From Data to Decisions

Part I: Data-Driven Robust Optimization

Nathan Kallus

15.094J Robust Modeling, Optimization and Computation

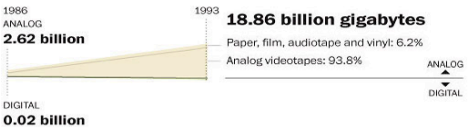
Massachusetts Institute of Technology

"Probability distributions are not known in practice; they exist in our imagination."

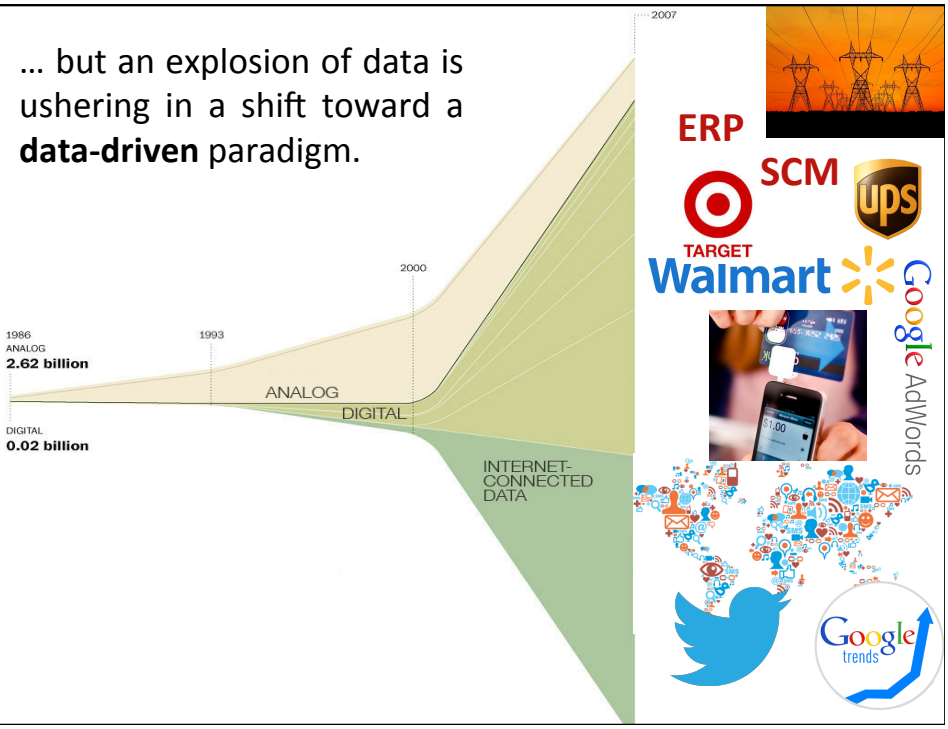
D. Bertsimas

Uncertainty sets are not known in practice either. Both are models of reality.

In data-poorer times, decision-making was necessarily **model-driven**…



… but an explosion of data is ushering in a shift toward a **data-driven** paradigm.



From Data to Decisions

Lecture 7: Designing uncertainty sets from data

Lecture 8: From predictive to prescriptive analytics

Based on papers
Data-Driven Robust Optimization. Bertsimas, Gupta, K.
Robust SAA. Bertsimas, Gupta, K.
From Predictive to Prescriptive Analytics. Bertsimas, K.

All available at www.nathankallus.com

"Probability distributions are not known in practice; they exist in our imagination."

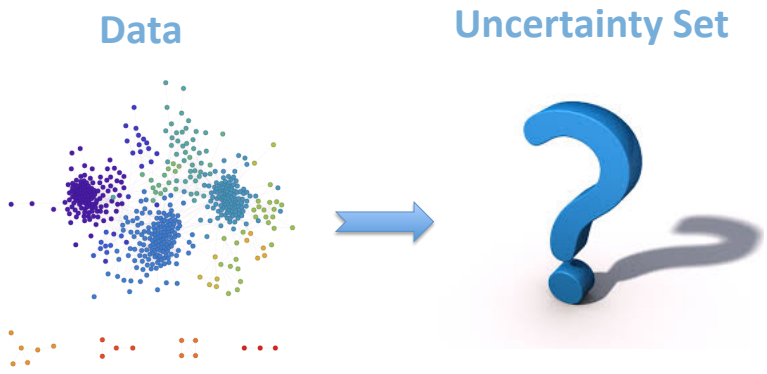
D. Bertsimas

Uncertainty sets are not known in practice, either. Both are models of reality.

*"What is available in practice is **data**."*

D. Bertsimas

Designing uncertainty sets from data

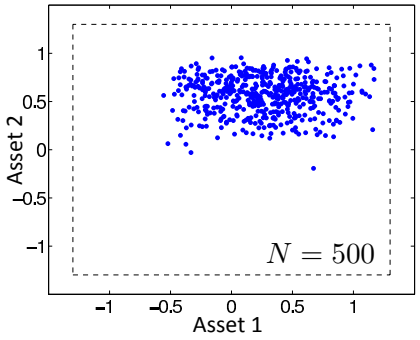


Example: Portfolio Allocation

- How much to invest in each of $d = 2$ assets?
  - No shorting, budget constraint
  - Maximize worst-case returns

$$\max_{\mathbf{x} \geq \mathbf{0}, t} \quad t$$
$$\text{s.t.} \quad u_1 x_1 + u_2 x_2 \geq t \quad \forall \mathbf{u} \in \mathcal{U}$$
$$x_1 + x_2 = 1.$$

How should we model $\mathcal{U}$?



Question:

- What features would we want a good uncertainty set to have?

Recap…

- Choice of $\mathcal{U}$ is crucial
- Many different (heuristic) constructions
  - Bounding ellipsoids, convex hull, CLT-type motivations, …
  - General purpose constructions:
    - Ben-Tal and Nemirovski [1999, 2009]; Bertsimas and Sim 2003; Chen et al 2010; Bertsimas and Brown 2009; Bertsimas and Bandi 2013.
- Many ways to define a "good" set
  - Tractability, probabilistic guarantees, volume, asymptotic behavior, performance relative to benchmark, application specific criteria

Outline

- Designing uncertainty sets from data
  - **Model setup**
  - Overview of schema
  - Building up the schema
    - Geometry
    - Statistics
    - The schema
  - Applications

Set up

- RO models uncertain constraints

$$f(x, \tilde{u}) \leq 0 \qquad \text{(e.g. } u^T x \leq b\text{)}$$

Decision variables — Uncertain (random) parameter

using robust constraints

$$f(x, u) \leq 0 \quad \forall u \in \mathcal{U}$$

- *Data-driven* RO:
  - $\tilde{\mathbf{u}} \sim \mathbb{P}^*$ has ***unknown*** distribution
  - We only have data $\mathcal{S} = \{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_N\}$ drawn iid from $\mathbb{P}^*$
  - And perhaps some a priori knowledge on $\mathbb{P}^*$ such as support information

Our goal…

- Methodology for constructing uncertainty sets from data
  - Yields *many* different data-driven sets tailored to different modeling situations
  - Connects to applied statistics
  - Significantly improves numerical performance
- Our new sets
  - Are **tractable**
  - Have **performance guarantee**
  - **Learn** their shape from features of the data
  - **Shrink** as more data become available
    - Reduces conservatism in robust models

Probabilistic guarantee

Fix $\epsilon > 0$.

$\mathcal{U}$ **implies a probabilistic guarantee** for $\mathbb{P}^*$ at level ϵ if for any $f(u, x)$ concave in u we have:

$$f(\mathbf{u}, \mathbf{x}) \leq 0 \quad \forall \mathbf{u} \in \mathcal{U} \implies \mathbb{P}^*(f(\mathbf{u}, \mathbf{x}) \leq 0) \geq 1 - \epsilon$$

- Usually have $f(u, x) = u^T x - b$
- "Robust feasibility implies feasibility with high probability"
- Defined with respect to **true** probability $\mathbb{P}^*$
- Most non-data-driven proposals for uncertainty sets satisfy a similar guarantee
- In data-driven setting, we don't know $\mathbb{P}^*$

The Data-Driven RO Problem

Given the data $\mathcal{S} = \{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_N\}$, find a convex set $\mathcal{U}(\mathcal{S})$ that implies a probabilistic guarantee with respect to the *unknown* $\mathbb{P}^*$.

Want:

Theorem: Fix $0 < \delta < 1$ and $0 < \epsilon < 1$. With probability at least $1 - \delta$ with respect to data collection, any set $\mathcal{U}(\mathcal{S})$ constructed according to our schema implies a probabilistic guarantee at level ϵ.

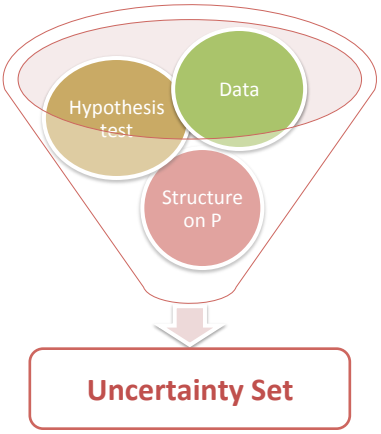
Outline

- Designing uncertainty sets from data
  - Model setup
  - **Overview of schema**
  - Building up the schema
    - Geometry
    - Statistics
    - The schema
  - Applications

A Schema for Data-Driven RO

Develop a novel and generic schema for going from data to uncertainty sets

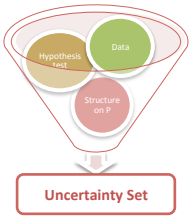
Schema



Key theoretical ingredients

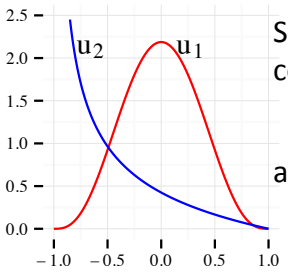
- **Hypothesis testing** to describe ambiguity about data
- **Risk theory** to describe the risk of infeasibility
- **Convex analysis** to describe a convex, tractable uncertainty set

Variety of Uncertainty Sets



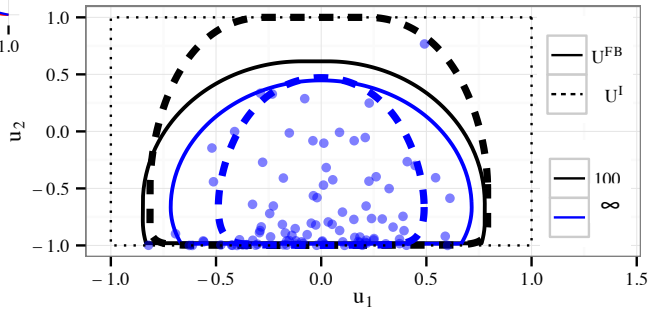
Assumptions on $\mathbb{P}$	Hypothesis Test	Geometric Description	Separation
Discrete support	χ^2-test	SOC	
Discrete support	G-test	Polyhedral*	
Independent marginals	KS Test	Polyhedral*	line search
Independent marginals	K Test	Polyhedral*	line search
Independent marginals	CvM Test	SOC*	
Independent marginals	W Test	SOC*	
Independent marginals	AD Test	EC	
Independent marginals	Chen et al. (2007)	SOC	closed-form
None	Marginal Samples	Box	closed-form
None	Linear Convex Ordering	Semi-infinite LP	linear optimizati
None	Shawe-Taylor & Cristianini (2003)	SOC	closed-form
None	Delage & Ye (2010)	LMI	

What do these sets look like?

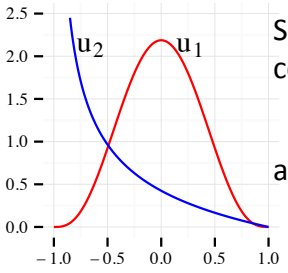


Suppose P has independent components with these PDFs

and we know components are independent

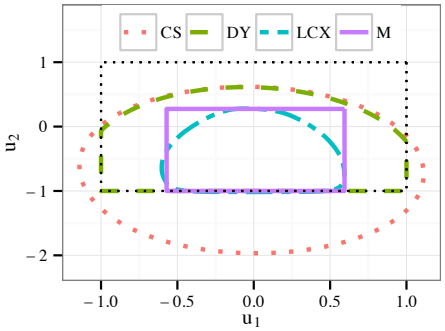


What do these sets look like?



Suppose P has independent components with these PDFs

and we *don't* know components are independent



Outline

- Designing uncertainty sets from data
  - Model setup
  - Overview of schema
  - **Building up the schema**
    - **Geometry**
    - Statistics
    - The schema
  - Applications

Probabilistic Guarantees

- Recall, $\mathcal{U}$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ if for any $f(u, x)$ concave in u we have:

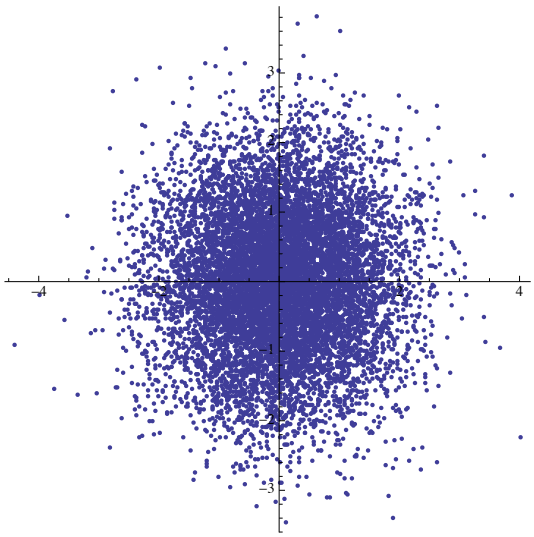
$$f(\mathbf{u}, \mathbf{x}) \leq 0 \;\; \forall \mathbf{u} \in \mathcal{U} \implies \mathbb{P}^*(f(\mathbf{u}, \mathbf{x}) \leq 0) \geq 1 - \epsilon$$

- What does this look like?
- Suppose we know $\tilde{\boldsymbol{u}} \sim \mathcal{N}(0, \boldsymbol{I}_d)$
- Let's try to develop a (non-data-driven) uncertainty set that implies a probabilistic guarantee
- Let's focus on $f(u, x) = u^T x - b$

Let's try…

$d = 2$

$\tilde{\boldsymbol{u}} \sim \mathcal{N}(0,\, \boldsymbol{I}_d)$



Let's try…

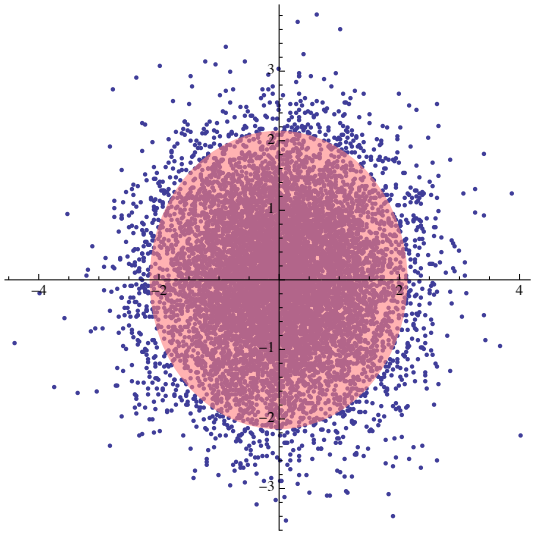
$d = 2$

$\mathcal{U}_1 = \{\|u\| \leq 2.145\}$

Fix $\boldsymbol{u}^T \boldsymbol{x} \leq b \quad \forall \boldsymbol{u} \in \mathcal{U}_1$

$\mathbb{P}(\tilde{\boldsymbol{u}} \in \mathcal{U}_1) = 0.9$

$\implies \mathbb{P}(\tilde{\boldsymbol{u}}^T x \leq b) \geq 0.9$



Let's try…

$$d = 2$$

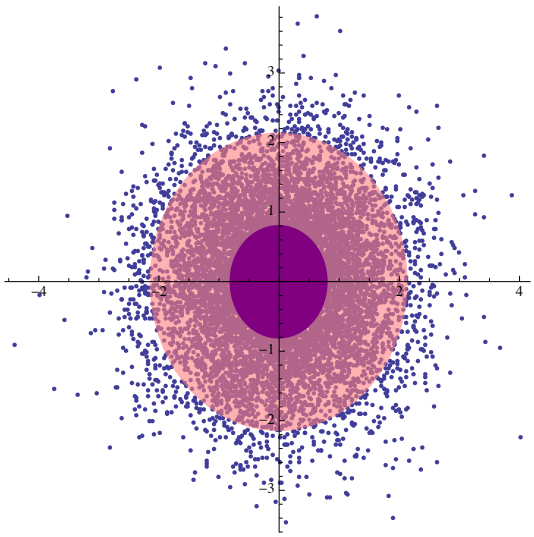
$$\mathcal{U}_2 = \{\|u\| \leq 0.816\}$$

Fix $\boldsymbol{u}^T\boldsymbol{x} \leq b \quad \forall \boldsymbol{u} \in \mathcal{U}_2$

$$\mathbb{P}(\tilde{\boldsymbol{u}} \in \mathcal{U}_2) = 0.28$$

And yet…

$$\mathbb{P}(\tilde{\boldsymbol{u}}^T x \leq b) = 0.9$$



Probabilistic Guarantees

- Recall, $\mathcal{U}$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ if for any $f(u, x)$ concave in u we have:

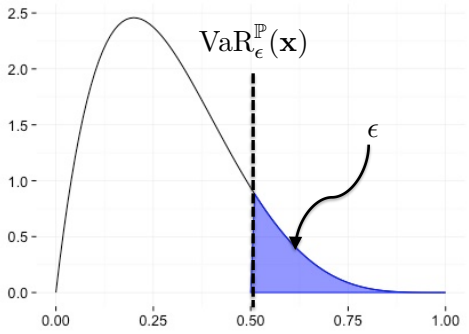
$$f(\mathbf{u}, \mathbf{x}) \leq 0 \;\; \forall \mathbf{u} \in \mathcal{U} \implies \mathbb{P}^*(f(\mathbf{u}, \mathbf{x}) \leq 0) \geq 1 - \epsilon$$

- Conclusion:
  - Covering probability mass is much too conservative!
  - Need to focus on the constraint and its special structure

Geometry of Guarantees: Value-at-Risk

- For any $\mathbf{x} \in \mathbb{R}^d$ and probability measure $\mathbb{P}$, define Value-at-Risk (VaR) of $\mathbf{x}$ at level ϵ as

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \equiv \inf\left\{t : \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x} > t) \leq \epsilon\right\}$$



Value-at-Risk Example

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \equiv \inf\left\{t : \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x} > t) \leq \epsilon\right\}$$

- Suppose $\tilde{\mathbf{u}} \sim N(\mathbf{0}, \mathbf{I})$ d = 3
- What is $\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x})$ for $\epsilon = 0.05$ and $\mathbf{x} = (1, 0, 2)^T$?

Answer:

- Note $\tilde{\mathbf{u}}^T\mathbf{x} \sim N(0, 5)$
- The 95% quantile of this distribution is

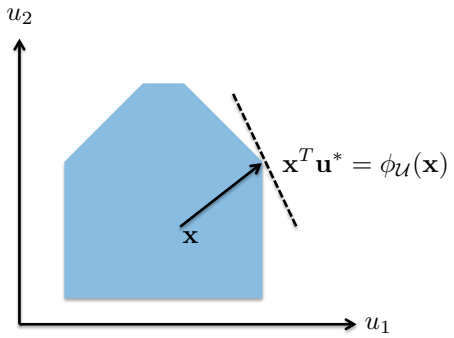
$$\sqrt{(5)}\Phi(.95) \approx 2.24 \times 1.65 \approx 3.69$$

Geometry of Guarantees: Support functions

- Recall our linear robust constraint

$$\mathbf{u}^T\mathbf{x} \leq 0 \quad \forall \mathbf{u} \in \mathcal{U} \iff \max_{\mathbf{u}\in\mathcal{U}} \mathbf{u}^T\mathbf{x} \leq 0$$

- We call this the *support function* $\phi_{\mathcal{U}}(\mathbf{x}) \equiv \max_{\mathbf{u}\in\mathcal{U}} \mathbf{x}^T\mathbf{u}$



Important facts about support functions:

$\phi_{\mathcal{U}}$ is convex and positively homogenous

For any convex, positively homogenous ϕ, there exists compact, non-empty, convex set $\mathcal{U}$ s.t. $\phi = \phi_{\mathcal{U}}$

Support Function Example

$$\phi_{\mathcal{U}}(\mathbf{x}) \equiv \max_{\mathbf{u}\in\mathcal{U}} \mathbf{x}^T\mathbf{u}$$

- Consider an ellipse: $\mathcal{U} = \{\mathbf{u} : (\mathbf{u}-\mathbf{c})^T\mathbf{M}(\mathbf{u}-\mathbf{c}) \leq d\}$
- What is $\phi_{\mathcal{U}}(\mathbf{x})$?

Answer: (See board work)

Geometry of Guarantees

Theorem:

Suppose $\mathcal{U}$ is nonempty, convex, compact set.
Then $\mathcal{U}$ implies a probabilistic guarantee at level ϵ,
(i.e., $f(\mathbf{u},\mathbf{x}) \le 0 \;\; \forall \mathbf{u} \in \mathcal{U} \implies \mathbb{P}^*(f(\mathbf{u},\mathbf{x}) \le 0) \ge 1-\epsilon$
for any $f(\mathbf{u},\mathbf{x})$ concave in $\mathbf{u}$)
if and only if

$$\phi_{\mathcal{U}}(x) \ge \mathrm{VaR}_\epsilon^{\mathbb{P}}(x) \quad \forall x \in \mathbb{R}^d$$

Geometry of Guarantees

- In words: $\mathcal{U}$ implies a probabilistic guarantee iff its support function upper bounds VaR
- **Idea**: Find a convex, positively homogenous bound, identify the set that has this as its support function
- **Bad**: Computing VaR involves determining the distribution of a sum of RVs, which is **hard** in general.
- **Good**: Variety of tight bounds exist, e.g.,

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \le \mathbb{E}[\tilde{\mathbf{u}}]^T\mathbf{x} + \sqrt{\frac{1}{\epsilon}-1}\sqrt{\mathbf{x}^T\mathrm{Cov}(\tilde{\mathbf{u}})\mathbf{x}}$$

- **Bad**: Given convex, positively homogenous ϕ, no general procedure to identify a set with such support fn.
- **Good**: Often easy to guess such a set.

Outline

- Designing uncertainty sets from data
  - Model setup
  - Overview of schema
  - **Building up the schema**
    - Geometry
    - **Statistics**
    - The schema
  - Applications

Background: hypothesis testing

- Use data to compare two hypotheses about the unknown distribution(s) that generated the data

  H_0 vs. H_A (null vs. alternative)

- Fixing $0 < \delta < 1$, a test gives
  - **Threshold** that depends on δ
  - **Statistic** that depends the data
- If **statistic** > **threshold**, then reject H_0
- Otherwise, "insufficient evidence"

Guarantee: If H_0 is true, with probability at least $1 - \delta$, we will not reject it.

Example: A test for the mean

- Suppose we knew (somehow) that
  $\mathrm{Cov}(\tilde{\mathbf{u}}) = \boldsymbol{\Sigma}$ and $\mathbb{P}\left((\boldsymbol{u} - \mathbb{E}\boldsymbol{u})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{u} - \mathbb{E}\boldsymbol{u}) \leq R^2\right) = 1$
- Hypotheses: $H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}_0$
  $H_A : \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}] \neq \boldsymbol{\mu}_0$
- Test statistic: $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$
- Threshold: $\frac{R^2}{N}\left(2 + \sqrt{2\log(1/\delta)}\right)^2$

Guarantee: If the true mean is actually $\boldsymbol{\mu}_0$, then with probability at least $1 - \delta$, we will not reject it.

(See Shawe-Taylor and Cristianini, 2003)

Thought experiment….

- Nishanth randomly generated some data and handed it to us…
- Martin claims to know that the true mean is $(1, 0, 2)^T$

- Let's test Martin's hypothesis
- What's the null and what's the alternative?
- Will we reject the null hypothesis?

Confidence regions

- A confidence region is the set of hypotheses which would pass the test given the data $\mathcal{S}$
- In our example, the confidence region is

$$\left\{\boldsymbol{\mu}_0 : (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \leq \frac{R^2}{N} \left(2 + \sqrt{2\log(1/\delta)}\right)^2\right\}$$

- Abuse notation: call the following set a confidence region, too

$$\mathcal{P}(\mathcal{S}) = \left\{\mathbb{P} : \mathbb{E}^{\mathbb{P}}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}_0, \quad (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \leq \frac{R^2}{N} \left(2 + \sqrt{2\log(1/\delta)}\right)^2\right\}$$

Why confidence regions?

- Recall from our thought experiment:

With probability at least $1 - \delta$, will not reject the true mean $\mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}]$.

$\Longleftrightarrow$

With probability at least $1 - \delta$, the true mean is in the confidence region defined by the data.

$\Longleftrightarrow$

With probability at least $1 - \delta$, $\mathbb{P}^* \in \mathcal{P}(\mathcal{S})$.

Why is this important?

- Using the confidence region, can identify a set of distributions that contains the true distribution with high probability
- This logic can be extended to *any* choice of hypothesis test
  - Obtain slightly different set of distributions

Outline

- Designing uncertainty sets from data
  - Model setup
  - Overview of schema
  - **Building up the schema**
    - Geometry
    - Statistics
    - **The schema**
  - Applications

The Schema

Fix $0 < \delta < 1$ and $0 < \epsilon < 1$

1. Choose a hypothesis test. Let $\mathcal{P}(\mathcal{S})$ be its $1-\delta$ confidence region.
2. Find a convex function $\phi(\mathbf{x}, \mathcal{S})$ s.t.
$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S})} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq \phi(\mathbf{x}, \mathcal{S})$$
3. Identify set $\mathcal{U}(\mathcal{S})$ whose support function is $\phi(\mathbf{x}, \mathcal{S})$

Theorem: Fix $0 < \delta < 1$ and $0 < \epsilon < 1$. With probability at least $1-\delta$, $\mathcal{U}(\mathcal{S})$ implies a probabilistic guarantee at level ϵ.

Example: A toy construction

- A priori assumptions:

  $\mathrm{Cov}(\tilde{\mathbf{u}}) = \boldsymbol{\Sigma}$ $\quad \mathbb{P}\left((\boldsymbol{u} - \mathbb{E}\boldsymbol{u})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{u} - \mathbb{E}\boldsymbol{u}) \leq R^2\right) = 1$

- Hypothesis test:

  $H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}_0$

  $H_A : \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}] \neq \boldsymbol{\mu}_0$

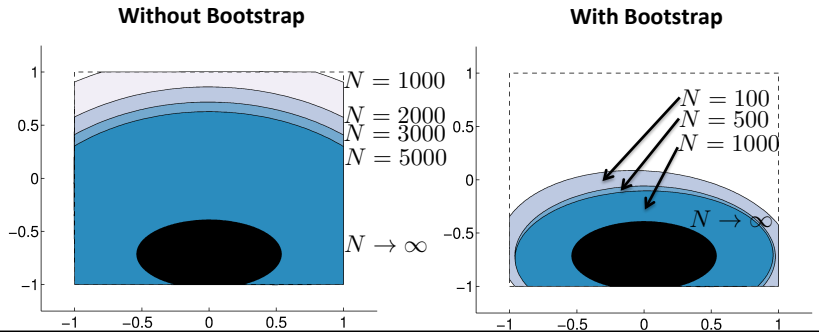
  $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$

  $\frac{R^2}{N}\left(2 + \sqrt{2\log(1/\delta)}\right)^2$ → Γ^2

- Follow steps of the schema…
  - Recall $\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq \mathbb{E}[\tilde{\mathbf{u}}]^T \mathbf{x} + \sqrt{\frac{1}{\epsilon} - 1}\sqrt{\mathbf{x}^T \mathrm{Cov}(\tilde{\mathbf{u}})\mathbf{x}}$
  - (See board work)

Why hypothesis tests?

- Various ways to construct of confidence regions
  - Concentration inequalities, phi-divergences, empirical process theory
- Hypothesis testing provides a unified perspective
  - "Duality between confidence regions and hypothesis tests"
- Well-developed tool in applied statistics



Outline

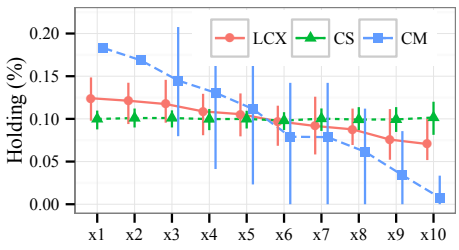
- Designing uncertainty sets from data
  - Model setup
  - Overview of schema
  - Building up the schema
    - Geometry
    - Statistics
    - The schema
  - **Applications**

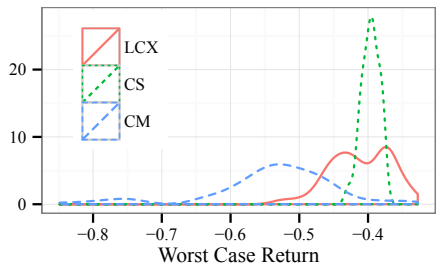
RO in Applications

- TONS of applications of RO
  - **Robust** Revenue Management (Perakis & Roels 2010, Ball & Queyranne 2009)
  - **Robust** Inventory Management (Bertsimas & Thiele 2006, Solyali et al. 2011)
  - **Robust** Supply Chain Management (Bertsimas & Thiele 2004)
  - **Robust** Unit Commitment (Bertsimas et al. 2013)
  - **Robust** Facility Location (Baron et al. 2011)
  - **Robust** Queuing (Bertsimas et al. 2011, Bandi et al. 2012)
  - **Robust** Portfolio Allocation (Goldfarb & Iyengar 2003)
  - …
- *Every* RO application depends on an uncertainty set
- Can retrofit these applications of RO with data-driven sets!

Ex: portfolio allocation

- 10 assets, 10%-worst-case return, 90%-confidence
- No a priori knowledge on P
- Reality: indep, 0% mean, 1% SD, differing skews
- Compare RO portfolio w/ data-driven uncertainty set to RO portfolio w/ Calafiore and Monastero (2012)

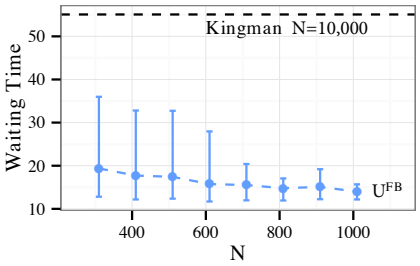




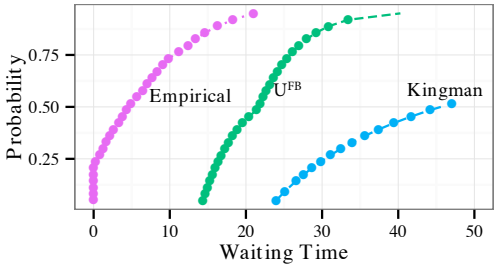
Ex: queueing

- G/G/1 queue (reality: M/Pareto/1, rho=0.9)
- Robust queuing model
  - Bertsimas et al. (2011), Bandi et al. (2012)
- Supplant uncertainty set with our data-driven sets
- v.s. data-driven Kingman bound + Markov inequality

Median waiting time



All quantiles



In the paper…

- Develop a great variety of sets in this way

Assumptions on $\mathbb{P}$	Hypothesis Test	Geometric Description	Separation
Discrete support	χ^2-test	SOC	
Discrete support	G-test	Polyhedral*	
Independent marginals	KS Test	Polyhedral*	line search
Independent marginals	K Test	Polyhedral*	line search
Independent marginals	CvM Test	SOC*	
Independent marginals	W Test	SOC*	
Independent marginals	AD Test	EC	
Independent marginals	Chen et al. (2007)	SOC	closed-form
None	Marginal Samples	Box	closed-form
None	Linear Convex Ordering	Semi-infinite LP	linear optimizati
None	Shawe-Taylor & Cristianini (2003)	SOC	closed-form
None	Delage & Ye (2010)	LMI	

Also:

- Computation and tractability
- Nonlinear and simultaneous constraints
- Applications

From Data to Decisions

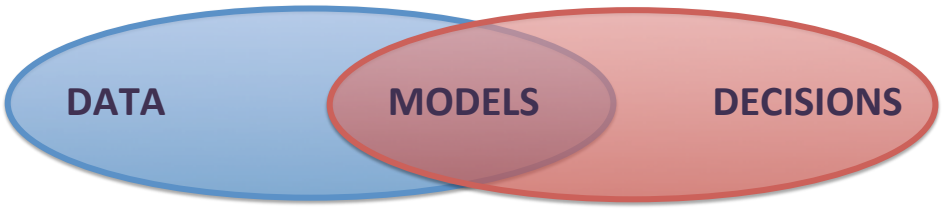
Part II: From Predictive to Prescriptive Analytics

Nathan Kallus

15.094J Robust Modeling, Optimization and Computation

Massachusetts Institute of Technology

"What is available in practice is data"



Statistics/ Machine learning

- Focus on leveraging *data* in various shapes and forms to build predictive models
- No decisions…
- To a great extent, ML is behind the *analytics edge*

Operations Research/ Management Science

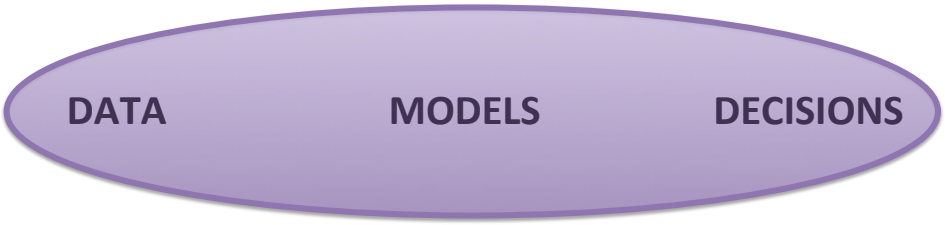
- Focus on modeling and optimization (decision making) under uncertainty
- Traditionally, little to no deference to data
- Especially, large-scale data

Applied Predictive Analytics

Data	Prediction Problem	Prescription Problem
Web Search	Predict video game demand (Goel et al. '10)	Inventory management for video game titles
Twitter	Predict box-office gross (Asur & Huberman '10)	Assign capacities (cinemas)
Blogs	Predict amazon book sales (Gruhl et al. '05)	Facility location, shipment planning
Twitter & News	Predict civil unrest (Kallus '14)	Supply chain setup

… … …

"What is available in practice is data"



Case Study: Distribution Arm of International Media Conglomerate

- Fortune Global 100 company
- 1 billion units of entertainment media shipped per year
- Sells 1/2 million different titles on CD/DVD/Bluray at over 50,000 retailers worldwide
- Vendor-managed inventory
- Scan-based trading
- Maximize number media sold

Case Study: Distribution Arm of International Media Conglomerate

- Difficult problem:
  - Limited shelf space at retail locations
  - Huge array of potential titles
  - Highly uncertain demand for new releases

Case Study: Distribution Arm of International Media Conglomerate

- Difficult problem…
- But, there's tons of data that might help
  - 4 years of sales data across a network of 50,000 retailers
  - Data harvested from public online sources

IMDb Rotten Tomatoes Google trends

- How to leverage all this data?

A general problem

- Data $y^1, \ldots, y^N$ on quantity(ies) of interest Y
  E.g. demands at locations/of products, % returns
- Data $x^1, \ldots, x^N$ on associated covariates X
  E.g. recent sales figures, search engine attention
- Decision $z \in \mathcal{Z}$ to minimize *uncertain* costs $c(z; Y)$ after observing $X = x$

What we'll look at

- **A new framework**
  - Unifies ML and OR/MS
  - General purpose
- **Theory**
  - Computational tractability
  - Asymptotic optimality
- **Performance metric**
  - Coefficient of prescriptiveness
- **Practice**
  - Case study of huge media distributor
  - Study *prescriptive* power of large-scale data

Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Case study

Standard Data-Driven Optimization

- Data $y^1, \ldots, y^N$ on quantity(ies) of interest Y
- Decision $z \in \mathcal{Z}$ to minimize *uncertain* costs $c(z; Y)$
- Hypothetical problem of interest is

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y)\right]$$

- But we only have data
  - "Distributions [are] in our imagination" – D. Bertsimas

Standard Data-Driven Optimization

- Hypothetical problem of interest is

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y)\right]$$

- Standard data-driven solution is sample average approximation

$$\hat{z}_N^{\mathrm{SAA}} \in \arg\min_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} c(z;y^i)$$

- Other approaches: SA (Robins 1951),
  Robust SAA (Bertsimas, Gupta, K 2014)
- General theorem: solutions from these approaches converge to the hypothetical full-info problem (cf. Shapiro et al. 2009)
- In our problem data-driven approaches like SAA account for uncertainty but not for auxiliary data

Standard Supervised Learning in ML

- Data $y^1, \ldots, y^N$ on quantity(ies) of interest Y
- Data $x^1, \ldots, x^N$ on associated covariates X
- Hypothetical problem of interest is to come up with a prediction $\hat{m}_N(x)$ for the a future value of Y after observing $X = x$ so that the difference (squared) between our best guess and the true value is small…
- I.e., the hypothetical target is $\mathbb{E}\left[Y \middle| X = x\right]$
- For example, a random forest!

Standard Supervised Learning in ML

- Hypothetical problem of interest is
$$\mathbb{E}\left[Y \middle| X = x\right]$$
- How to use for decision-making?
- Fit a ML predictive model $\hat{m}_N(x) \approx \mathbb{E}\left[Y \middle| X = x\right]$ (e.g. a random forest) and, since it's such a good guess, solve a deterministic problem
$$\hat{z}_N^{\text{point-pred}}(x) \in \arg\min_{z \in \mathcal{Z}} c(z; \hat{m}_N(x))$$
- In our problem, this point-prediction-driven decision accounts for auxiliary data but not for uncertainty

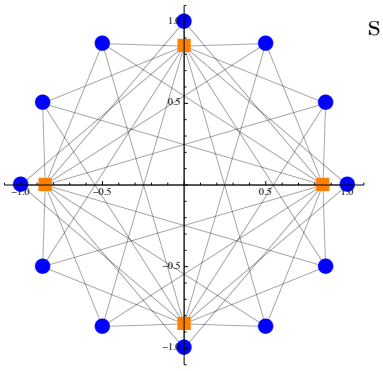
The predictive prescription problem

- *Our* problem of interest:
$$z^*(x) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y) \middle| X = x\right]$$
- Hypothetical full-information optimum
  - Uses knowledge of $\mu_{X,Y}$ to leverage $X = x$ to greatest possible extent in reducing costs
- **Our task**:
use data $S_N = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ to construct a data-driven predictive prescription
$$\hat{z}_N(x) : \mathcal{X} \to \mathcal{Z}$$

Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week

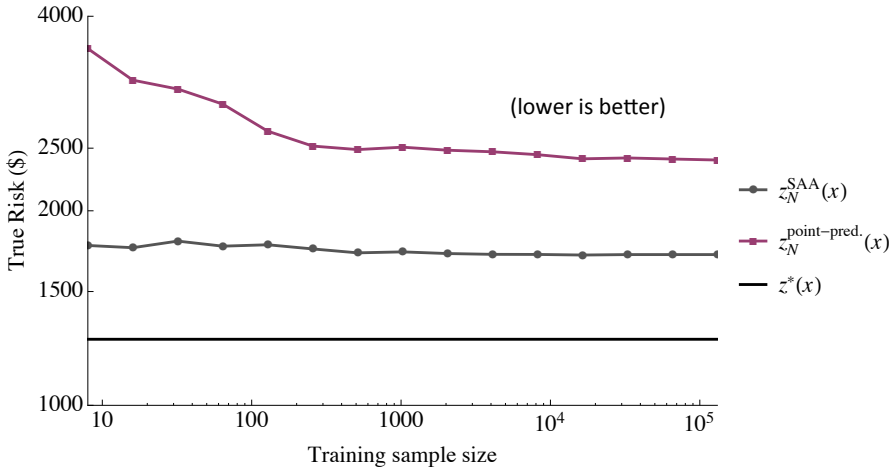
$$c(z;y) = p_1 \sum_{i=1}^{d_z} z_i \quad + \quad \min \left(p_2 \sum_{i=1}^{d_z} t_i + \sum_{i=1}^{d_z} \sum_{j=1}^{d_y} c_{ij} s_{ij} \right)$$



$$\text{s.t. } t_i \geq 0 \qquad \forall i$$
$$s_{ij} \geq 0 \qquad \forall i, j$$
$$\sum_{i=1}^{d_z} s_{ij} \geq y_j \qquad \forall j$$
$$\sum_{j=1}^{d_y} s_{ij} \leq z_i + t_i \qquad \forall i$$

Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Portfolio example

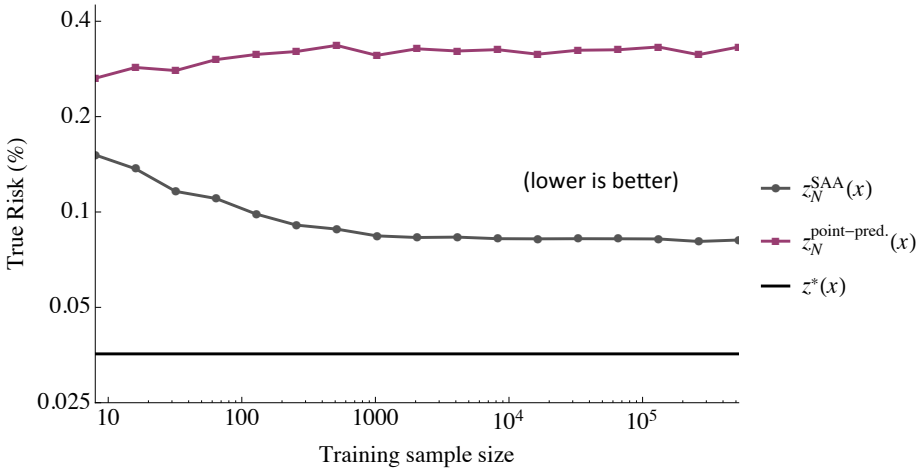
- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns

$$c((z,\beta);y) = \beta + \frac{1}{\epsilon}\max\left\{-z^T y - \beta, 0\right\} - \lambda z^T y$$

$$\mathcal{Z} = \{\beta \in \mathbb{R},\ z \geq 0,\ \sum_{i=1}^{d_z} z_i = 1\}$$

Portfolio example

- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns



Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Case study

Our approach

- Construct predictive prescriptions of the form

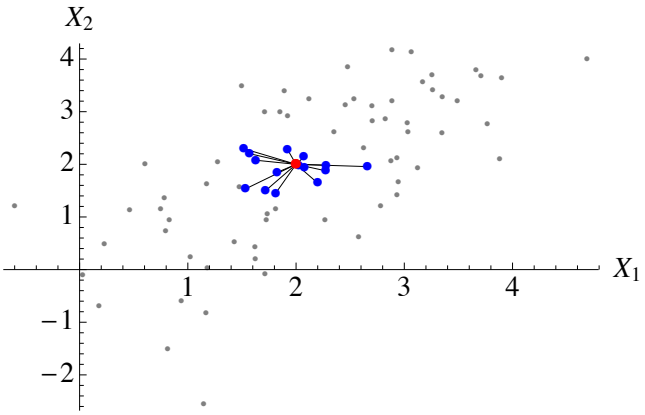
$$\hat{z}_N(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} w_N^i(x) c(z; y^i)$$

- Re-weight Y data using data-driven weights

Thm: if $c(z;y)$ is convex + subgrad oracle, $\mathcal{Z}$ convex + sep oracle, then computing $\hat{z}_N(x)$ possible in polynomial time + oracle calls.

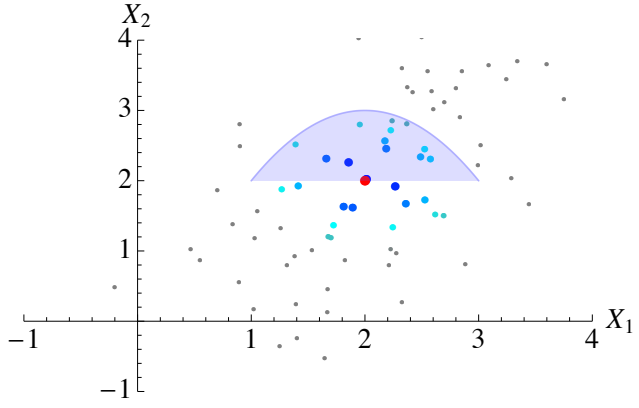
*k*NN

$$\hat{z}_N^{k\mathrm{NN}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{x^i \text{ is } k\mathrm{NN} \text{ of } x} c(z; y^i)$$



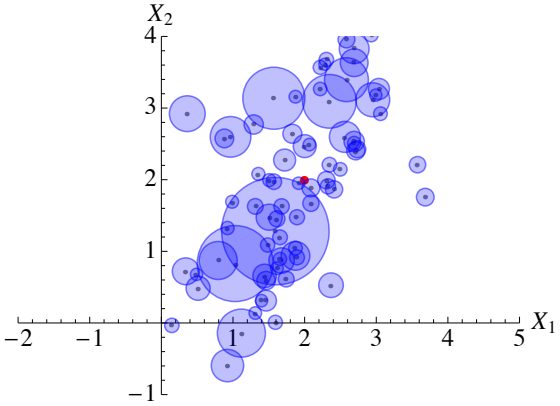
Parzen windows

$$\hat{z}_N^{\mathrm{KR}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} K((x^i - x)/h_N) c(z; y^i)$$



Recursive Parzen windows

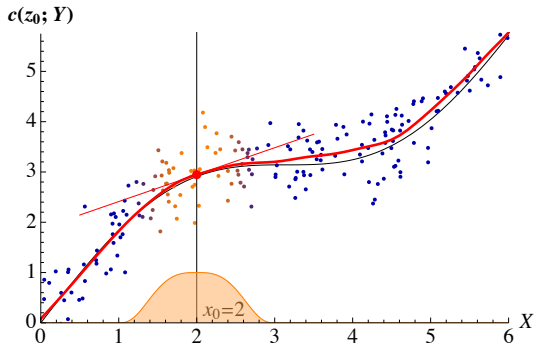
$$\hat{z}_N^{\text{Rec-KR}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} K((x^i - x)/h_i) c(z; y^i)$$



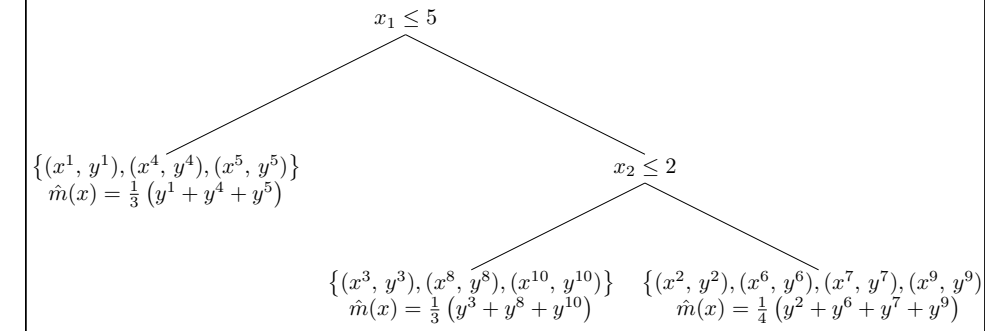
Local linear regression

$$\hat{z}_N^{\text{LOESS}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} k_i(x) \left(1 - \sum_{j=1}^{n} k_j(x)(x^j - x)^T \Xi(x)^{-1}(x^i - x)\right) c(z; y^i)$$

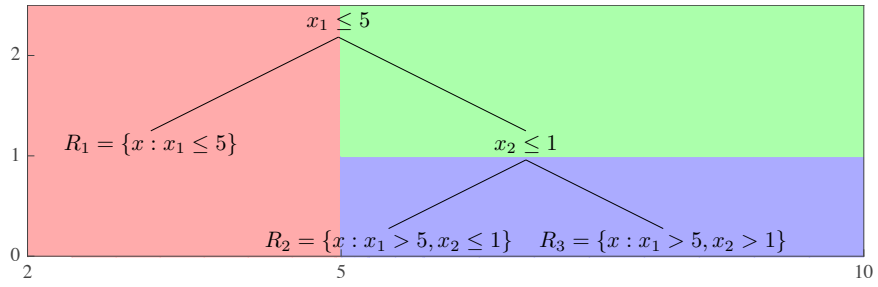
$$\Xi(x) = \sum_{i=1}^{n} k_i(x)(x^i - x)(x^i - x)^T \quad k_i(x) = \left(1 - \left(||x^i - x|| / h_N\right)^3\right)^3 \mathbb{I}\left[||x^i - x|| \leq h_N\right]$$



CART



CART



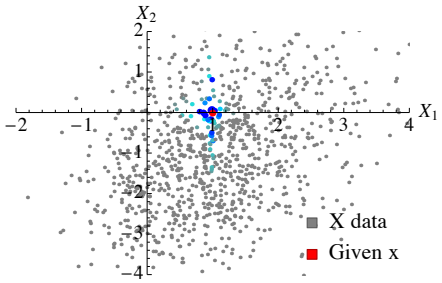
Implied binning rule $R(x) = (j \text{ s.t. } x \in R_j)$

$$\hat{z}_N^{\text{CART}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{\mathcal{R}(x^i) = \mathcal{R}(x)} c(z; y^i)$$

Random Forest

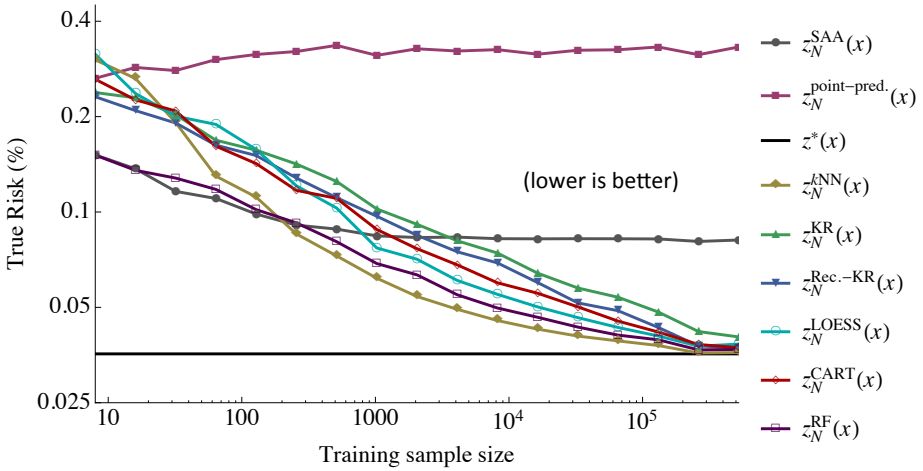
- Train T trees on bootstrapped samples and randomly selected feature subsets
- Get T binning rules $R^t(x) = \left(j \text{ s.t. } x \in R_j^t\right)$

$$\hat{z}_N^{\mathrm{RF}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{t=1}^{T} \frac{1}{|\{j : R^t(x^j) = R^t(x)\}|} \sum_{R^t(x^i) = R^t(x)} c(z; y^i)$$



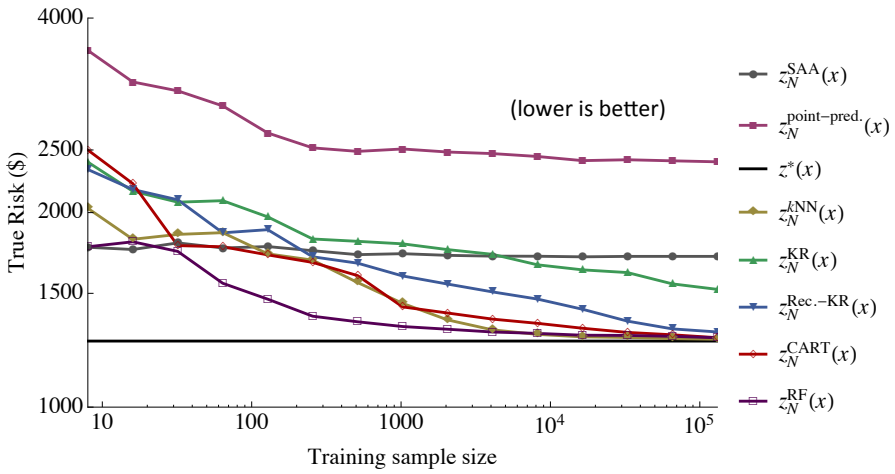
Portfolio example

- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns



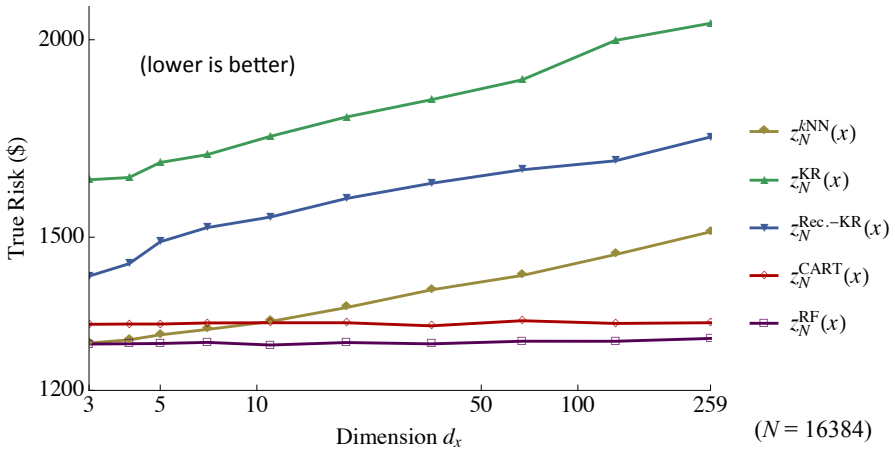
Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Case study

Asymptotic Optimality

- Want

Def: predictive prescription $\hat{z}_N(x)$ is ***asymptotically optimal*** if, with probability 1, for almost everywhere x, as $N \to \infty$

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X=x\right] = \min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x\right]$$

$$L\left(\{\hat{z}_N(x) : N \in \mathbb{N}\}\right) \subset \arg\min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x\right]$$

- Need

Assumption 1: The full-info problem is well defined, i.e.,
$\mathbb{E}\left[|c(z;Y)|\right] < \infty$

Assumption 2: $c(z;y)$ is equicontinuous in z.

Assumption 3: $\mathcal{Z}$ is closed and either (a) also bounded, (b) $c(z;y)$ is coercive, or (c) $c(z;y)$ is convex.

Data collection as a mixing process

- Instead of IID consider a data collection process
  $$(x_1, y_1), (x_2, y_2), \ldots$$
  that is a stationary mixing process
- I.e. as the lag ℓ gets bigger,
  $$(x_1, y_1), \ldots, (x_t, y_t)$$
  and
  $$(x_{t+\ell}, y_{t+\ell}), (x_{t+\ell+1}, y_{t+\ell+1}), \ldots$$
  look more and more independent.
- Encompasses ARMA, GARCH, Markov processes.
- Can represent more realistic data collection from interdependent weekly demands, stock returns, volume of Google searches on a topic, ...

Asymptotic Optimality: *k*NN

$$\hat{z}_N^{k\mathrm{NN}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{x^i \text{ is } k\mathrm{NN} \text{ of } x} c(z; y^i)$$

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is IID, and $k = \min\{\lceil CN^\delta \rceil, N-1\}$ with $0 < \delta < 1$.
Then $\hat{z}_N^{k\mathrm{NN}}(x)$ is asymptotically optimal.

Asymptotic Optimality: Parzen Windows

$$\hat{z}_N^{\mathrm{KR}}(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} K((x^i - x)/h_N)c(z; y^i)$$

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is mixing, costs satisfy $\mathbb{E}\left[|c(z;Y)|\left(\log|c(z;Y)|\right)_+\right] < \infty$, K is one of given kernels, and $h_N = CN^{-\delta}$, $0 < \delta < 1/d_x$. Then $\hat{z}_N^{\mathrm{KR}}(x)$ is asymptotically optimal.

Asymptotic Optimality: Recursive Parzen Windows

$$\hat{z}_N^{\text{Rec-KR}}(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} K((x^i - x)/h_i)c(z; y^i)$$

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is mixing, K is one of given kernels, and $h_i = Ci^{-\delta}$, $0 < \delta < 1/(2d_x)$. Then $\hat{z}_N^{\text{Rec-KR}}(x)$ is asymptotically optimal.

Asymptotic Optimality: LOESS

$$\hat{z}_N^{\mathrm{LOESS}}(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} k_i(x)\left(1-\sum_{j=1}^{n} k_j(x)(x^j-x)^T\Xi(x)^{-1}(x^i-x)\right)c(z;y^i)$$

$$\Xi(x)=\sum_{i=1}^{n} k_i(x)(x^i-x)(x^i-x)^T \quad k_i(x)=\left(1-\left(||x^i-x||/h_N\right)^3\right)^3 \mathbb{I}\left[||x^i-x||\le h_N\right]$$

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is mixing, μ_X abs. continuous, costs bounded $|c(z;y)| \le g(z)$, and $h_N = CN^{-\delta}$, $0 < \delta < 1/d_x$.
Then $\hat{z}_N^{\mathrm{LOESS}}(x)$ is asymptotically optimal.

Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Case study

Value of a Prescription

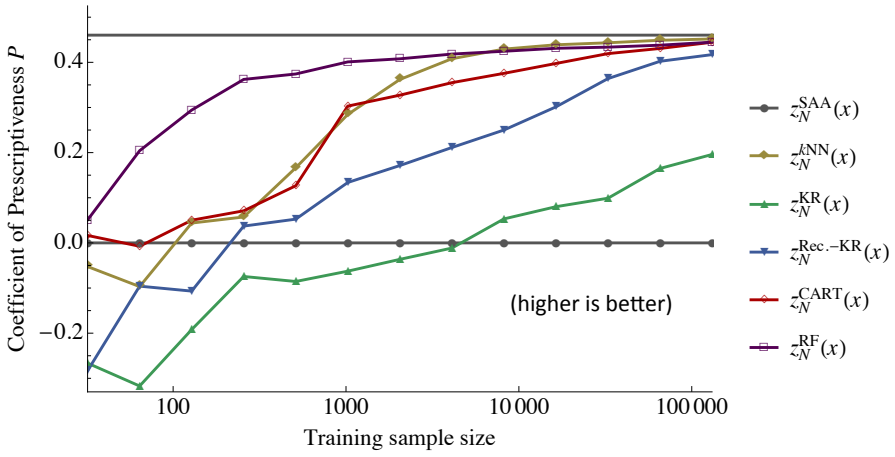
- *Coefficient of Prescriptiveness*

$$P = \frac{\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} c(z; y^i) - \sum_{i=1}^{N} c(\hat{z}_N(x^i); y^i)}{\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} c(z; y^i) - \sum_{i=1}^{N} \min_{z\in\mathcal{Z}} c(z; y^i)} \quad \begin{matrix} \leq 1 \\ \to [0,1] \end{matrix}$$

- Measures the prescriptive value of X and of the of the prescription trained
- In particular when measured OoS

Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Case study

Back to our media distribution application

- Recall: want to maximize number of items sold.
- Focus on video media, Europe

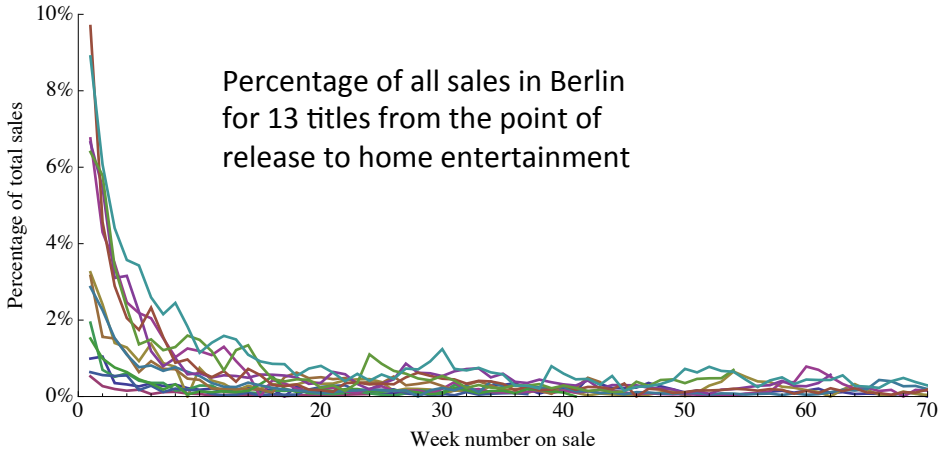
$$\max \quad \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j,\, z_{trj}\} \,\middle|\, X = x_{tr}\right]$$

$$\text{s.t.} \quad \sum_{j=1}^{d} z_{trj} \leq K_r$$

$$z_{trj} \geq 0 \qquad \forall j = 1, \ldots, d$$

Internal Company Data

- Sales by item/location, 2010 to present
- ~50GB *after* aggregating transaction records by week



Dealing with Censored Data

- Observe sales, not demand (quantity of interest Y)

$$U = \min\{Y, V\}$$

- Adjust weights for right-censored data

$$\tilde{w}_{N,(i)}(x) = \begin{cases} \left(\frac{w_{N,(i)}(x)}{\sum_{\ell=i}^{N} w_{N,(\ell)}(x)}\right) \prod_{k \le i-1 \,:\, u^{(k)} < v^{(k)}} \left(\frac{\sum_{\ell=k+1}^{N} w_{N,(\ell)}(x)}{\sum_{\ell=k}^{N} w_{N,(\ell)}(x)}\right) & \text{if } u^{(i)} < v^{(i)}, \\ 0 & \text{otherwise.} \end{cases}$$

Thm: Under same assumptions as before and if in addition (a) Y and V conditionally independent given X, (b) Y and V share no atoms, and (c) upper support of V greater than that of Y given $X = x$, then $\hat{z}_N(x)$ is ***asymptotically optimal***.

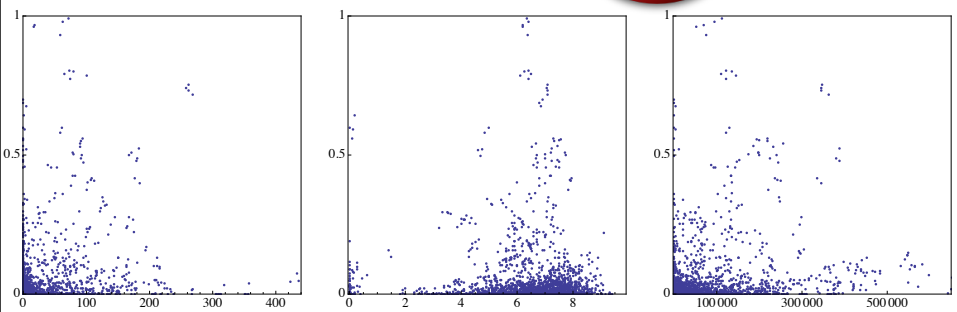
Internal Company Data

- Sales by item/location, 2010 to present
- ~50GB *after* aggregating transaction records by week
- Location info:
  - Address
    - Google Geocoding API
- Item info:
  - Medium (DVD/BLU)
  - Obfuscated title
    - Disambiguation

Beyond internal company data: Harvesting public data (more X)

- **Movie/series**
- **Actors** (find actor communities; Blondel et al 2008)
- **Plot summary** (cosine similarities, hierarchically clustered)
- **Box office gross, US**
- **Oscar wins and nominations** and other awards
- **Professional (meta-)ratings**, **user ratings**
- **Num of user ratings**
- **Genre** (can be multiple)
- **MPAA rating**

Beyond internal company data: Harvesting public data (more X)

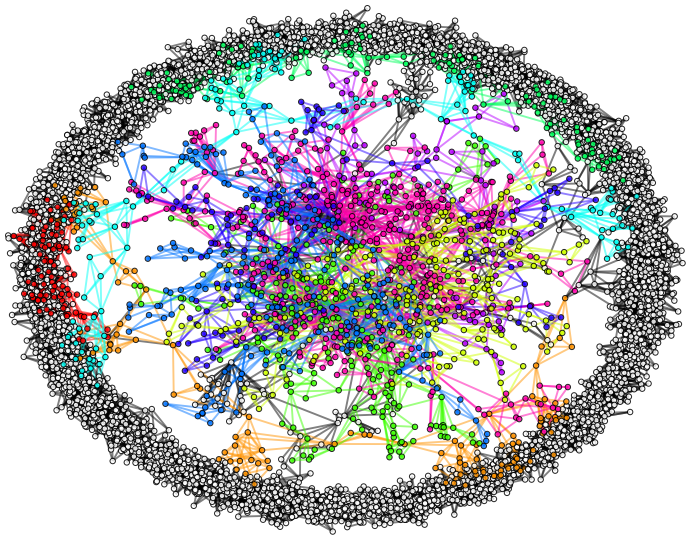


Box office gross
ρ = 0.32

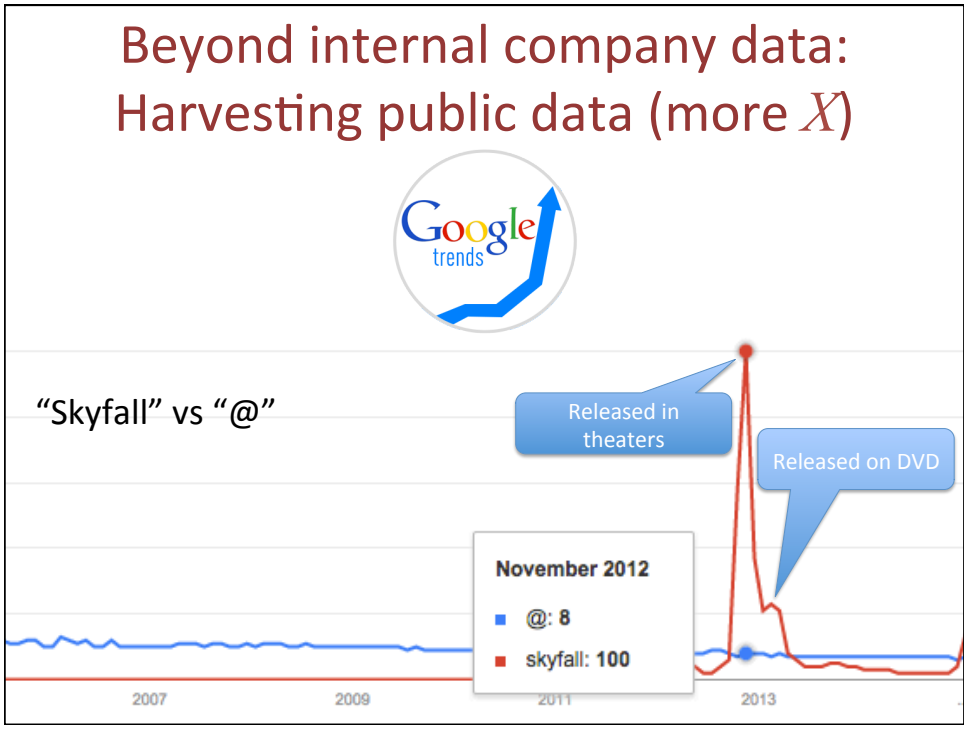
IMDb rating
ρ = 0.02

Number user votes
ρ = 0.25

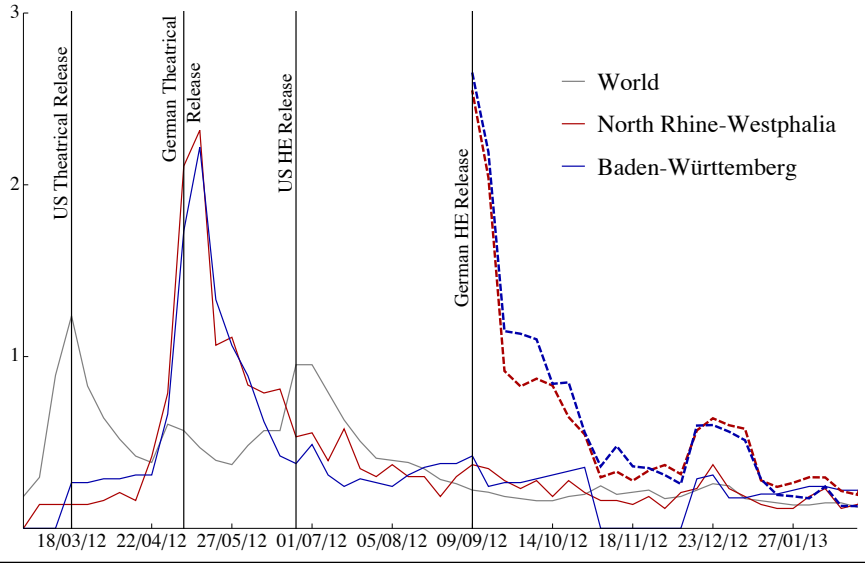
Beyond internal company data: Harvesting public data (more X)



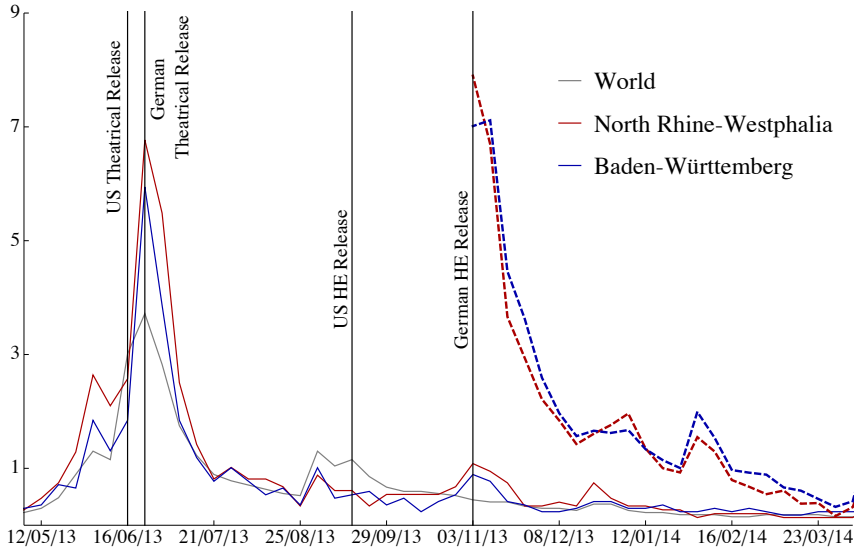
Beyond internal company data: Harvesting public data (more X)



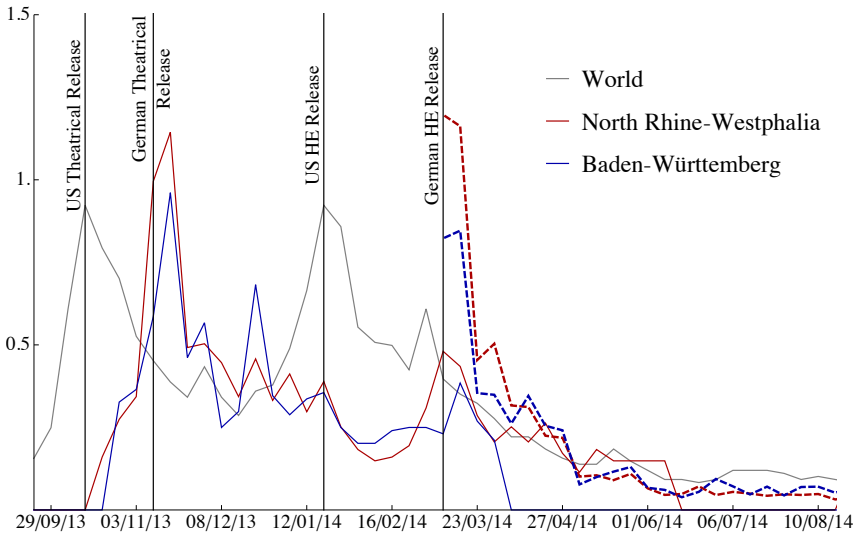
Beyond internal company data: Harvesting public data (more X)



Beyond internal company data: Harvesting public data (more X)

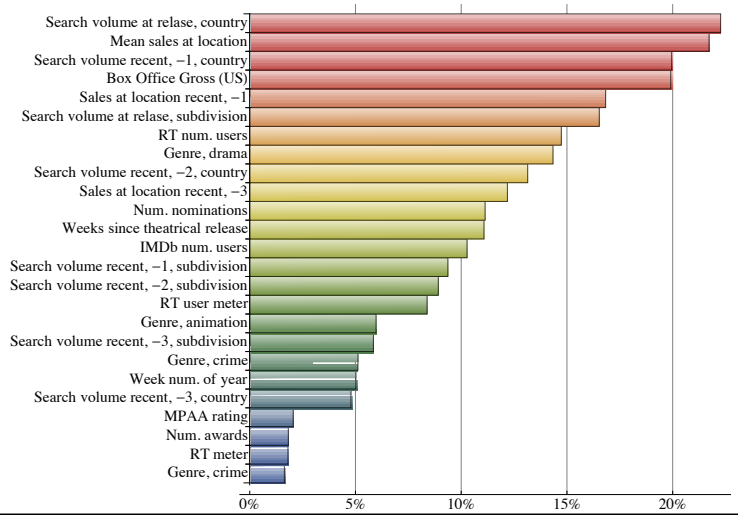


Beyond internal company data: Harvesting public data (more X)



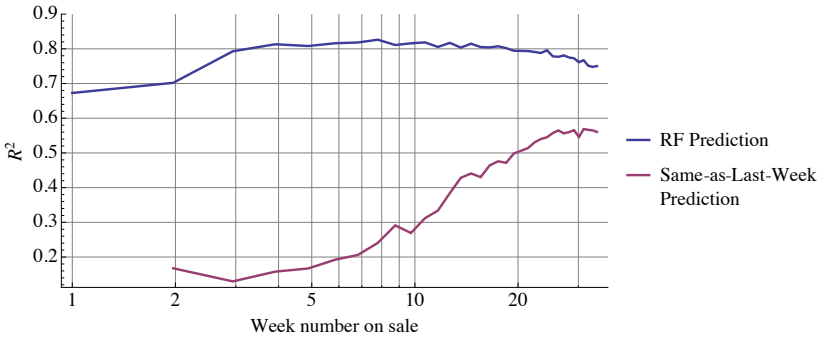
Predicting Demand

- Random forest regressor
- ***New*** titles: out-of-sample R^2 = 0.67



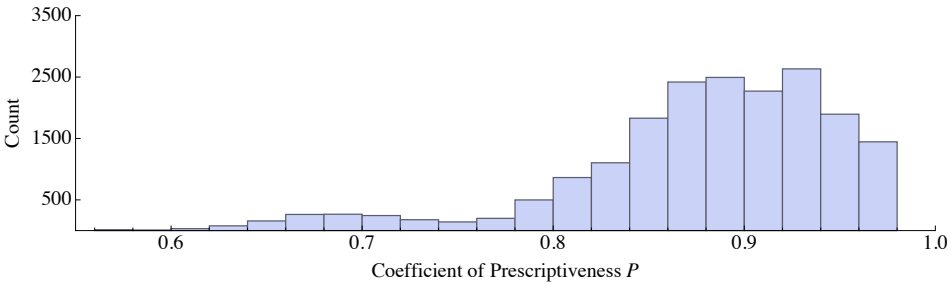
Predicting Demand

- Random forest regressor
- ***New*** titles: out-of-sample R^2 = 0.67



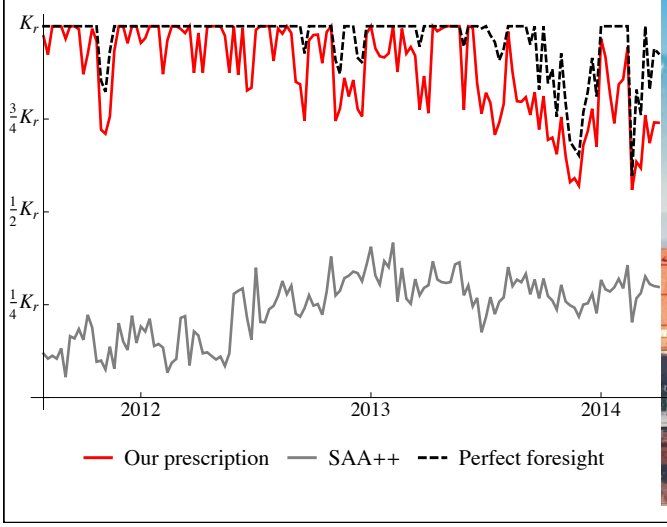
Prescribing Order Quantities

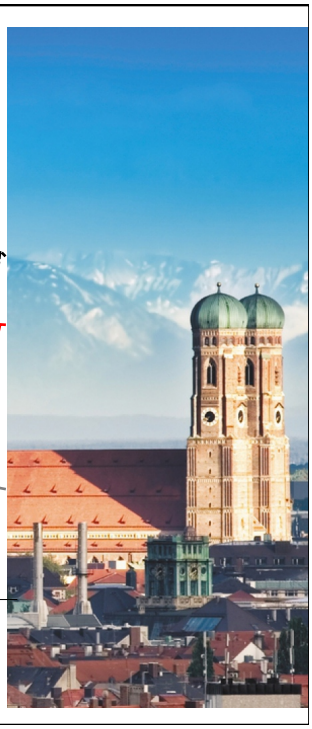
- Construct a predictive prescription based on our random forest…
  - ***Out-of-sample***
    P = 0.88

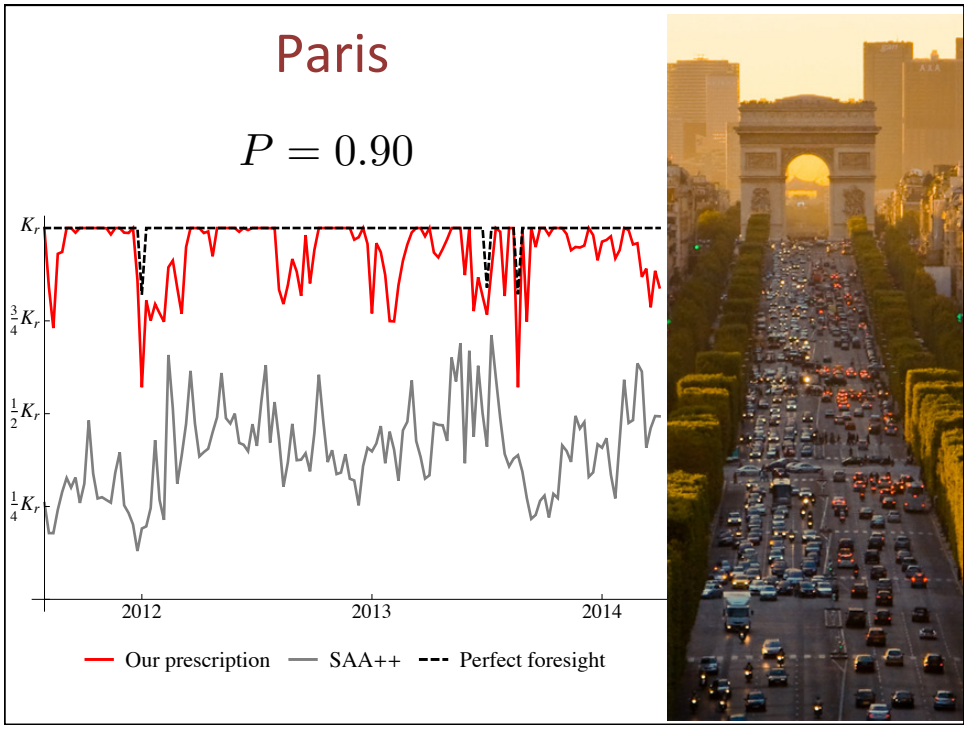


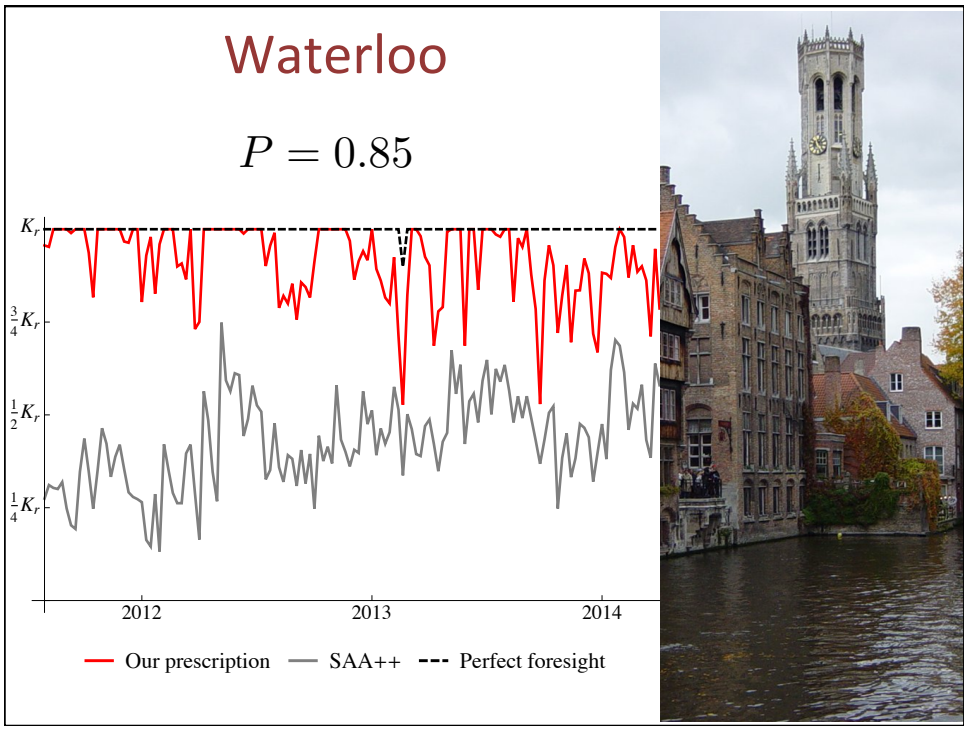
Munich

$P = 0.89$



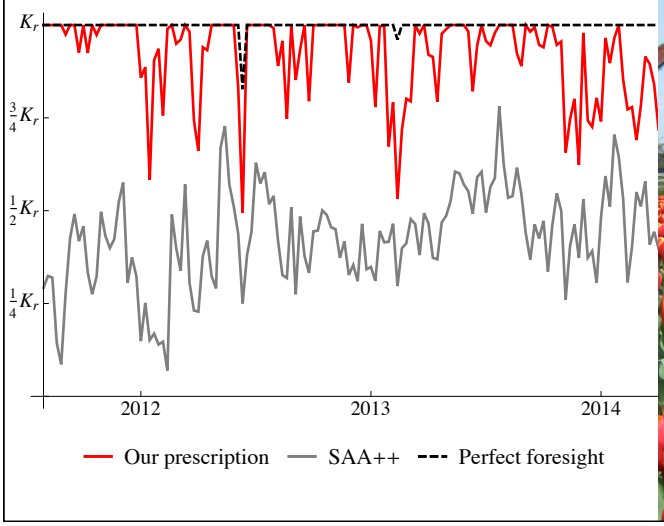



Paris
$P = 0.90$
K_r
$\frac{3}{4}K_r$
$\frac{1}{2}K_r$
$\frac{1}{4}K_r$
2012
2013
2014
Our prescription
SAA++
Perfect foresight


Waterloo
$P = 0.85$
K_r
$\frac{3}{4}K_r$
$\frac{1}{2}K_r$
$\frac{1}{4}K_r$
2012
2013
2014
Our prescription
SAA++
Perfect foresight

The Hague

$$P = 0.86$$



What we've looked at

- **A new framework**
  - Unifies ML and OR/MS
  - General purpose
- **Theory**
  - Computational tractability
  - Asymptotic optimality
- **Performance metric**
  - Coefficient of prescriptiveness
- **Practice**
  - Case study of huge media distributor
  - Study *prescriptive* power of large-scale data