Inhomogeneous large-scale data: maximin effects

Nicolai Meinshausen
ETH Zürich

based on joint work with

Peter Bühlmann

Classical regression

For target $Y \in \mathbb{R}^n$ and predictor matrix $X \in \mathbb{R}^{p \times n}$,

$$Y = X\beta^* + \text{noise}$$

with n iid samples of p predictor variables and optimal fixed linear approximation $\beta^* \in \mathbb{R}^p$.
...analogous for graph estimation, classification etc.

Challenges for large-scale data analysis:
a combination of one or all of

(i) computational issues due to large n and/or p.

(ii) inhomogeneous data

(iii) ...

Challenge Ia): computational issues due to large p

With large p,

- trade bias and variance by fitting a sparse approximating model to data (optimizing statistical efficiency)
- trade computational and statistical efficiency by using convex relaxations

for regression:

$$\text{Least squares: } \operatorname{argmin}_\beta \|Y - X\beta\|_2^2$$

$$\text{Model selection: } \operatorname{argmin}_\beta \|Y - X\beta\|_2^2 \text{ such that } \|\beta\|_0 \le s$$

$$\text{Lasso: } \operatorname{argmin}_\beta \|Y - X\beta\|_2^2 \text{ such that } \|\beta\|_1 \le \tau$$

Challenge Ib): computational issues due to large n

With large n,

- trade computational efficiency and variance by retaining just a random subset of the data

loss of efficiency can be exactly controlled if data are iid. If data are iid, retaining a few thousand samples will often be “good enough”

Challenge II) inhomogeneous data

Simple iid-model:

For target $Y \in \mathbb{R}^n$ and predictor matrix $X \in \mathbb{R}^{p \times n}$,

$$Y = X\beta^* + \textit{noise}$$

with n **iid** *samples of p predictor variables and optimal* **fixed** *linear approximation $\beta^* \in \mathbb{R}^p$.*

might be very wrong!

primary concern

large-scale data is "always" inhomogeneous!

we expect
batch effects, different populations,
unwanted variation (Bartsch & Speed, 2012–current), ...

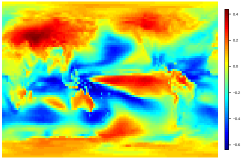
→ ignoring them can give very misleading results

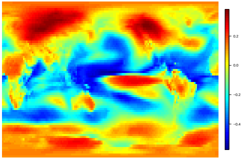
→ addressing them can be computationally very expensive/impossible

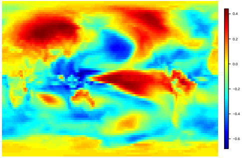
Example I: Climate models

(Knutti et al. at ETHZ)
different global circulation models produce similar (but not identical) results

ACCESS1 model *CNRM model* *IPSL-CM5A model*





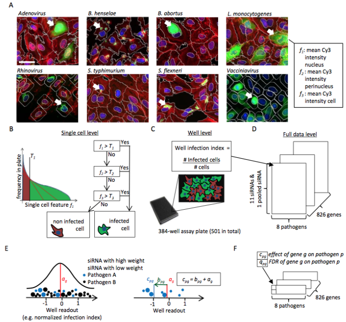


models are not idential
what are the common effects in them?

Example II: pathogen ("virus") entry into human cells

(InfectX project, ongoing (PB); Drewek, Schmich, ..., Beerenwinkel, PB, Dehio)
study 8 different viruses

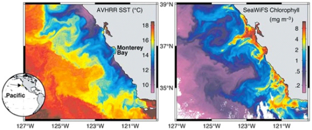
Figure 1

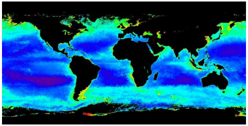


what are the "common effects" present in all 8 virus sources?

Example III: biomass models

(Gruber et al. at ETHZ)
fit 6 different models for biomass based on satellite data, simulation models, historic ground-based measurements etc.





Chlorophyll maps from different sources

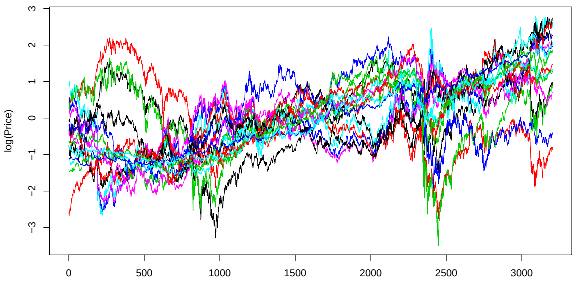
infer the common effects among all possible models that use different data sources previous examples: sources or groups are known

we will also deal with cases where the
sources/groups are unknown

e.g. when we expect
“batch effects”, “different populations”,
“unwanted variation” ,
...

Example IV: financial time-series with changepoints

Time-series operate in different regimes with (unknown) change-points



scaled log-prices of 17 financial instruments over 16 years.

- which effects stay constant over time?
- can we find the common, constant effects without having to do a full change-point analysis?

challenges

1) construction of reasonably simple models which capture potential inhomogeneities
2) computation (and memory requirements)

$\rightsquigarrow$ structure of the talk:

- model
- statistical properties
- computation

First “naive” thoughts (mainly regarding computation)

reduce computational load by subsampling

- naive subsampling by
  random subsets $\mathcal{S}_1, \ldots, \mathcal{S}_B \subset \{1, \ldots, n\}$ and
  computing model parameters $\hat{\theta}_{\mathcal{S}_b}$ for $b = 1, \ldots, B$

  $\rightsquigarrow$ trivial implementation for distributed computing!
- aggregation of estimates $\hat{\theta}_{\mathcal{S}_1}, \ldots, \hat{\theta}_{\mathcal{S}_B}$ (cf. Breiman, 1996)

gain insight about the distribution/variability of the estimated model parameters $\hat{\theta}_{\mathcal{S}_1}, \ldots, \hat{\theta}_{\mathcal{S}_B}$ under subsampling

in particular: how stable are the estimates in presence of outliers, batch effects, inhomogeneities ?

arising questions

(i) is naive subsampling a valid approach?

(ii) how should we aggregate the subsampled estimates? averaging like in Bagging and Random Forests (Breiman, 2001)?

quick answers

(i) is naive subsampling a valid approach?

$\rightarrow$ naive subsampling is good for “i.i.d. data” but usually the wrong probability mechanism for data with “inhomogeneity structure”

$\rightarrow$ naive subsampling will nevertheless be shown to be useful in connection with adequate aggregation

(ii) how should we aggregate the subsampled estimates? averaging like in Bagging and Random Forests (Breiman, 2001)?

$\rightarrow$ mean or median aggregation of $\hat{\theta}_{\mathcal{S}_1}, \ldots, \hat{\theta}_{\mathcal{S}_B}$ often “inadequate”

$\rightarrow$ “maximin” aggregation is more suitable and “robust”

“classical” approaches (to deal with inhomogeneities)

(i) robust methods (Huber, 1964; 1973)

(ii) mixed or random effects models (when groups are known)
(Pinheiro & Bates, 2000)

(iii) time-varying coefficient models
(Hastie & Tibshirani, 1993; Fan & Zhang, 1999)
shift over time (Hand, 2006)

(iv) mixture models (when groups are unknown)
(Aitkin & Rubin, 1985; McLachlan & Peel, 2004)

here we ask for

- **less** than the models above provide (just estimate the common constant effects)
- but we want it **faster** (with less computational effort).

“classical” approaches (to deal with inhomogeneities)

(i) robust methods (Huber, 1964; 1973)

(ii) mixed or random effects models (when groups are known)
(Pinheiro & Bates, 2000)

(iii) time-varying coefficient models
(Hastie & Tibshirani, 1993; Fan & Zhang, 1999)
shift over time (Hand, 2006)

(iv) mixture models (when groups are unknown)
(Aitkin & Rubin, 1985; McLachlan & Peel, 2004)

here we ask for

- **less** than the models above provide (just estimate the common constant effects)
- but we want it **faster** (with less computational effort).

mixture model – without requiring to fit such a model

linear model context:

$$Y_i = X_i^T \underbrace{B_i}_{p\times 1} + \varepsilon_i \ \ (i = 1, \ldots, n),$$
$$B_i \sim F_B$$

$\mathbb{E}[\varepsilon^T X] = 0$ (errors uncorrelated from predictors)
B_i's independent of X, ε, but not necessarily i.i.d.

Example 1 (clustered regression)
finite support of F_B with $|\mathrm{supp}(F_B)| = G$
$\rightsquigarrow$ observations from G different groups (either known or unknown) with $B_i \equiv b_g \ \forall i \in$ group g ($g = 1, \ldots, G$)

Example 2
positively correlated B_i's $\rightsquigarrow$ "smooth behavior w.r.t. index i"
e.g. time-varying coefficient model

motivation for “maximin” or “common” effects

we do not want to fit the entire mixture model because:

1) no gain for prediction (if no information in X on mixture component)
2) only want to learn the “effects which are consistent/stable” across the mixture components
3) computationally cumbersome

regarding the second point: our proposal is to maximize the explained variance under the worst adversarial scenario

explained variance

consider linear model with fixed $b \in \mathrm{supp}(F_B)$ and random design X with covariance Σ:

$$Y_i = X_i^T b + \varepsilon_i \ \ (i = 1, \ldots, n)$$

in short: $$Y = Xb + \varepsilon$$

explained variance when choosing parameter vector β:

$$V_{b,\beta} = \mathbb{E}_Y \|Y\|_2^2/n - \mathbb{E}_{Y,X}\|Y - X\beta\|_2^2/n = 2\beta^T \Sigma b - \beta^T \Sigma \beta$$

Definition: (NM & Bühlmann, 2014)
maximize explained variance under most adversarial scenario

maximin effects: $b_{\mathrm{maximin}} = \mathrm{argmax}_{\beta} \min_{b \in \mathrm{supp}(F_B)} V_{b,\beta}$

Example (clustered regression)
G groups each with its own regr. parameter b_g ($g = 1, \ldots, G$)

$$\min_{b \in \mathrm{supp}(F_B)} V_{b,\beta} = \min_{g} V_{b_g,\beta} = \min_{g} 2\beta^T \Sigma b_g - \beta^T \Sigma \beta$$

$=$ explained variance, when choosing β, in worst case (group)

in general: $\mathrm{supp}(F_B)$ does not need to be finite
(i.e. not necessarily G points from G groups)

maximin effects are

(i) very different from pooled effects:

$$b_{\text{pool}} = \text{argmax}_{\beta} \mathbb{E}_B[V_{B,\beta}]$$

best "on average over $B \sim F_B$"

(ii) somewhat different from corresponding prediction

$$b_{\text{pred}-\text{maximin}} = \text{argmin}_{\beta} \max_{b} \mathbb{E}_X \|Xb - X\beta\|_2^2/n$$

regarding the latter:

$$\begin{aligned} V_{\beta,b} &= \mathbb{E}_Y \|Y\|_2^2/n - \mathbb{E}_{Y,X}\|Y - X\beta\|_2^2/n \\ &= \underbrace{b^T \Sigma b}_{\neq\ \text{const.}} - \mathbb{E}_X \|Xb - X\beta\|_2^2/n \end{aligned}$$

$\rightsquigarrow b_{\text{maximin}} \neq b_{\text{pred}-\text{maximin}}$

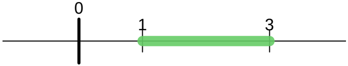
all the same for constant coefficients

if $|\mathrm{supp}(F_B)| = 1 \rightsquigarrow$

$$b_{\mathrm{maximin}} = b_{\mathrm{pred-maximin}} = b_{\mathrm{pool}}$$

b_{maximin} versus $b_{\text{pred-maximin}}$ (explaining variance versus prediction)

toy example I



$$\text{supp}(F_B) = \{1, 3\}$$
$$b_{\text{maximin}} = 1$$
$$b_{\text{pred-maximin}} = 2$$
$$b_{\text{pool}} \in (1, 3)$$

- $b_{\text{maximin}} = 1$:
  point in the convex hull of support closest to zero
- $b_{\text{pred-maximin}} = 2$:
  mid-point of the convex hull of support
- $b_{\text{pool}} \in (1, 3)$:
  weighted mean of support points

red statements are “true in general”

b_{maximin} versus $b_{\text{pred-maximin}}$ (explaining variance versus prediction)

toy example II

$$\text{supp}(F_B) = \{-4, 1\}$$
$$b_{\text{maximin}} = 0$$
$$b_{\text{pred-maximin}} = -1.5$$
$$b_{\text{pool}} \in (-4, 1)$$

- $b_{\text{maximin}} = 0$:
  point in the convex hull of support closest to zero
- $b_{\text{pred-maximin}} = -1.5$:
  mid-point of the convex hull of support
- $b_{\text{pool}} \in (-4, 1)$:
  weighted mean of support points

red statements are "true in general"

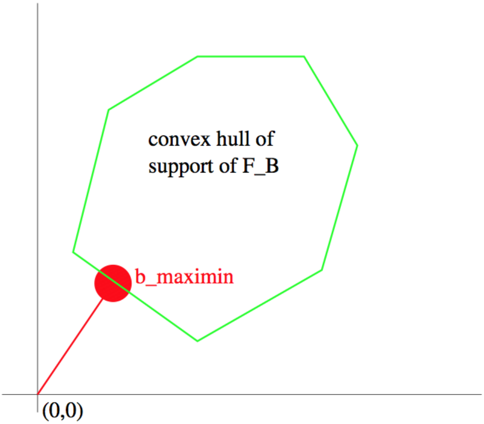
maximin effects: the value zero plays a special role

and we think that this makes most sense:
if some coefficients are negative and some are positive, we want to state that there is no “worst case effect”, i.e., assign the value zero

in general (NM & Bühlmann, 2014)

$b_{\mathrm{maximin}} =$ point in convex hull of $\mathrm{supp}(F_B)$ closest to zero

"closest" w.r.t. $d(\beta,\gamma)^2 = (\beta-\gamma)^T\Sigma(\beta-\gamma)$



different characterization

let

$$\text{Predictions} := X\beta$$
$$\text{Residuals} := Y - X\beta.$$

then

$$b_{\text{maximin}} = \text{argmax}_b \, E\left(\|\text{Predictions}\|_2^2\right) \text{ such that } \min_{b \in \text{supp}(F_B)} E\left(\text{Predictions} \cdot \text{Residuals}\right) \geq 0.$$

$\rightsquigarrow$ make maximally large prediction such that you never “get it wrong” the target parameter is b_{maximin}

and we can directly estimate it
without complicated fitting of the entire mixture model

assume G known groups/clusters for the samples $i = 1, \ldots, n$
within each group g: $B_i \equiv b_g \; \forall i \in g \; (g = 1, \ldots, G)$

(regularized) maximin estimator for known groups

$$\hat{\beta} = \mathrm{argmin}_{\beta} \max_{g} - \hat{V}_{\beta}^{g} \quad (+\lambda \|\beta\|_1)$$

(or with Ridge penalty $\lambda \|\beta\|_2^2$)

where empirical counterpart to $V_{b_g,\beta} = 2\beta^T \Sigma b_g - \beta^T \Sigma \beta$ in group g is

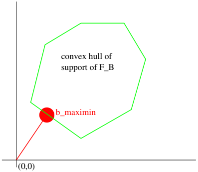
$$\hat{V}_{\beta}^{g} = \frac{2}{n_g} \beta^T X_g^T Y_g - \beta^T \underbrace{\hat{\Sigma}_g}_{n_g^{-1} X_g^T X_g} \beta$$

Closely related: maximin aggregation

Magging (PB & Meinshausen, 2014)

assume we know the G groups
$\rightsquigarrow$ assume we know true regression parameter b_g in every group g:

$$b_{\text{maximin}} = \text{argmin}_{\beta \in H} \beta^T \Sigma \beta,$$
$$H = \text{ convex hull of } \text{supp}(F_B)$$



$$\Longleftrightarrow$$

$$b_{\text{maximin}} = \sum_{g=1}^{G} w_g^* b_g \text{ (convex combination)}$$
$$w^* = \text{argmin}_{w_g} \sum_{g,g'} w_g w_{g'} b_g^T \Sigma b_{g'} \text{ s.t. } w_g \geq 0, \sum_g w_g = 1$$

Magging (**PB & Meinshausen, 2014**)

assume we know the G groups
$\rightsquigarrow$ assume we estimate true regression parameter b_g by $\hat{b}_g$ in every group g (using least-squares, Ridge or Lasso, ...)

$$b_{\text{maximin}} = \sum_{g=1}^{G} w_g^* b_g \text{ (convex combination)}$$

$$w^* = \text{argmin}_{w_g} \sum_{g,g'} w_g w_{g'} b_g^T \Sigma b_{g'} \text{ s.t. } w_g \geq 0, \sum_g w_g = 1$$

Use plug-in idea

$$\hat{b}_{\text{maximin}} = \sum_{g=1}^{G} \hat{w}_g^* \hat{b}_g \text{ (convex combination)}$$

$$\hat{w}^* = \text{argmin}_{w_g} \sum_{g,g'} w_g w_{g'} \hat{b}_g^T \hat{\Sigma} \hat{b}_{g'} \text{ s.t. } w_g \geq 0, \sum_g w_g = 1$$

Magging: convex maximin aggregating

$$\hat{b}_{\text{magging}} = \sum_{g=1}^{G} \hat{w}_g \hat{b}_g$$

$$\hat{w} = \text{argmin}_{w_g} \| \sum_{g=1}^{G} w_g X \hat{b}_g \|_2^2$$

$$\text{s.t. } w_g \geq 0, \sum_g w_g = 1$$

only a *G*-dimensional quadratic program
$\leadsto$ very fast to solve (if G is small or moderate)

very generic:
can e.g. use the Lasso for estimators $\hat{b}_g$ in each group g

in R-software environment:

computation of the aggregation weights

```
library(quadprog)
theta ← cbind(theta1,...,thetaG)   #the regression estimates
hatS ← t(X)
H ← t(theta) %*% hatS %*% theta
A ← rbind(rep(1,G),diag(1,G))
b ← c(1,rep(0,G))
d ← rep(0,G)
w ← solve.QP(H,d,t(A),b, meq = 1)
```

question on previous slide:

“how should we aggregate the subsampled estimates?”

$\leadsto$ answer for known groups:
maximin aggregation with convex combination weights $\hat{w}_b$

maximin effects estimator for unknown groups

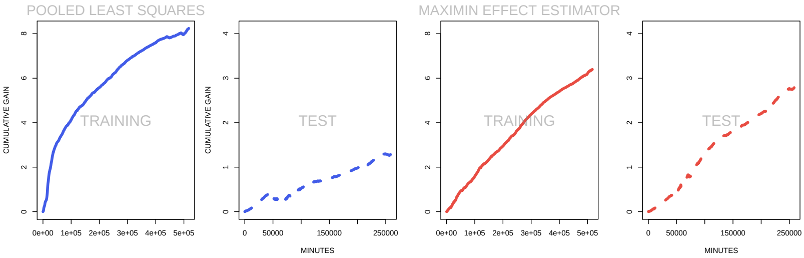
with e.g. time ordering:
build groups of consecutive observations
implicitly assuming “smooth” or “locally-constant” behavior of B_i w.r.t. index i

and use maximin estimator (or magging) from before

Example: minute returns of Euro-Dollar exchange rate

$p = 60$: twelve financial instruments,
with 5 lagged values each
$n_{\mathrm{train}} \approx 500'000$ consecutive observations
$n_{\mathrm{test}} \approx 250'000$ consecutive observations

maximin effects estimator with 3 groups of consecutive observations
cumul. plots: $\sum_{i=1}^{t} Y_i \hat{Y}_i$ vs. t, measuring "explaining variance"



$\rightsquigarrow$ maximin eff. estimator is substantially better than OLS

without any information on groups

randomly sample G groups of equal size $n_g \equiv m$

with these randomly sampled groups: maximin estimator as before

$$\hat{\beta} = \text{argmin}_{\beta} \max_{g} - \hat{V}_{\beta}^{g} + \lambda \|\beta\|_1$$

or magging

very easy to do!
is it too naive?

question on previous slide:

"is naive subsampling a valid approach?"

$\rightsquigarrow$ answer:
yes, in case of no structure

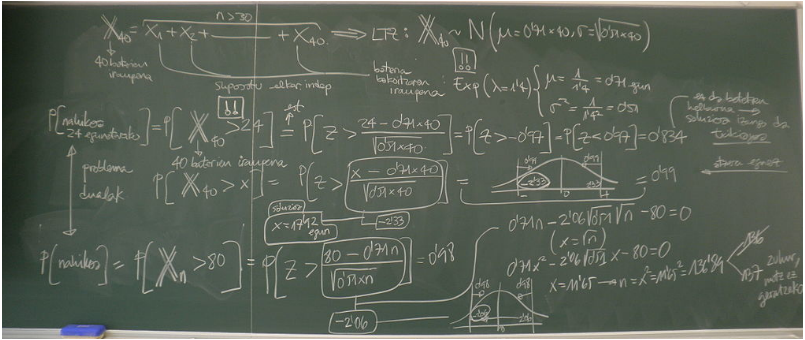
Summary

(A) in case of known groups: use these groups

(B) with time structure: build groups of consecutive observations

(C) without any information: randomly sample groups

Statistical theory (NM & Bühlmann, 2014)



oracle inquality for known groups

Assume $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. with sub-Gaussian distribution and, for simplicity: $n_g \equiv m \; \forall g$. For $\lambda \asymp \sigma \sqrt{\log(pG)/m}$, with high probability,

$$\text{perform. with estimator} \le \text{perform. with oracle} \quad + \text{error}$$

$$\underbrace{\max_g -V_{b_g,\hat{\beta}}}_{\max_{b \in \mathrm{supp}(F_B)} -V_{b,\hat{\beta}}} \le \underbrace{V^*}_{\min_\beta \max_{b \in \mathrm{supp}(F_B)} -V_{b,\beta}} \quad + \text{error}$$

$$\text{where error} = O(\kappa \sqrt{\frac{\log(pG)}{m}}) + O(\kappa^2 D)$$

$$\text{with } \kappa = \max(\|b_{\mathrm{maximin}}\|_1, \max_g \|b_g\|_1) \qquad (\text{“}\ell_1\text{-sparsity”})$$

$$\text{and } D = \max_g \|\hat{\Sigma}_g - \Sigma\|_\infty \overset{\text{typically}}{=} O(\sqrt{\frac{\log(G)}{m}}) \qquad (\text{cov.estimation})$$

for e.g. Gaussian design: sharper rate with $O(\kappa D)$

General case

Similar results are valid in more settings

(i) known groups with const. coeff. in each group (as shown)

(ii) chronological observations with a jump process

(iii) contaminated samples

Example (i)

known groups with const. coeff. in each group g

$$|\mathrm{supp}(F_B)| = G$$

$\rightsquigarrow$ magging and maximin estimation successful with shown oracle rates.

Example (ii)

chronological observations with a jump process
$|\mathrm{supp}(F_B)| = J$

$$B_i = \begin{cases} B_{i-1} & \text{with prob. } 1-\delta, \\ U_i & \text{otherwise} \end{cases}$$

$$U_1, \ldots, U_n \text{ i.i.d. Unif.}(\mathrm{supp}(F_B))$$

build groups of consecutive observations of equal size
Previous bound holds with probability $\geq 1-\gamma$ if

$$G \geq 4n\delta J/\gamma,$$
$$\delta(n-1)/J \geq 1/\log(2J/\gamma)$$

$\leadsto$ works well if $n\delta$ sufficiently large and $G \geq O(n\delta)$

Example (iii): contaminated samples ($|\mathrm{supp}(F_B)| = \infty$)

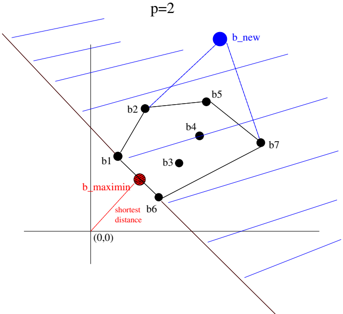
assume that $B_1, \ldots, B_n$ i.i.d. with

$$B = \begin{cases} b_{\mathrm{maximin}} = \text{“true parameter” “}= \beta^0\text{”} & \text{with prob. } 1-\delta, \\ U & \text{otherwise} \end{cases}$$

$U \sim$ a distribution on $\mathbb{R}^p$ such that

$$(u - b_{\mathrm{maximin}})^T \Sigma b_{\mathrm{maximin}} \geq 0 \; \forall u \in \mathrm{supp}(U)$$

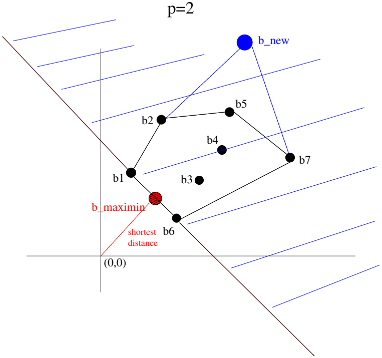
for $\delta > 0$ small $\rightsquigarrow$ small amount of contamination



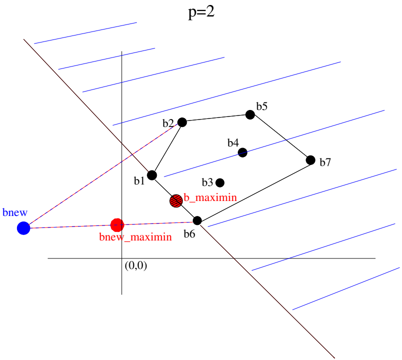
randomly sample G groups of equal size m without replacement within groups and with replacement between groups

with $m = O(1/|\log(1-\delta)|)$ and $G \geq O(|\log(\gamma)|)$
$\rightsquigarrow$ Pareto condition holds with probability $1-\gamma$
too large G pays a price for estimation error

illustration:



if
$(u - b_{\text{maximin}})^T \Sigma b_{\text{maximin}} \geq 0 \; \forall u \in \text{supp}(U)$
fails, estimate will just shrink towards origin



$\rightsquigarrow$ good breakdown point/robustness of the maximin effects est.

Robustness in a simulation experiment

sparse linear model with $p = 500$, $s_0 = 10$ active variables
5% or 17% outliers: different coefficients for same act. var.
sample size $n = 300$

use magging:
Lasso for each $G = 6$ randomly subsampled groups of 100 obs.

relative improvements over pooled Lasso

	method	out-sample L_2	$\|\hat{\beta} - \beta^0\|_1$	$\|\hat{\beta} - \beta^0\|_2$
5% outliers	magging	13.7%	36%	31.7%
	pooled Lasso	0%	0%	0%
	mean $\overline{Y}$	-2.5%	–	–
17% outliers	magging	6.9%	47.2%	49.8%
	pooled Lasso	0%	0%	0%
	mean $\overline{Y}$	-6.0%	–	–

$\rightsquigarrow$ easy and efficient way to achieve robustness!

negative Example (iv)

with continuous support for F_B or
discrete but growing support ($G = G_n$)

$\leadsto$ random sampling of groups of equal size m
will in general not be good enough

data example (Kogan et al., 2009)

predicting risk from financial reports ("fundamental company data") with regression

response: stock return volatility in twelve month period after the release of reports (for thousands of publicly traded U.S. companies)

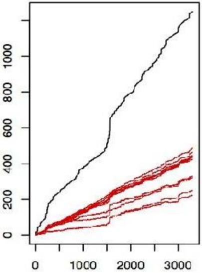
predictor variables: unigrams and bigrams of word frequencies in the reports

$p \approx 4.27 \cdot 10^6, \; n \approx 19'000$

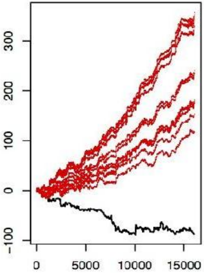
training set: first 3'000 observations
test set: remaining 16'000 observations

maximin effects estimator based on G groups of consecutive observations
(reports are ordered chronologically)

cumulative plots of $\sum_{i=1}^{t} Y_i \hat{Y}_i$ versus t



training set

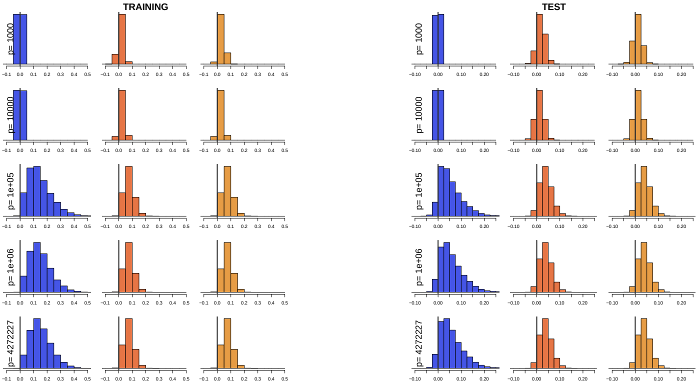


test set

black: pooled Ridge estimator

red: maximin effects est with Ridge ℓ_2-norm penalty for different number of groups G

$\rightsquigarrow$ fitting a group of outliers is bad for pooled Ridge plots: histograms for $\sum_{i \in \mathcal{I}} Y_i \hat{Y}_i$, $\mathcal{I}$ random subset of size 500



orange: maximin effects estimator with Lasso ℓ_1-norm penalty and $G = 3$ groups of consecutive observations

yellow: as above but G cross-validated

blue: pooled Lasso estimation

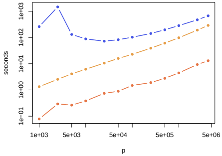
$\rightsquigarrow$ maximin effects estimator exhibits less variability

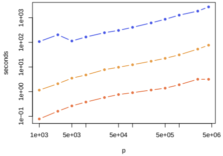
Computational properties

maximin effects est. much faster than pooled
Lasso (`glmnet` Friedman, Hastie & Tibshirani, 2008) or Ridge

CPU as function of p, $n = 3000$

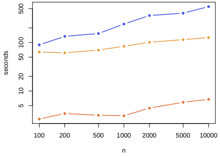
ℓ_1-norm regul. ℓ_2-norm regul.

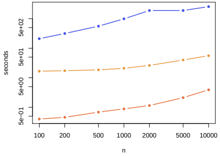




CPU as function of n, $p = 10^6$

ℓ_1-norm regul. ℓ_2-norm regul.





maximin $G = 3$, maximin CV-optim. G, Lasso/Ridge (CV)

memory requirements

for ℓ_1-norm regularized estimation:

- maximin effects est. with "maximal" penalty $\lambda \to \lambda_{\max}$
  (for "high-dimensional, noisy scenarios")
  memory of order $O(pG)$
- pooled Lasso:
  memory of order $O(\min(np, p^2))$

with few groups G: maximin eff. est. needs much less memory than Lasso

Conclusions

random subsampling and maximin aggregation/estimation:
statistically powerful and computationally efficient
for robust inference in large-scale, inhomogeneous data

- for fitting the “stable/consistent” maximin effects in a heterogeneous mixture regression model
  without fitting the mixture model!
- with good statistical properties

and there remain many open issues...

Thank you!

References:

Meinshausen, N. and Bühlmann, P. (2014). Maximin effects in inhomogeneous large-scale data. Preprint arXiv:1406.0596.

Bühlmann, P. and Meinshausen, N. (2014). Magging: maximin aggregation for inhomogeneous large-scale data. Preprint arXiv:1409.2638