Convex Optimization Applications

Stephen Boyd Steven Diamond
Junzi Zhang Akshay Agrawal

EE & CS Departments

Stanford University

Outline

Outline

Portfolio allocation vector

- invest fraction w_i in asset i, $i = 1, \ldots, n$
- $w \in \mathbf{R}^n$ is *portfolio allocation vector*
- $\mathbf{1}^T w = 1$
- $w_i < 0$ means a *short position* in asset i
  (borrow shares and sell now; must replace later)
- $w \geq 0$ is a *long only* portfolio
- $\|w\|_1 = \mathbf{1}^T w_+ + \mathbf{1}^T w_-$ is *leverage*
  (many other definitions used ...)

Asset returns

- investments held for one period
- initial prices $p_i > 0$; end of period prices $p_i^+ > 0$
- asset (fractional) returns $r_i = (p_i^+ - p_i)/p_i$
- portfolio (fractional) return $R = r^T w$
- common model: r is a random variable, with mean $\mathbf{E}\, r = \mu$, covariance $\mathbf{E}(r-\mu)(r-\mu)^T = \Sigma$
- so R is a RV with $\mathbf{E}\, R = \mu^T w$, $\mathbf{var}(R) = w^T \Sigma w$
- $\mathbf{E}\, R$ is (mean) *return* of portfolio
- $\mathbf{var}(R)$ is *risk* of portfolio
  (risk also sometimes given as $\mathbf{std}(R) = \sqrt{\mathbf{var}(R)}$)
- two objectives: high return, low risk

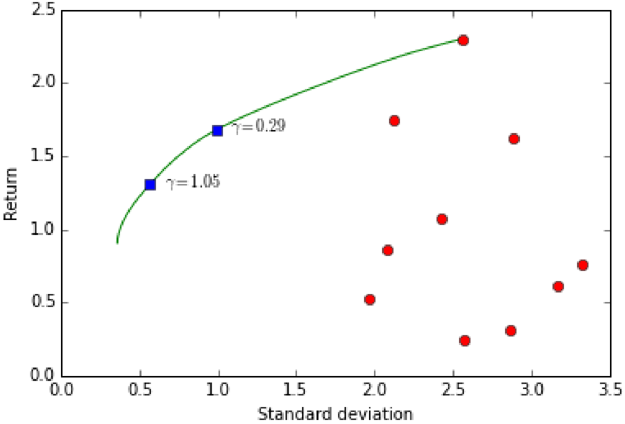
Classical (Markowitz) portfolio optimization

$$\begin{array}{ll} \text{maximize} & \mu^T w - \gamma w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1, \quad w \in \mathcal{W} \end{array}$$

- variable $w \in \mathbf{R}^n$
- $\mathcal{W}$ is set of allowed portfolios
- common case: $\mathcal{W} = \mathbf{R}^n_+$ (long only portfolio)
- $\gamma > 0$ is the *risk aversion parameter*
- $\mu^T w - \gamma w^T \Sigma w$ is *risk-adjusted return*
- varying γ gives optimal *risk-return trade-off*
- can also fix return and minimize risk, *etc.*

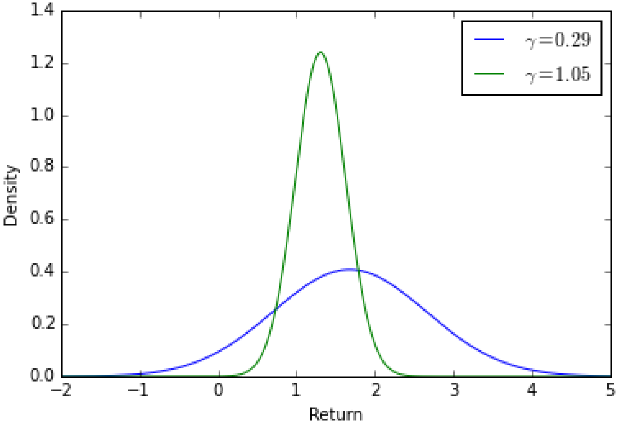
Example

optimal risk-return trade-off for 10 assets, long only portfolio



Example

return distributions for two risk aversion values



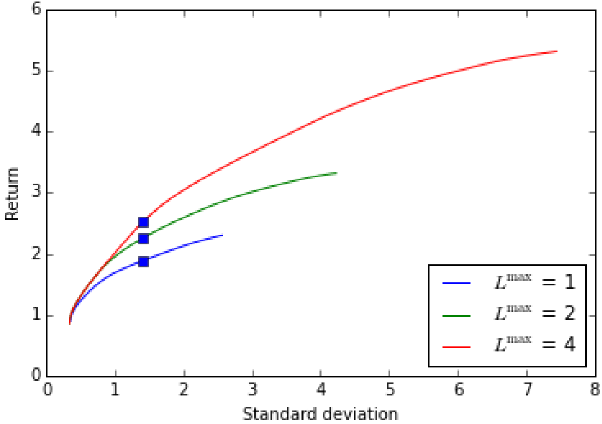
Portfolio constraints

- $\mathcal{W} = \mathbf{R}^n$ (simple analytical solution)
- leverage limit: $\|w\|_1 \leq L^{\max}$
- *market neutral*: $m^T \Sigma w = 0$
  - m_i is capitalization of asset i
  - $M = m^T r$ is *market return*
  - $m^T \Sigma w = \mathbf{cov}(M, R)$

  i.e., market neutral portfolio return is uncorrelated with market return

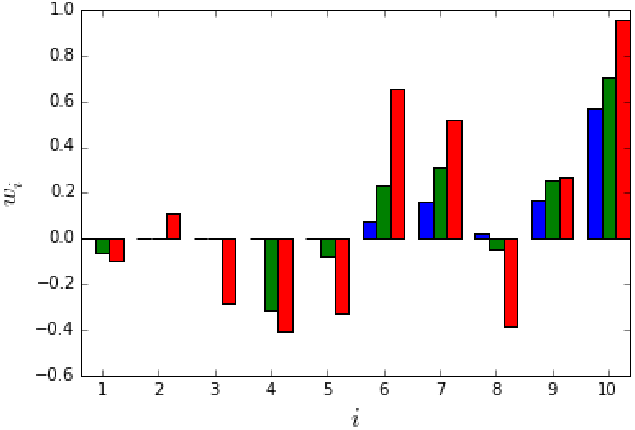
Example

optimal risk-return trade-off curves for leverage limits $1, 2, 4$



Example

three portfolios with $w^T \Sigma w = 2$, leverage limits $L = 1, 2, 4$



Variations

- require $\mu^T w \geq R^{\min}$, minimize $w^T \Sigma w$ or $\|\Sigma^{1/2} w\|_2$
- include (broker) cost of short positions,

$$s^T (w)_-, \quad s \geq 0$$

- include transaction cost (from previous portfolio w^{prev}),

$$\kappa^T |w - w^{\mathrm{prev}}|^\eta, \quad \kappa \geq 0$$

common models: $\eta = 1,\ 3/2,\ 2$

Factor covariance model

$$\Sigma = F\tilde{\Sigma}F^T + D$$

- $F \in \mathbf{R}^{n\times k}$, $k \ll n$ is *factor loading matrix*
- k is number of factors (or sectors), typically 10s
- F_{ij} is loading of asset i to factor j
- D is diagonal matrix; $D_{ii} > 0$ is *idiosyncratic risk*
- $\tilde{\Sigma} > 0$ is the *factor covariance matrix*

- $F^T w \in \mathbf{R}^k$ gives portfolio *factor exposures*
- portfolio is *factor j neutral* if $(F^T w)_j = 0$

Portfolio optimization with factor covariance model

$$
\begin{array}{ll}
\text{maximize} & \mu^T w - \gamma \left(f^T \tilde{\Sigma} f + w^T D w \right) \\
\text{subject to} & \mathbf{1}^T w = 1, \quad f = F^T w \\
& w \in \mathcal{W}, \quad f \in \mathcal{F}
\end{array}
$$

- variables $w \in \mathbf{R}^n$ (allocations), $f \in \mathbf{R}^k$ (factor exposures)
- $\mathcal{F}$ gives factor exposure constraints

- computational advantage: $O(nk^2)$ vs. $O(n^3)$

Example

- 50 factors, 3000 assets
- leverage limit $= 2$
- solve with covariance given as
  - single matrix
  - factor model
- CVXPY/OSQP single thread time

covariance	solve time
single matrix	173.30 sec
factor model	0.85 sec

Outline

Covariance uncertainty

- single period Markowitz portfolio allocation problem
- we have fixed portfolio allocation $w \in \mathbf{R}^n$
- return covariance Σ not known, but we believe $\Sigma \in \mathcal{S}$
- $\mathcal{S}$ is convex set of possible covariance matrices
- risk is $w^T \Sigma w$, a *linear function of* Σ

Worst-case risk analysis

- what is the worst (maximum) risk, over all possible covariance matrices?
- worst-case risk analysis problem:

$$\begin{array}{ll} \text{maximize} & w^T \Sigma w \\ \text{subject to} & \Sigma \in \mathcal{S}, \quad \Sigma \succeq 0 \end{array}$$

  with variable Σ
- ... a convex problem with variable Σ

- if the worst-case risk is not too bad, you can worry less
- if not, you'll confront your worst nightmare

Example

- $w = (-0.01, 0.13, 0.18, 0.88, -0.18)$
- optimized for Σ^{nom}, return 0.1, leverage limit 2
- $\mathcal{S} = \{\Sigma^{\mathrm{nom}} + \Delta \, : \, |\Delta_{ii}| = 0, \; |\Delta_{ij}| \leq 0.2\}$,

$$\Sigma^{\mathrm{nom}} = \begin{bmatrix} 0.58 & 0.2 & 0.57 & -0.02 & 0.43 \\ 0.2 & 0.36 & 0.24 & 0 & 0.38 \\ 0.57 & 0.24 & 0.57 & -0.01 & 0.47 \\ -0.02 & 0 & -0.01 & 0.05 & 0.08 \\ 0.43 & 0.38 & 0.47 & 0.08 & 0.92 \end{bmatrix}$$

Example

- nominal risk $= 0.168$
- worst case risk $= 0.422$

$$\text{worst case } \Delta = \begin{bmatrix} 0 & 0.04 & -0.2 & -0. & 0.2 \\ 0.04 & 0 & 0.2 & 0.09 & -0.2 \\ -0.2 & 0.2 & 0 & 0.12 & -0.2 \\ -0. & 0.09 & 0.12 & 0 & -0.18 \\ 0.2 & -0.2 & -0.2 & -0.18 & 0 \end{bmatrix}$$

Outline

Ad display

- m advertisers/ads, $i = 1, \ldots, m$
- n time slots, $t = 1, \ldots, n$
- T_t is total traffic in time slot t
- $D_{it} \geq 0$ is number of ad i displayed in period t
- $\sum_i D_{it} \leq T_t$
- contracted minimum total displays: $\sum_t D_{it} \geq c_i$
- goal: choose D_{it}

Clicks and revenue

- C_{it} is number of clicks on ad i in period t
- click model: $C_{it} = P_{it}D_{it}$, $P_{it} \in [0,1]$
- payment: $R_i > 0$ per click for ad i, up to budget B_i
- ad revenue

$$S_i = \min\left\{R_i \sum_t C_{it}, B_i\right\}$$

  . . . a concave function of D

Ad optimization

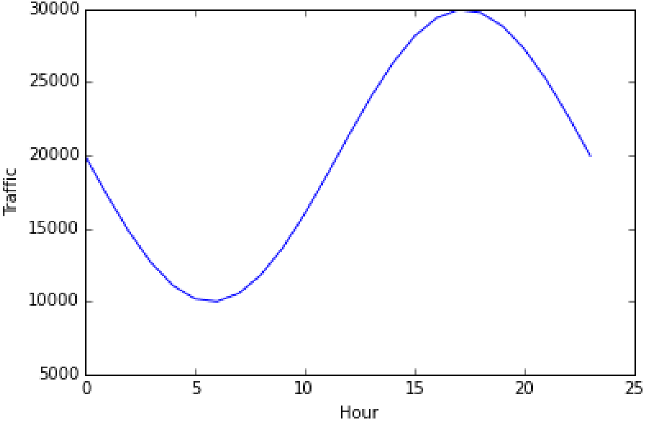
- choose displays to maximize revenue:

$$\begin{array}{ll} \text{maximize} & \sum_i S_i \\ \text{subject to} & D \geq 0, \quad D^T \mathbf{1} \leq T, \quad D\mathbf{1} \geq c \end{array}$$

- variable is $D \in \mathbf{R}^{m \times n}$
- data are T, c, R, B, P

Example

- 24 hourly periods, 5 ads (A–E)
- total traffic:



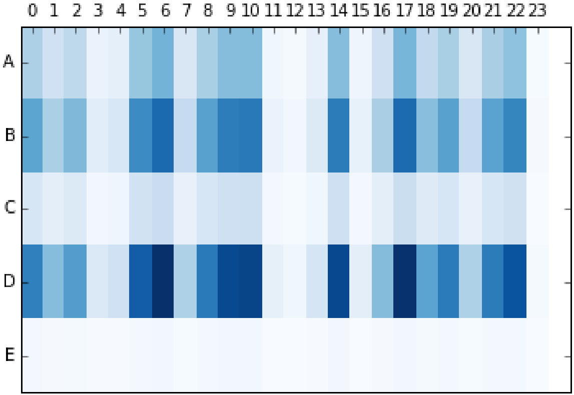
Example

- ad data:

Ad	A	B	C	D	E
c_i	61000	80000	61000	23000	64000
R_i	0.15	1.18	0.57	2.08	2.43
B_i	25000	12000	12000	11000	17000

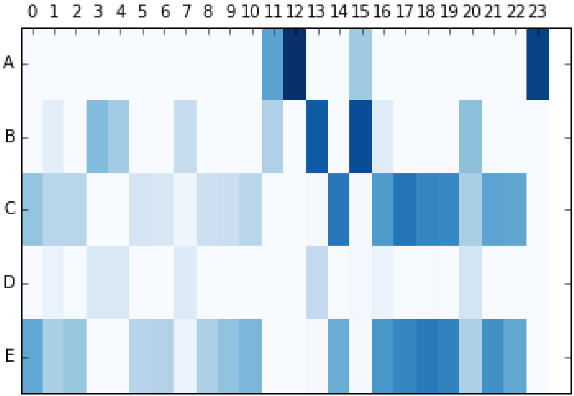
Example

P_{it}



Example

optimal D_{it}



Example

ad revenue

Ad	A	B	C	D	E
c_i	61000	80000	61000	23000	64000
R_i	0.15	1.18	0.57	2.08	2.43
B_i	25000	12000	12000	11000	17000
$\sum_t D_{it}$	61000	80000	148116	23000	167323
S_i	182	12000	12000	11000	7760

Outline

Standard regression

- given data $(x_i, y_i) \in \mathbf{R}^n \times \mathbf{R}$, $i = 1, \ldots, m$
- fit linear (affine) model $\hat{y}_i = \beta^T x_i - v$, $\beta \in \mathbf{R}^n$, $v \in \mathbf{R}$
- residuals are $r_i = \hat{y}_i - y_i$
- least-squares: choose β, v to minimize $\|r\|_2^2 = \sum_i r_i^2$
- mean of optimal residuals is zero
- can add (Tychonov) regularization: with $\lambda > 0$,

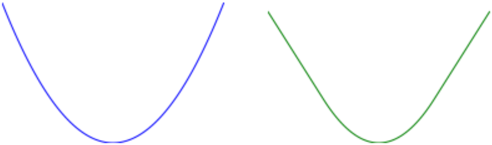
$$\text{minimize} \quad \|r\|_2^2 + \lambda \|\beta\|_2^2$$

Robust (Huber) regression

- replace square with *Huber function*

$$\phi(u) = \begin{cases} u^2 & |u| \leq M \\ 2Mu - M^2 & |u| > M \end{cases}$$

$M > 0$ is the Huber threshold



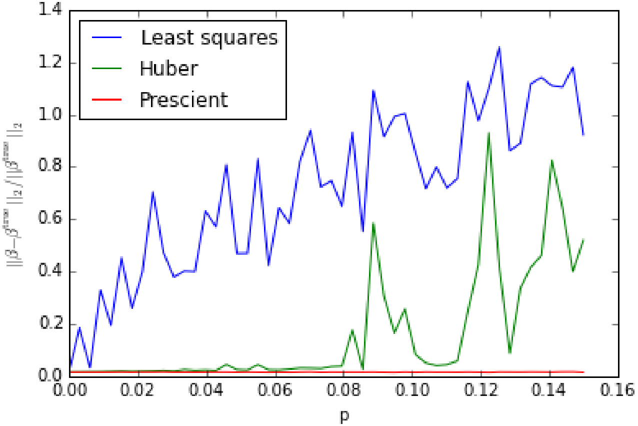
- same as least-squares for small residuals, but allows (some) large residuals

Example

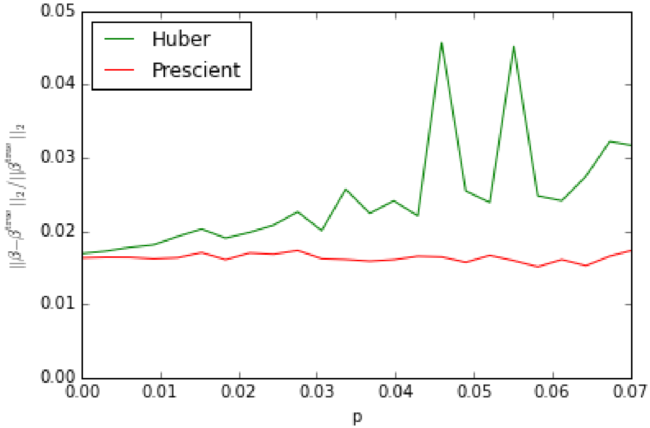
- $m = 450$ measurements, $n = 300$ regressors
- choose β^{true}; $x_i \sim \mathcal{N}(0, I)$
- set $y_i = (\beta^{\mathrm{true}})^T x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$
- with probability p, replace y_i with $-y_i$
- data has fraction p of (non-obvious) wrong measurements
- distribution of 'good' and 'bad' y_i are the same
- try to recover $\beta^{\mathrm{true}} \in \mathbf{R}^n$ from measurements $y \in \mathbf{R}^m$
- 'prescient' version: we know which measurements are wrong

Example

50 problem instances, p varying from 0 to 0.15



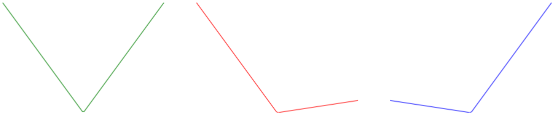
Example


Huber
Prescient
0.05
0.04
0.03
0.02
0.01
0.00
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
p

Quantile regression

- *tilted ℓ_1 penalty*: for $\tau \in (0,1)$,

$$\phi(u) = \tau(u)_+ + (1-\tau)(u)_- = (1/2)|u| + (\tau - 1/2)u$$



- *quantile regression*: choose β, v to minimize $\sum_i \phi(r_i)$
- $\tau = 0.5$: equal penalty for over- and under-estimating
- $\tau = 0.1$: $9\times$ more penalty for under-estimating
- $\tau = 0.9$: $9\times$ more penalty for over-estimating

Quantile regression

- for $r_i \neq 0$,

$$\frac{\partial \sum_i \phi(r_i)}{\partial v} = \tau \left|\{i : r_i > 0\}\right| - (1-\tau) \left|\{i : r_i < 0\}\right|$$

- (roughly speaking) for optimal v we have

$$\tau \left|\{i : r_i > 0\}\right| = (1-\tau) \left|\{i : r_i < 0\}\right|$$

- and so for optimal v, $\tau m = \left|\{i : r_i < 0\}\right|$
- τ-quantile of optimal residuals is zero
- hence the name quantile regression

Example

- time series x_t, $t = 0, 1, 2, \ldots$
- auto-regressive predictor:

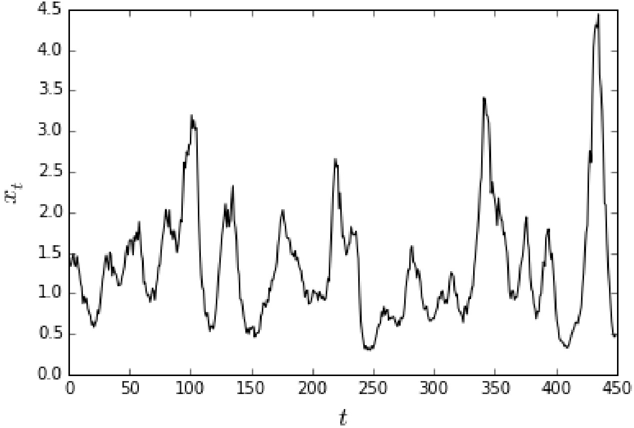
$$\hat{x}_{t+1} = \beta^T(x_t, \ldots, x_{t-M}) - v$$

- $M = 10$ is memory of predictor
- use quantile regression for $\tau = 0.1, 0.5, 0.9$
- at each time t, gives three one-step-ahead predictions:

$$\hat{x}^{0.1}_{t+1}, \qquad \hat{x}^{0.5}_{t+1}, \qquad \hat{x}^{0.9}_{t+1}$$

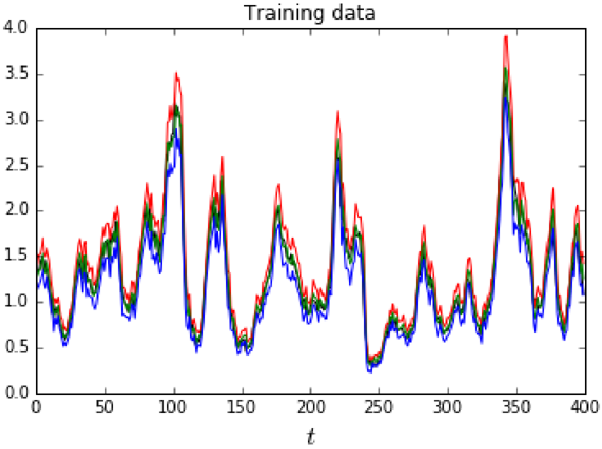
Example

time series x_t



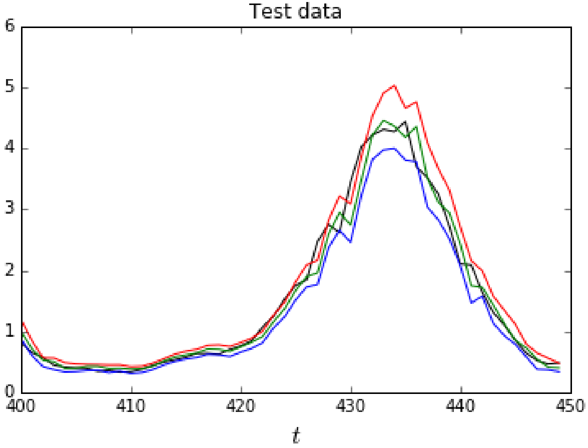
Example

x_t and predictions $\hat{x}_{t+1}^{0.1}$, $\hat{x}_{t+1}^{0.5}$, $\hat{x}_{t+1}^{0.9}$ (training set, $t = 0, \ldots, 399$)



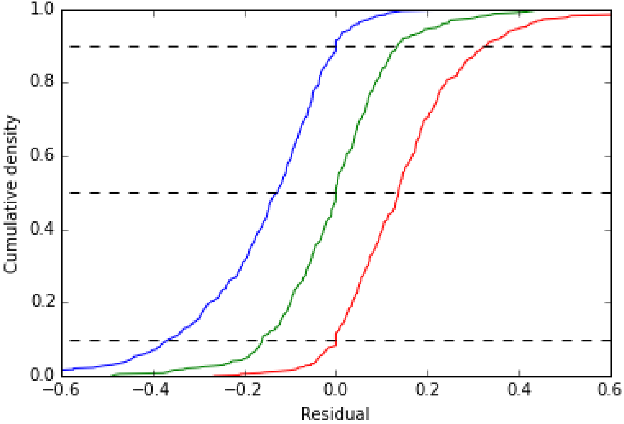
Example

x_t and predictions $\hat{x}^{0.1}_{t+1}$, $\hat{x}^{0.5}_{t+1}$, $\hat{x}^{0.9}_{t+1}$ (test set, $t = 400, \ldots, 449$)



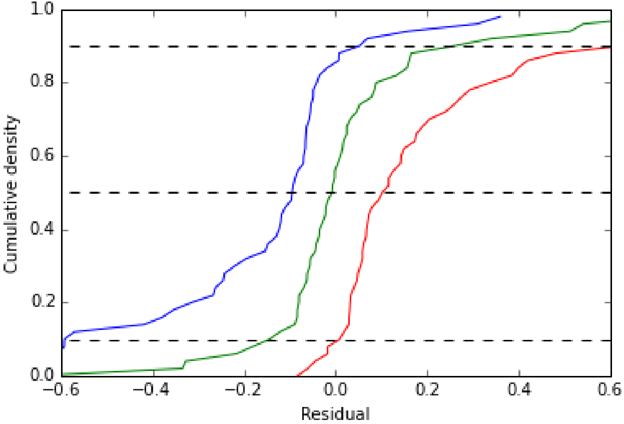
Example

residual distributions for $\tau = 0.9$, 0.5, and 0.1 (training set)



Example

residual distributions for $\tau = 0.9$, 0.5, and 0.1 (test set)



Outline

Data model

- given data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, m$
- for $\mathcal{X} = \mathbf{R}^n$, x is *feature vector*
- for $\mathcal{Y} = \mathbf{R}$, y is (real) *outcome* or *label*
- for $\mathcal{Y} = \{-1, 1\}$, y is (boolean) outcome

- find *model* or *predictor* $\psi : \mathcal{X} \to \mathcal{Y}$ so that $\psi(x) \approx y$ *for data* (x, y) *that you haven't seen*
- for $\mathcal{Y} = \mathbf{R}$, ψ is a *regression model*
- for $\mathcal{Y} = \{-1, 1\}$, ψ is a *classifier*
- we choose ψ based on observed data, prior knowledge

Loss minimization model

- data model parametrized by $\theta \in \mathbf{R}^n$
- *loss function* $L : \mathcal{X} \times \mathcal{Y} \times \mathbf{R}^n \to \mathbf{R}$
- $L(x_i, y_i, \theta)$ is loss (miss-fit) for data point (x_i, y_i), using model parameter θ
- choose θ; then model is

$$\psi(x) = \operatorname*{argmin}_y L(x, y, \theta)$$

Model fitting via regularized loss minimization

- *regularization* $r : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$
- $r(\theta)$ measures model complexity, enforces constraints, or represents prior
- choose θ by minimizing *regularized loss*

$$(1/m) \sum_i L(x_i, y_i, \theta) + r(\theta)$$

- for many useful cases, this is a convex problem
- model is $\psi(x) = \mathrm{argmin}_y \, L(x, y, \theta)$

Examples

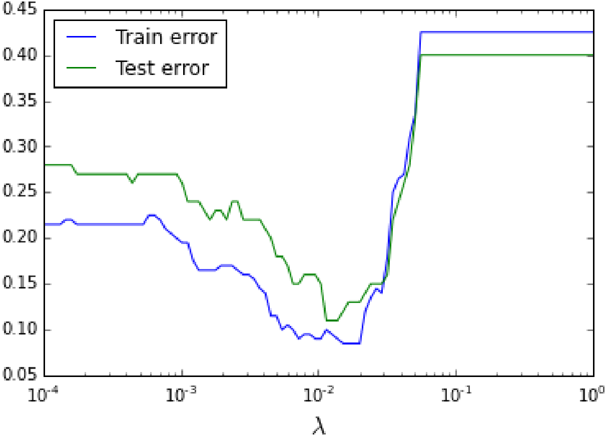
model	$L(x,y,\theta)$	$\psi(x)$	$r(\theta)$
least-squares	$(\theta^T x - y)^2$	$\theta^T x$	0
ridge regression	$(\theta^T x - y)^2$	$\theta^T x$	$\lambda\|\theta\|_2^2$
lasso	$(\theta^T x - y)^2$	$\theta^T x$	$\lambda\|\theta\|_1$
logistic classifier	$\log(1 + \exp(-y\theta^T x))$	$\mathrm{sign}(\theta^T x)$	0
SVM	$(1 - y\theta^T x)_+$	$\mathrm{sign}(\theta^T x)$	$\lambda\|\theta\|_2^2$

- $\lambda > 0$ scales regularization
- all lead to convex fitting problems

Example

- original (boolean) features $z \in \{0, 1\}^{10}$
- (boolean) outcome $y \in \{-1, 1\}$
- new feature vector $x \in \{0, 1\}^{55}$ contains all products $z_i z_j$ (co-occurence of pairs of original features)
- use logistic loss, ℓ_1 regularizer
- training data has $m = 200$ examples; test on 100 examples

Example



Train error
Test error
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
λ

Example

selected features $z_i z_j$, $\lambda = 0.01$