Suppose we have an ensemble of regression coefficient estimates $\hat{\theta}_g \in \mathbb{R}^p$ $(g = 1, \ldots, G)$, where each estimate has been obtained from the data in group $\mathcal{G}_g$, possibly in a computationally distributed fashion. The goal is to aggregate these estimators into a single estimator $\hat{\theta}_{\text{aggr}}$.

A. Mean Aggregation and Bagging

Bagging [1] simply averages the ensemble members with equal weight to get the aggregated estimator

$$\text{mean aggregation:} \quad \hat{\theta}_{\text{aggr}} := \sum_{g=1}^{G} w_g \hat{\theta}_g, \qquad \text{where } w_g = \frac{1}{G}, \quad \text{for all } g = 1, \ldots, G.$$

One could equally average the predictions $X\hat{\theta}_g$ to obtain the predictions $X\hat{\theta}_{\text{aggr}}$. The advantage of bagging is the simplicity of the procedure, its variance reduction property [4], and the fact that it is not making use of the data, which allows simple evaluation of its performance. The term "bagging" stands for **B**ootstrap **agg**regat**ing** (mean aggregation) where the ensemble members $\hat{\theta}_g$ are fitted on bootstrap samples of the data, that is, the groups $\mathcal{G}_g$ are sampled with replacement from the whole data.

B. Stacking

Wolpert [14] and Breiman [2] propose the idea of "stacking" estimators. The general idea is in our context as follows. Let $\hat{Y}(g) = X\hat{\theta}_g \in \mathbb{R}^n$ be the prediction of the gth member in the ensemble. Then, the stacked estimator is found as

$$\text{stacked aggregation:} \quad \hat{\theta}_{\text{aggr}} := \sum_{g=1}^{G} w_g \hat{\theta}_g, \qquad \text{where } w := \arg\min_{w \in W} \left\| Y - \sum_g \hat{Y}(g) w_g \right\|_2$$

where the space of possible weight vectors is typically of one of the following forms:

$$\text{(ridge constraint): } W = \left\{ w : \|w\|_2 \le s \right\} \text{ for some } s > 0$$
$$\text{(sign constraint): } W = \{ w : \min_g w_g \ge 0 \}$$
$$\text{(convex constraint):} W = \left\{ w : \min_g w_g \ge 0 \text{ and } \sum_g w_g = 1 \right\}.$$

If the ensemble of initial estimators $\hat{\theta}_g$ $(g = 1, \ldots, G)$ is derived from an independent data set, the framework of stacked regression has also been analyzed in [5]. Typically, though, the groups on which the ensemble members are derived use the same underlying data set as the aggregation. Then, the predictions $\hat{Y}(g)$ are for each sample point $i = 1, \ldots, n$ defined as being generated with $\hat{\theta}_g^{(-i)}$, which is the same estimator as $\hat{\theta}_g$ with observation i left out of group $\mathcal{G}_g$ (and consequently $\hat{\theta}_g^{(-i)} = \hat{\theta}_g$ if $i \notin \mathcal{G}_g$). Instead of a leave-one-out procedure, one could also use other leave-one-out schemes, such as, e.g., the out-of-bag method [3]. To this end, we just average for a given sample over all estimators that did not use this sample point in their construction, effectively setting $\hat{\theta}_g^{(-i)} \equiv 0$ if $i \in \mathcal{G}_g$. The idea of "stacking" is thus to find the optimal linear or convex combination of all ensemble members. The optimization is G-dimensional and is a quadratic programming problem with linear inequality constraints, which can be solved efficiently with a general-purpose quadratic programming solver. Note that only the inner products $\hat{Y}(g)^t \hat{Y}_{g'}$ and $\hat{Y}(g)^t Y$ for $g, g' \in \{1, \ldots, G\}$ are necessary for the optimization.

Whether stacking or simple mean averaging as in bagging provides superior performance depends on a range of factors. Mean averaging, as in bagging, certainly has an advantage in terms of simplicity. Both schemes are, however, questionable when the data are inhomogeneous. It is then not evident why the estimators should carry equal aggregation weight (as in bagging) or why the fit should be assessed by weighing each observation identically in the squared error loss sense (as in stacked aggregation).

C. Magging: Maximin Aggregation for Heterogeneous Data

We propose here **M**aximin **agg**regat**ing**, called magging, for heterogeneous data: the concept of maximin estimation has been proposed in [10], and we present a connection in Section III. The differences and similarities to mean and stacked aggregation are:

1) the aggregation is a weighted average of the ensemble members (as in both stacked aggregation and bagging);
2) the weights are nonuniform in general (as in stacked aggregation);
3) the weights do not depend on the response Y (as in bagging).

The last property makes the scheme almost as simple as mean aggregation as we do not have to develop elaborate leave-one-out schemes for estimation (as in, e.g., stacked regression). Magging is choosing the weights as a convex combination to minimize the ℓ_2-norm of the fitted values

$$\text{magging:}\quad \hat{\theta}_{\text{aggr}} := \sum_{g=1}^{G} w_g \hat{\theta}_g,$$

$$\text{where } w := \arg\min\nolimits_{w \in C_G} \left\| \sum_g \hat{Y}(g) w_g \right\|_2,$$

$$\text{and } C_G := \left\{ w\colon \min_g w_g \geq 0 \text{ and } \sum_g w_g = 1 \right\}. \tag{1}$$

If the solution is not unique, we take the solution with the lowest ℓ_2-norm of the weight vector among all solutions.

The optimization and computation can be implemented in a very efficient way. The estimators $\hat{\theta}_g$ are computed in each group of data $\mathcal{G}_g$ separately, and this task can be easily performed in parallel. In the end, the estimators only need to be combined by calculating optimal convex weights in G-dimensional space (where typically $G \ll n$ and $G \ll p$) with quadratic programming; some pseudocode in R [12] for these convex weights is presented in the Appendix. Computation of magging is thus computationally fast and simple. Furthermore, magging is very generic (e.g., one can choose its own favored regression estimator $\hat{\theta}_g$ for the gth group) and also straightforward to use in more general settings beyond linear models.

The magging scheme will be motivated in Section III with a model for inhomogeneous data, and it will be shown that it corresponds to maximizing the minimally "explained variance" among all data groups. The main idea is that if an effect is common across all groups $\mathcal{G}_g$ ($g = 1, \ldots, G$), then we cannot "average it away" by searching for a specific convex combination of the weights. The common effects will be present in all groups and will thus be retained even after the minimization of the aggregation scheme.

The construction of the groups $\mathcal{G}_g$ ($g = 1, \ldots, G$) for magging in presence of inhomogeneous data is rather specific and described in Section III-D1 for various scenarios. There, Examples 1 and 2 represent the setting where the data within each group are (approximately) homogeneous, whereas Example 3 is a case with randomly subsampled groups, despite the fact of inhomogeneity in the data.

1) Relation to the Maximin Effect Estimator: Magging is related to the maximin effect estimator proposed in [10]. Both of them aim to estimate the maximin effect parameter which is explained in Section III. Magging though is much more general than the maximin effect estimator. The former is a general aggregation technique which is easy to implement: in particular, when aggregating predictions with weights

$$w = \arg\min\nolimits_{w \in C_G} \left\| \sum_g \hat{Y}(g) w_g \right\|_2$$

it can be used with any nonlinear base learner for $\hat{Y}(g)$. On the other hand, the maximin effect estimator and its efficient computation is specifically designed for linear models with ℓ_1- or ℓ_2-norm regularization.

III. INHOMOGENEOUS DATA AND MAXIMIN EFFECTS

We motivate in the following why magging (maximin aggregation) can be useful for inhomogeneous data when the interest is in effects that are present in all groups of data.

In the linear model setting, we consider the framework of a mixture model

$$Y_i = X_i^t B_i + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2}$$

where Y_i is a univariate response variable, X_i is a p-dimensional covariable, B_i is a p-dimensional regression parameter, and ε_i is a stochastic noise term with mean zero and which is independent of the (fixed or random) covariable. Every sample point i is allowed to have its own and different regression parameter; hence, the inhomogeneity occurs because of changing parameter vectors, and we have a mixture model where, in principle, every sample arises from a different mixture component. The model in (2) is often too general: we make the assumption that the regression parameters $B_1, \ldots, B_n$ are realizations from a distribution F_B

$$B_i \sim F_B, \qquad i = 1, \ldots, n \tag{3}$$

where B_i's do not need to be independent of each other. However, we assume that B_i's are independent from X_i's and ε_i's.

Example 1 (Known Groups): Consider the case where there are known groups $\mathcal{G}_g$ with $B_i \equiv b_g$ for all $i \in \mathcal{G}_g$. Thus, this is a clusterwise regression problem (with known clusters) where every group $\mathcal{G}_g$ has the same (unknown) regression parameter vector b_g. We note that the groups $\mathcal{G}_g$ are the ones for constructing the magging estimator described in the previous section.

Example 2 (Smoothness Structure): Consider the situation where there is a smoothly changing behavior of B_i's with respect to the sample indices i: this can be achieved by positive correlation among B_i's. In practice, the sample index often corresponds to time. There are no true (unknown) groups in this setting.

Example 3 (Unknown Groups): This is the same setting as in Example 1 but the groups $\mathcal{G}_g$ are unknown. From an estimation point of view, there is a substantial difference from Example 1 [10].

A. Maximin Effects

In model (2) and in Examples 1–3, we have a "multitude" of regression parameters. We aim for a single p-dimensional parameter, which contains the common components among all B_i's (and essentially sets the noncommon components to the value zero). This can be done by the idea of so-called maximin effects which we explain next.

Consider a linear model with the fixed p-dimensional regression parameter b which can take values in the support of F_B from (3)

$$Y_i = X_i^t b + \varepsilon_i, \qquad i = 1, \ldots, n \tag{4}$$

where X_i and ε_i are as in (2) and assumed to be i.i.d. We will connect the random variables B_i in (2) to the values b via a worst case analysis as described below: for that purpose, the parameter b is assumed to not depend on the sample index i. The variance which is explained by choosing a parameter vector β in the linear model (4) is

$$V_{\beta,b} := \mathbb{E}|Y|^2 - \mathbb{E}|Y - X^t\beta|^2 = 2\beta^t\Sigma b - \beta^t\Sigma\beta$$

where Σ denotes the covariance matrix of X. We aim for maximizing the explained variance in the worst (most adversarial) scenario; this is the definition of the maximin effects.

Definition [10]: The maximin effect parameter is

$$b_{\text{maximin}} = \arg\min_\beta \max_{b \in \text{supp}(F_B)} -V_{\beta,b}$$

and note that the definition uses the negative explained variance $-V_{\beta,b}$.

The maximin effects can be interpreted as an aggregation among the support points of F_B to a single parameter vector, i.e., among all B_i's (e.g., in Example 2) or among all the clustered values b_g (e.g., in Examples 1 and 3); see also Fact 1. The maximin effect parameter is different from the pooled effects $b_{\text{pool}} = \arg\min_\beta \ \mathbb{E}_B[-V_{\beta,B}]$ and, a bit surprisingly, also rather different from the prediction analog

$$b_{\text{pred-maximin}} = \arg\min_\beta \max_{b \in \text{supp}(F_B)} \mathbb{E}\left[(X^t b - X^t\beta)^2\right].$$

In particular, the value zero has a special status for the maximin effect parameter b_{maximin}, unlike for $b_{\text{pred-maximin}}$ or b_{pool}; see [10]. The following is an important "geometric" characterization which indicates the special status of the value zero; see also Fig. 1.

Fact 1 [10]: Let H be the convex hull of the support of F_B. Then

$$b_{\text{maximin}} = \arg\min_{\gamma \in H} \gamma^t\Sigma\gamma.$$

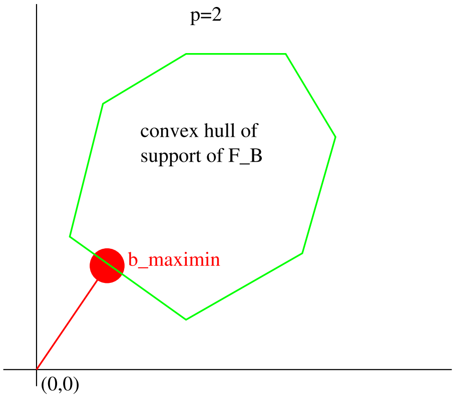

Fig. 1. ***Illustration of Fact 1 in dimension $p = 2$.***

That is, the maximin effect parameter b_{maximin} is the point in the convex hull H which is closest to zero with respect to the distance $d(u,v) = (u-v)^t\Sigma(u-v)$; in particular, if the value zero is in H, the maximin effect parameter equals $b_{\text{maximin}} \equiv 0$.

The characterization in Fact 1 leads to an interesting robustness issue which we will discuss in Section III-B.

The connection to magging (maximin aggregation) can be made most easily for the setting of Example 1 with known groups and constant regression parameter b_g within each group $\mathcal{G}_g$. We can rewrite, using Fact 1

$$\begin{aligned} b_{\text{maximin}} &= \sum_{g=1}^{G} w_g^0 b_g \\ w^0 &= \left(w_1^0, \ldots, w_G^0\right) \\ &= \arg\min\nolimits_{w\in C_G} \sum_{g,g'=1}^{G} w_g w_{g'} b_g^T \Sigma b_g \\ &= \arg\min\nolimits_{w\in C_G} \mathbb{E}_X \left\| \sum_{g=1}^{G} w_g X b_g \right\|_2^2 \end{aligned}$$

where C_G is as in (1). The magging estimator is then using the plug-in principle with estimates $\hat{\theta}_g$ for b_g and $\|\sum_g w_g \hat{Y}(g)\|_2^2$ for $\mathbb{E}_X \|\sum_{g=1}^G w_g X b_g\|_2^2$.

B. Robustness

It is instructive to see how the maximin effect parameter is changing if the support of F_B is extended, possibly rendering the support nonfinite. There are two possibilities, illustrated in Fig. 2. In the first case, illustrated in the left panel of Fig. 2, the new parameter vector b_{new} is not changing the point in the convex hull of the support of F_B that is closest to the origin. The maximin effect parameter is then unchanged. The second situation is illustrated in the right panel of Fig. 2. The addition of a new support point here does change the convex hull of the support such that there is now a point in the support closer to the origin. Consequently, the maximin effect parameter will shift to this new value. The maximin effect parameter thus is either unchanged or is moving closer to the origin. Therefore, maximin effect parameters and their estimation exhibit an excellent robustness feature with respect to breakdown properties.

The question might occur whether the maximin effects are then very conservative in nature and in fact coincide often with the origin. We will show that this is not the case for high-dimensional spaces. Note that the maximin effects do not coincide with the origin if all $b_g, g = 1, \ldots, G$, lie on one hemisphere, that is, there exists a direction $\gamma \in \mathbb{R}^p$ such that $\gamma^t b_g > 0$ for all $g \in 1, \ldots, G$. As an extreme scenario, we can imagine b_g being sampled independently uniform on the sphere. Even in this scenario, the probability that the maximin effects do not vanish is given by Wendel's theorem [13] as

$$2^{-G+1} \sum_{k=0}^{p-1} \binom{G-1}{k}$$

which is for $G = p+1$ (that is one group more than there are dimensions) equal to

$$1 - 2^{-p}$$

even higher for smaller number G of groups, and for $G = 2p$, we obtain the probability 0.5. The chance that the maximin effects vanish identically is thus very small in higher dimensional spaces, even if the groups are sampled centro-symmetric around the origin for each group, as long as the number of groups G is not exceeding the number of dimensions by a factor larger than one.

C. Maximin Effects Under Time-Varying Effects

To illustrate the maximin effects with a small example and compare with pooled estimation, we can use a simple model, where one part of the signal is constant and another part of the signal is changing in strength over time. Specifically, assume that one direction v has a constant effect strength, while the effect in a direction w has a linear trend $b_i = v + w[\alpha_0 + (i-1)(\alpha_1 - \alpha_0)/(n-1)]$ for the training data $i = 1, \ldots, n$, where v and w are two directions that are orthogonal with respect to Σ. The pooled population effects over the training data are $b_{\text{pool}} = v + w\alpha_{\text{pool}}$, where $\alpha_{\text{pool}} = (\alpha_0 + \alpha_1)/2$. Consider G groups of consecutive observations. The population maximin effects are given by $b_{\text{maximin}} = v + w\alpha_{\text{maximin}}$, where $\alpha_{\text{maximin}} = \arg\min_{x\in[\alpha_{\min},\alpha_{\max}]} |x|$, using Fact 1, where $[\alpha_{\min}, \alpha_{\max}]$ is the range of the population effects in the direction of w among all the groups $g = 1, \ldots, G$.

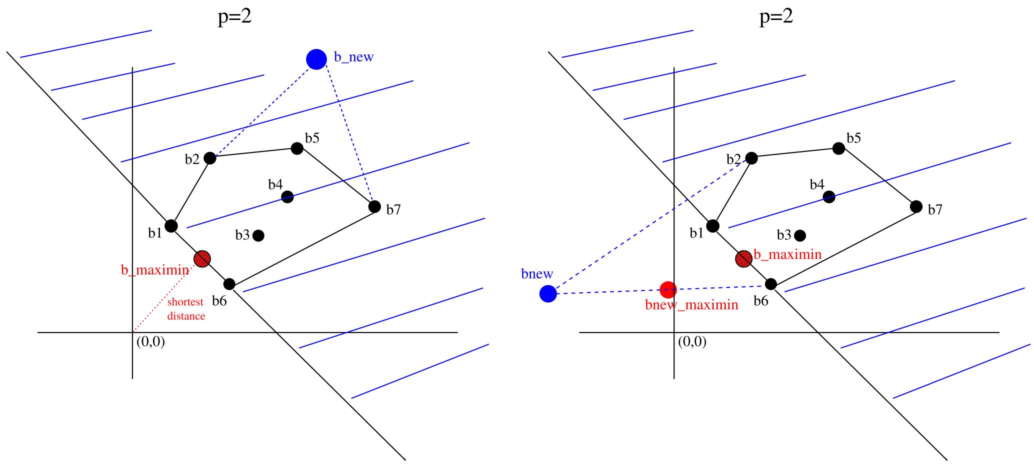

Fig. 2. ***Illustration of the case with a finite number of possible values for B. (Left) The values $b_1, \ldots, b_7$ are possible realizations of B_i, and b_{maximin} is the closest point to zero in the convex hull of $\{b_1, \ldots, b_7\}$ (in black). When adding a new additional realization b_{new}, the convex hull becomes larger (in dashed blue). As long as the new support point is in the blue shaded half-space, the maximin effect parameter b_{maximin} remains the same regardless of how far away the new support point is added. (Right) A new additional realization b_{new} arises which does not lie in the blue shaded half-space, the convex hull becomes larger (in dashed blue), and the new maximin effect parameter becomes $b_{\text{new,maximin}}$. Since the new convex hull (in dashed blue) gets enlarged by a new realized value b_{new}, the corresponding new maximin effect parameter $b_{\text{new,maximin}}$ must be closer to the origin than the original parameter b_{maximin}. Thus, it is impossible to shift b_{maximin} away from zero by placing new realizations at arbitrary positions.***

In the limit of a large number G of groups, we have $\alpha_{\min} = \min\{\alpha_0, \alpha_1\}$ and $\alpha_{\max} = \max\{\alpha_0, \alpha_1\}$. If the linear trend in the direction w continues to hold and we calculate the test mean squared error over the samples $i = n+1, n+2, \ldots$, we have

$$E_{\text{test}}\left(\|Y - Xb_{\text{maximin}}\|_2^2\right) \leq E_{\text{test}}\left(\|Y - Xb_{\text{pool}}\|_2^2\right)$$

if and only if $|\alpha_1| \leq |\alpha_0|$, that is, if the effect strength at the end of the training data is weaker than at the start of the training phase (intuitively as the effects in direction of w will then continue to weaken on the test data or have already changed sign compared with the start of the training phase). The inequality is strict if $|\alpha_1| < |\alpha_0|$ and the difference between the test set errors becomes large with increasing value of $|\alpha_0| - |\alpha_1|$, meaning an increasing amount of heterogeneity. Moreover, looking at the explained variance, let $F_{\text{trend}} = \{\beta : \beta = v + w\tau \text{ with } \tau \in \mathbb{R}\}$ be the set of all coefficients where direction v is constant and direction w variable. It is then easy to show that for all α_0, α_1 and number G of groups

$$\min_{b \in F_{\text{trend}}} E(V_{b, b_{\text{maximin}}}) \geq \min_{b \in F_{\text{trend}}} E(V_{b, b_{\text{pool}}})$$

that is, the worst case explained variance over F_{trend} (which is a larger set than the one seen for training) is always better under the maximin effects than under the pooled effects. Magging will, in general, shrink a direction more if the signal is varying strongly in this direction, while it will retain all directions where the effect is stable. Of course, one price to pay for this is increased estimation noise compared to the pooled estimator as magging operates in each group on a smaller sample size, which we will quantify in the following.

D. Statistical Properties of Magging

We will derive now some statistical properties of magging, the maximin aggregation scheme, proposed in (1). They depend also on the setting-specific construction of the groups $\mathcal{G}_1, \ldots \mathcal{G}_G$ which is described in Section III-D1.

Assumptions: Consider the model (2) and that there are G groups $\mathcal{G}_g$ $(g = 1, \ldots, G)$ of data samples. Denote by Y_g and X_g the data values corresponding to group $\mathcal{G}_g$.

1) Let b_g^* be the optimal regression vector in each group, that is $b_g^* = \mathbb{E}_B[|\mathcal{G}_g|^{-1} \sum_{i \in \mathcal{G}_g} B_i]$. Assume that b_{maximin} is in the convex hull of $\{b_1^*, \ldots, b_G^*\}$.
2) We assume random design with a mean-zero random predictor variable X with covariance matrix Σ and let $\hat{\Sigma}_g = |\mathcal{G}_g|^{-1} X_g^t X_g$ be the empirical Gram matrices. Let $\hat{\theta}_g$ $(g = 1, \ldots, G)$ be the estimates in each group. Assume that there exists some $\eta_1, \eta_2 > 0$ such that

$$\max_g \left(\hat{\theta}_g - b_g^*\right)^t \Sigma \left(\hat{\theta}_g - b_g^*\right) \leq \eta_1$$

$$\max_g \|\hat{\Sigma}_g - \Sigma\|_\infty \leq \eta_2$$

where $m = \min_g |\mathcal{G}_g|$ is the minimal sample size across all groups.

3) The optimal and estimated vectors are sparse in the sense that there exists some $\kappa > 0$ such that

$$\max_g \left\| b_g^* \right\|_1 \leq \kappa \quad \text{and} \quad \max_g \|\hat{\theta}_g\|_1 \leq \kappa.$$

Assumption A1) is fulfilled for known groups, where the convex hull of $\{b_1^*, \ldots, b_G^*\}$ is equal to the convex hull of the support of F_B and the maximin-vector b_{maximin} is hence contained in the former. Example 1 is fulfilling the requirement, and we will discuss generalizations to the settings in Examples 2 and 3 in Section III-D1. Assumptions A2) and A3) are relatively mild: the first part of A3) is an assumption that the underlying model is sufficiently sparse. If we consider standard Lasso estimation with sparse optimal coefficient vectors and assuming bounded predictor variables, then A2) is fulfilled with high probability for η_1 of the order $\kappa(\log(pG)/m)^{1/2}$ (faster rates are possible under a compatibility assumption) and η_2 of order $\log(pG)/m$, where $m = \min_g |\mathcal{G}_g|$ denotes the minimal sample size across all groups; see, for example, [10].

Define for $x \in \mathbb{R}^p$ the norm $\|x\|_\Sigma^2 = x^t \Sigma x$ and let $\hat{\theta}_{\text{magging}}$ be the magging estimator (1).

Theorem 1: Assume A1)–A3). Then

$$\|\hat{\theta}_{\text{magging}} - b_{\text{maximin}}\|_\Sigma^2 \leq 6\eta_1 + 4\eta_2\kappa^2.$$

A proof is given in the Appendix.

The result implies that the maximin effect parameter can be estimated with good accuracy by magging (maximin aggregation) if the individual effects in each group can be estimated accurately with standard methodology (e.g., penalized regression methods).

1) Construction of Groups and Their Validity for Different Settings: Theorem 1 hinges mainly on assumption A1). We discuss the validity of the assumption for the three discussed settings under appropriate (and setting-specific) sampling of the data groups.

Example 1 (Known Groups Continued): Obviously, the groups $\mathcal{G}_g$ $(g = 1, \ldots, G)$ are chosen to be the true known groups.

Assumption A1) is then trivially fulfilled with known groups and constant regression parameter within groups (clusterwise regression).

Example 2 (Smoothness Structure Continued): We construct G groups of nonoverlapping consecutive observations. For simplicity, we would typically use equal group size $m = \lfloor n/G \rfloor$ so that $\mathcal{G}_1 = \{1, 2, \ldots, m\}$, $\mathcal{G}_2 = \{m+1, \ldots, 2m\}, \ldots, \mathcal{G}_G = \{(G-1)m+1, \ldots, n\}$.

When taking sufficiently many groups and for a certain model of smoothness structure, condition A1) will be fulfilled with high probability [10]: it is shown there that it is rather likely to get some groups of consecutive observations where the optimal vector is approximately constant and the convex hull of these "pure" groups will be equal to the convex hull of the support of F_B.

Example 3 (Unknown Groups Continued): We construct G groups of equal size m by random subsampling: sample without replacement within a group and with replacement between groups.

This random subsampling strategy can be shown to fulfill condition A1) when assuming an additional so-called Pareto condition [10]. As an example, a model with a fraction of outliers fulfills A1) and one obtains an important robustness property of magging which is closely connected to Section III-B.

E. Numerical Example

We illustrate the difference between mean aggregation and maximin aggregation (magging) with a simple example. We are recording, several times, data in a time domain. Each recording (or group of observations) contains a common signal, a combination of two frequency components, shown in the top left of Fig. 3. On top of the common signal, seven out of a total of 100 possible frequencies (bottom left in Fig. 3) add to the recording in each group with a random phase. The 100 possible frequencies are the first frequencies $2\pi j/P, j = 1, \ldots, 100$, for periodic signal with periodicity P defined by the length of the recordings. They form the dictionary used for least squares estimation of the signal, and the magging estimate is formed in this Fourier space. In total, $G = 50$ recordings are made, of which the first 11 are shown in the second column of Fig. 3. The estimated signals are shown in the third column, removing most of the noise but leaving the random contribution from the noncommon signal in place. Averaging over all estimates in the mean sense yields little resemblance with the common effects. The same holds true if we estimate the coefficients by pooling all data into a single group (first two panels in the rightmost column of Fig. 3). Magging (maximin aggregation) and the closely related but less generic maximin estimation [10], on the other hand, approximate the common signal in all groups quite well (bottom two panels in the rightmost column of Fig. 3). The magging estimate retains substantial weight only on dictionary elements corresponding to low frequencies, showing that the common signal is to be

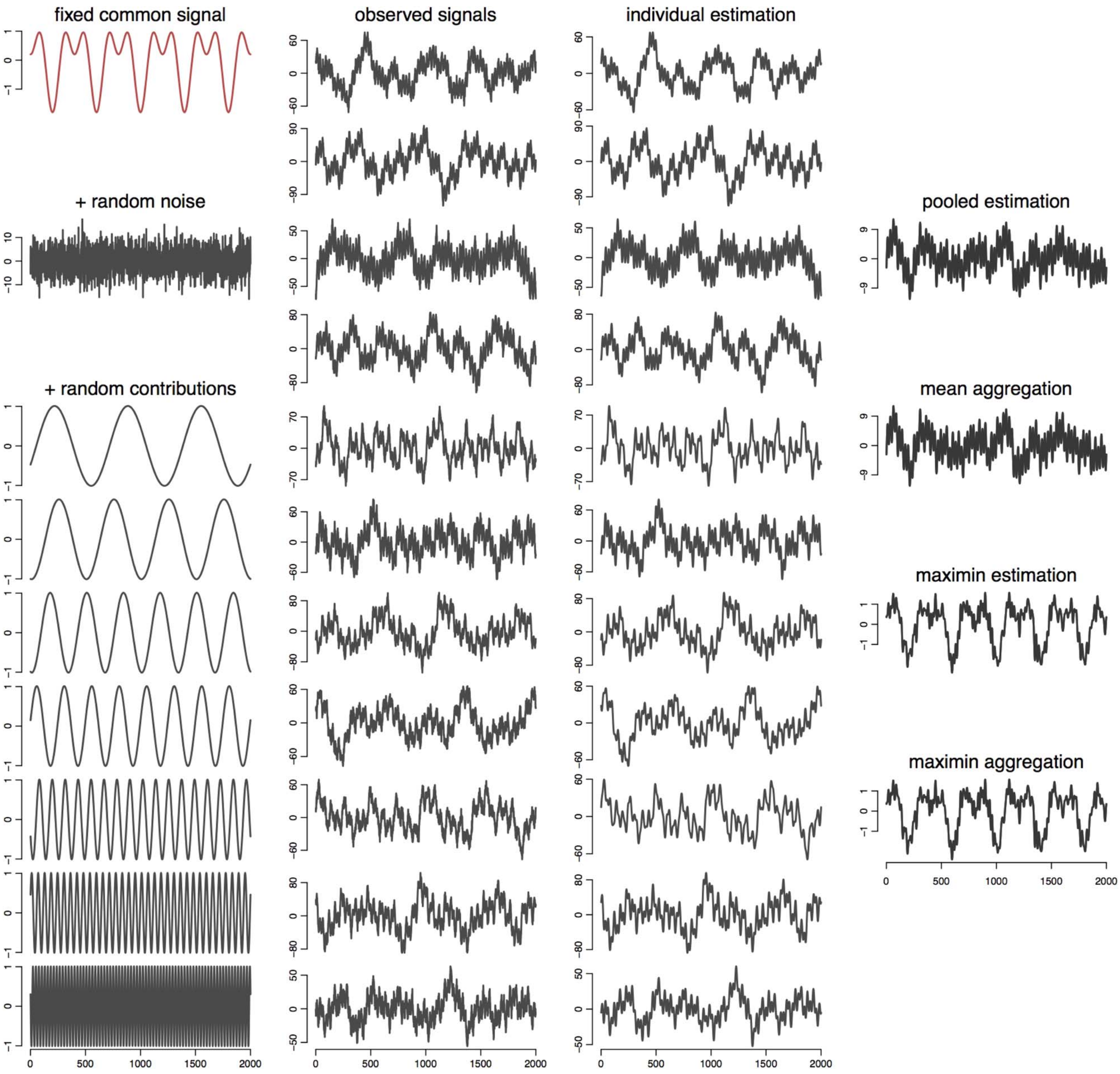

Fig. 3. *Left column shows the data generation. Each group has the same fixed common effect (shown in red at the top left), and gets random noise as well as other random periodic contributions added (with random phase), where the latter two contributions are drawn independently for all groups $g = 1, \ldots, G = 50$. The second column shows the realizations of Y_g for the first groups $g = 1, \ldots, 11$, while the third shows the least squares estimates of the signal when projecting onto the space of periodic signals in a certain frequency range. The last column shows from top to bottom: (a) the pooled estimate one obtains when adding all groups into one large data set and estimating the signal on all data simultaneously (the estimate does not match closely the common effects shown in red); (b) the mean aggregated data obtained by averaging the individual estimates (here identical to pooled estimation); (c) the (less generic) maximin effect estimator from [10]; and (d) magging: maximin aggregated estimators (1), both of which match the common effects quite closely.*

found in the low-frequency component and that higher frequencies do not show much commonality between the settings.

Meinshausen *et al.* [10] provide other real data results where maximin effect estimation leads to better out-of-sample predictions in two financial applications.

IV. CONCLUSION

Large-scale and "Big" data poses many challenges from a statistical perspective. One of them is to develop algorithms and methods that retain optimal or reasonably good statistical properties while being computationally cheap to compute. Another is to deal with inhomogeneous data which might contain outliers, shifts in distributions, and other effects that do not fall into the classical framework of identically distributed or stationary observations. Here we have shown how magging ("maximin aggregation") can be a useful approach addressing both of the two challenges. The whole task is split into several smaller data sets (groups), which can be processed trivially in parallel. The standard solution is then to average the results from all tasks, which we call "mean aggregation" here. In contrast, we show that finding a certain convex combination, we can detect the signals which are common in all subgroups of the data. While "mean aggregation" is easily confused by signals that shift over time or which are not present in all groups, magging ("maximin aggregation") eliminates as much as possible these inhomogeneous effects and just retains the common signals which is an interesting feature in its own right and often improves out-of-sample prediction performance. ■

APPENDIX

Proof of Theorem 1: Define for $w \in C_G$ [where $C_G \subset \mathbb{R}^G$ is as defined in (1) the set of positive vectors that sum to one]

$$\hat{\theta}(w) := \sum_{g=1}^{G} w_g \hat{\theta}_g \quad \text{and} \quad \theta(w) := \sum_{g=1}^{G} w_g b_g^*$$

and let for $\hat{\Sigma} = n^{-1} X^t X$

$$\hat{L}(w) := \hat{\theta}(w)^t \hat{\Sigma} \hat{\theta}(w) \quad \text{and} \quad L(w) := \theta(w)^t \Sigma \theta(w).$$

Then, $w^* = \arg\min_w L(w)$ and $b_{\text{maximin}} = \theta(w^*)$ and $\hat{w} = \arg\min_w \hat{L}(w)$ and $\hat{\theta}_{\text{magging}} = \hat{\theta}(\hat{w})$. Now, using A3)

$$\begin{aligned} \sup_{w \in C_G} \left| \hat{L}(w) - L(w) \right| &\leq \sup_{w \in C_G} \left| \theta(w)^t (\Sigma - \hat{\Sigma}) \theta(w) \right| \\ &\quad + \max_g \left\| b_g^* - \hat{b}_g \right\|_\Sigma^2 \\ &\leq \eta_2 \left(\max_{w \in C_G} \|\theta(w)\|_1 \right)^2 + \eta_1. \end{aligned}$$

Hence, as $w^* = \arg\min_{w \in C_G} L(w)$ and $\hat{w} = \arg\min_w \hat{L}(w)$

$$L(\hat{w}) \leq L(w^*) + 2(\eta_1 + \eta_2 \kappa^2). \tag{5}$$

For $\Delta := \theta(\hat{w}) - \theta(w^*)$

$$\begin{aligned} L(\hat{w}) = \|\theta(\hat{w})\|_\Sigma^2 &= (\theta(w^*) + \Delta)^t \Sigma (\theta(w^*) + \Delta) \\ &= \theta(w^*)^t \Sigma \theta(w^*) + 2\Delta^t \Sigma \theta(w^*) + \Delta^t \Sigma \Delta \\ &\geq L(w^*) + \|\Delta\|_\Sigma^2 \end{aligned}$$

where $\Delta^t \Sigma \theta(w^*) \geq 0$ follows by the definition of the maximin vector $\theta(w^*) = b_{\text{maximin}}$. Combining the last inequality with (5)

$$\|\theta(\hat{w}) - \theta(w^*)\|_\Sigma^2 \leq 2(\eta_1 + \eta_2 \kappa^2) \tag{6}$$

Furthermore, by A3)

$$\sup_{w \in C_G} \left\| \hat{\theta}(w) - \theta(w) \right\|_\Sigma^2 \leq \eta_1.$$

Using the equality for $\hat{\theta}_{\text{magging}} = \hat{\theta}(\hat{w})$

$$\left\| \hat{\theta}(\hat{w}) - \theta(\hat{w}) \right\|_\Sigma^2 \leq \eta_1. \tag{7}$$

Combining (6) and (7)

$$\begin{aligned} &\|\hat{\theta}_{\text{magging}} - b_{\text{maximin}}\|_\Sigma^2 \\ &\quad = \left\| \hat{\theta}(\hat{w}) - \theta(w^*) \right\|_\Sigma^2 \\ &\quad \leq 2\left(\left\| \hat{\theta}(\hat{w}) - \theta(\hat{w}) \right\|_\Sigma^2 + \|\theta(\hat{w}) - \theta(w^*)\|_\Sigma^2 \right) \\ &\quad \leq 2\big(\eta_1 + 2(\eta_1 + \eta_2 \kappa^2)\big) \\ &\quad = 6\eta_1 + 4\eta_2 \kappa^2 \end{aligned}$$

which completes the proof.

Implementation of Magging in R*:* At the top of the next page, we present some pseudocode for computing the weights $w_1, \ldots, w_G$ in magging (1), using quadratic programming in the R-software environment.

```
library(quadprog)
theta< cbind(theta1,...,thetaG)       #matrix with G columns :
                                      #each column is a regression estimate

hatS<-t(X) %*% X/n                    #empirical covariance matrix of X
H <- t(theta) %*% hatS %*% theta      #assume that it is positive definite
                                      #(use H+xi * I, xi > 0 small, otherwise)
A <- rbind(rep(1,G),diag(1,G))        #constraints
b < c(1,rep(0,G))
d <- rep(0,G)                         #linear term is zero
w <- solve.QP(H,d,t(A),b,meq = 1)     #quadratic programming solution to
                                      #argmin(x^t H x) such that Ax >= b and
                                      #first inequality is an equality
```

REFERENCES

[1] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[2] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, pp. 49–64, 1996.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[4] P. Büuhlmann and B. Yu, "Analyzing bagging," *Ann. Stat.*, vol. 30, pp. 927–961, 2002.

[5] B. Bunea, A. Tsybakov, and M. Wegkamp, "Aggregation for Gaussian regression," *Ann. Stat.*, vol. 35, pp. 1674–1697, 2007.

[6] V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," *Proc. Nat. Acad. Sci.*, vol. 110, pp. E1181–E1190, 2013.

[7] W. DeSarbo and W. Cron, "A maximum likelihood methodology for clusterwise linear regression," *J. Classification*, vol. 5, pp. 249–282, 1988.

[8] M. W. Mahoney, "Randomized algorithms for matrices and data," *Found. Trends Mach. Learn.*, vol. 3, pp. 123–224, 2011.

[9] G. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2004.

[10] N. Meinshausen and P. Büuhlmann, "Maximin effects in inhomogeneous large-scale data," *Ann. Stat.*, vol. 43, pp. 1801–1830, 2015.

[11] J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-PLUS*. New York, NY, USA: Springer-Verlag, 2000.

[12] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2014.

[13] J. Wendel, "A problem in geometric probability," *Mathematica Scandinavia*, vol. 11, pp. 109–111, 1962.

[14] D. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, pp. 241–259, 1992.

ABOUT THE AUTHORS

Peter Bühlmann received the Ph.D. degree in statistics from ETH Zurich, Zurich, Switzerland, in 1993.

After three years at the University of California Berkeley, Berkeley, CA, USA (1994-1997), he was appointed as Assistant and then Associate Professor, and in 2004, he became a Full Professor at the Department of Mathematics, ETH Zurich. Since 2013, he has been Chair of the Department of Mathematics, ETH Zurich. His research interests are in causal and high-dimensional statistical inference, machine learning, and applications in life sciences.

Dr. Bühlmann is a Fellow of the Institute of Mathematical Statistics and an elected Member of the International Statistical Institute. He presented a Medallion Lecture at the Joint Statistical Meetings 2009, a read paper to the Royal Statistical Society in 2010, and various named lectures. He was an Editor of the *Annals of Statistics* from 2010 to 2012.

Nicolai Meinshausen received the Ph.D. degree in statistics from ETH Zurich, Zurich, Switzerland, in 2005.

After a postdoctoral stay at the University of California Berkeley, Berkeley, CA, USA, he became a University Lecturer in 2007 and a Professor in 2012 at the University of Oxford, Oxford, U.K. In 2013, he moved to ETH Zurich as a Full Professor. His research interests are in machine learning, high-dimensional learning, causality, and applications, mostly in the physical sciences.

Dr. Meinshausen received a Leverhulme Prize in 2010, presented a read paper to the Royal Statistical Society in 2010, was awarded a Guy Medal in Bronze in 2011, and gave a Medallion Lecture at the Joint Statistical Meetings in 2015.