Decision-based model selection

Arnoud den Boer

joint work with Dirk Sierag

Workshop on Data Driven Operations Management
Eindhoven, October 24, 2016

When are

simple & simplified models

preferable to

complex & realistic models?

When are

simple & simplified models

preferable to

complex & realistic models?

Trade-off between modeling, estimation, and optimization errors.

Many criteria exist to select a model

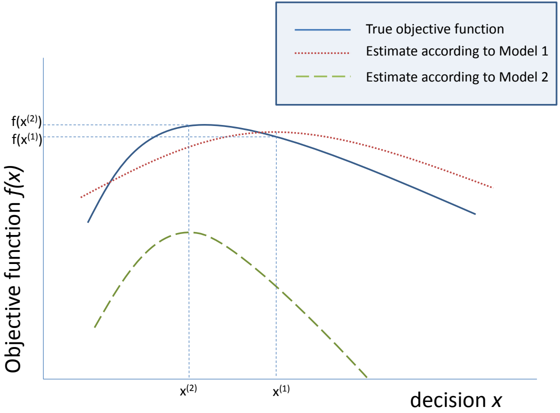
Many criteria exist to select a model

- Akaike information criterion
- Bayesian information criterion
- Deviance information criterion
- Hypothesis testing
- ...

Many criteria exist to select a model

- Akaike information criterion
- Bayesian information criterion
- Deviance information criterion
- Hypothesis testing
- . . .

Based on ‘goodness-of-fit’ on current or future data.



True objective function
Estimate according to Model 1
Estimate according to Model 2
f(x(2))
f(x(1))
Objective function f(x)
x(2)
x(1)
decision x

Models in data-driven optimization

Models in data-driven optimization

$$\max_{x \in \mathcal{X}} \mathbb{E}[r(x, Y(x))] = \max_{x \in \mathcal{X}} \int r(x, y) \mathrm{d}F_{Y(x)}(y)$$

Models in data-driven optimization

$$\max_{x \in \mathcal{X}} \mathbb{E}[r(x, Y(x))] = \max_{x \in \mathcal{X}} \int r(x, y) \mathrm{d}F_{Y(x)}(y)$$

x decision variable, $\mathcal{X}$ admissible decisions
$Y(x)$ random variable, supported on $\mathcal{Y}$.
$F_{Y(x)}$ unknown cdf of Y(x)
$r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ reward function

A model M describes the unknown cdf's $F_{Y(x)}$ for all x:

$$M = \{F_{x,\theta} \text{ cdf on } \mathcal{Y} \mid x \in \mathcal{X},\, \theta \in \Theta\}$$

where Θ may be infinite dimensional.

A model M describes the unknown cdf's $F_{Y(x)}$ for all x:

$$M = \{F_{x,\theta} \text{ cdf on } \mathcal{Y} \mid x \in \mathcal{X},\, \theta \in \Theta\}$$

where Θ may be infinite dimensional.

If $F_{x,\theta^*} \equiv F_{Y(x)}$ for some $\theta^* \in \Theta$, then M is *correctly specified*.

Data is a set $d \in (\mathcal{X} \times \mathcal{Y})^n$, for some $n \in \mathbb{N}$.

Decision x	Observation of $Y(x)$
x_1	y_1
$\vdots$	$\vdots$
x_i	y_i
$\vdots$	$\vdots$
x_n	y_n

An estimator maps data to parameters:

$$\tau : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$$

An estimator maps data to parameters:

$$\tau : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$$

An optimization algorithm maps parameters to decisions

$$\chi : \Theta \to \mathcal{X}$$

An estimator maps data to parameters:

$$\tau : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$$

An optimization algorithm maps parameters to decisions

$$\chi : \Theta \to \mathcal{X}$$

Typically, $\chi(\theta)$ maximizes the reward as function of θ
Sometimes χ is an heuristic.

Model M + data d $\to$ estimate $\tau(d)$ $\to$ decision $\chi(\tau(d))$.

Now we have $K+1$ models

$$M^{(k)} = \{F^{(k)}_{x,\theta} \text{ cdf on } \mathcal{Y} \mid x \in \mathcal{X},\, \theta \in \Theta^{(k)}\}, \quad k = 0, 1, \ldots, K;$$

Now we have $K + 1$ models

$$M^{(k)} = \{F^{(k)}_{x,\theta} \text{ cdf on } \mathcal{Y} \mid x \in \mathcal{X},\, \theta \in \Theta^{(k)}\}, \quad k = 0, 1, \ldots, K;$$

'complex' model $M^{(0)}$ with estimator $\tau^{(0)}$ and algorithm $\chi^{(0)}$,

'simplified' models $M^{(k)}$ with estimator $\tau^{(k)}$ and algorithm $\chi^{(k)}$, $k = 1, \ldots, K$.

Now we have $K+1$ models

$$M^{(k)} = \{F^{(k)}_{x,\theta} \text{ cdf on } \mathcal{Y} \mid x \in \mathcal{X},\, \theta \in \Theta^{(k)}\}, \quad k = 0, 1, \ldots, K;$$

'complex' model $M^{(0)}$ with estimator $\tau^{(0)}$ and algorithm $\chi^{(0)}$,

'simplified' models $M^{(k)}$ with estimator $\tau^{(k)}$ and algorithm $\chi^{(k)}$, $k = 1, \ldots, K$.

and a data set d_0.

Now we have $K + 1$ models

$$M^{(k)} = \{F^{(k)}_{x,\theta} \text{ cdf on } \mathcal{Y} \mid x \in \mathcal{X},\, \theta \in \Theta^{(k)}\}, \quad k = 0, 1, \ldots, K;$$

'complex' model $M^{(0)}$ with estimator $\tau^{(0)}$ and algorithm $\chi^{(0)}$,

'simplified' models $M^{(k)}$ with estimator $\tau^{(k)}$ and algorithm $\chi^{(k)}$, $k = 1, \ldots, K$.

and a data set d_0.

Which model should we use?

Assume $M^{(0)}$ is correctly specified, with true but unknown parameter θ^*.

Assume $M^{(0)}$ is correctly specified, with true but unknown parameter θ^*.

Write $r(x, \theta) = \int_y r(x, y) dF^{(0)}_{x,\theta}(y)$ for the reward function under $M^{(0)}$.

Assume $M^{(0)}$ is correctly specified, with true but unknown parameter θ^*.

Write $r(x, \theta) = \int_y r(x, y) dF^{(0)}_{x,\theta}(y)$ for the reward function under $M^{(0)}$.

Model k suggests decision $x^{(k)} := \chi^{(k)}(\tau^{(k)}(d_0))$, for each k.

Assume $M^{(0)}$ is correctly specified, with true but unknown parameter θ^*.

Write $r(x, \theta) = \int_y r(x, y) dF^{(0)}_{x,\theta}(y)$ for the reward function under $M^{(0)}$.

Model k suggests decision $x^{(k)} := \chi^{(k)}(\tau^{(k)}(d_0))$, for each k.

Goal: estimate

$$\arg\max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}, \theta^*)$$

Goal: estimate

$$\underset{k \in \{0,1,\ldots,K\}}{\arg\max}\ r(x^{(k)}, \theta^*) \tag{1}$$

Goal: estimate

$$\arg\max_{k\in\{0,1,\ldots,K\}} r(x^{(k)}, \theta^*) \tag{1}$$

Naïve estimator

$$\arg\max_{k\in\{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(0)}(d_0))$$

Goal: estimate

$$\underset{k \in \{0,1,\dots,K\}}{\arg\max}\ r(x^{(k)}, \theta^*) \tag{1}$$

Naïve estimator

$$\underset{k \in \{0,1,\dots,K\}}{\arg\max}\ r(x^{(k)}, \tau^{(0)}(d_0))$$

does not work, since typically $x^{(0)}$ maximizes $r(\cdot, \tau^{(0)}(d_0))$.

Goal: estimate

$$\arg\max_{k\in\{0,1,\ldots,K\}} r(x^{(k)}, \theta^*) \tag{1}$$

Naïve estimator

$$\arg\max_{k\in\{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(0)}(d_0))$$

does not work, since typically $x^{(0)}$ maximizes $r(\cdot, \tau^{(0)}(d_0))$.

Our proposal:

Goal: estimate

$$\underset{k\in\{0,1,\ldots,K\}}{\arg\max}\; r(x^{(k)},\theta^*) \tag{1}$$

Naïve estimator

$$\underset{k\in\{0,1,\ldots,K\}}{\arg\max}\; r(x^{(k)},\tau^{(0)}(d_0))$$

does not work, since typically $x^{(0)}$ maximizes $r(\cdot,\tau^{(0)}(d_0))$.

Our proposal:
draw a new data set, according to $\tau^{(0)}(d_0)$;

Goal: estimate

$$\underset{k\in\{0,1,\ldots,K\}}{\arg\max}\ r(x^{(k)},\theta^*) \tag{1}$$

Naïve estimator

$$\underset{k\in\{0,1,\ldots,K\}}{\arg\max}\ r(x^{(k)},\tau^{(0)}(d_0))$$

does not work, since typically $x^{(0)}$ maximizes $r(\cdot,\tau^{(0)}(d_0))$.

Our proposal:

draw a new data set, according to $\tau^{(0)}(d_0)$;

re-estimate θ^* based on the resampled data set;

Goal: estimate

$$\arg\max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}, \theta^*) \tag{1}$$

Naïve estimator

$$\arg\max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(0)}(d_0))$$

does not work, since typically $x^{(0)}$ maximizes $r(\cdot, \tau^{(0)}(d_0))$.

Our proposal:

draw a new data set, according to $\tau^{(0)}(d_0)$;
re-estimate θ^* based on the resampled data set;
replace θ^* in (1) by the re-estimated parameter

Decision-based model selection

Decision-based model selection

- Write $\theta_0 = \tau^{(0)}(x_1, y_1, \ldots, x_n, y_n)$.

Decision-based model selection

- Write $\theta_0 = \tau^{(0)}(x_1, y_1, \ldots, x_n, y_n)$.
- Replace y_i by $y_i^r \sim F^{(0)}_{x_i,\theta_0}$, $i = 1, \ldots, n$.

Decision-based model selection

- Write $\theta_0 = \tau^{(0)}(x_1, y_1, \ldots, x_n, y_n)$.
- Replace y_i by $y_i^r \sim F^{(0)}_{x_i,\theta_0}$, $i = 1, \ldots, n$.
- Let $\theta_r = \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r)$.

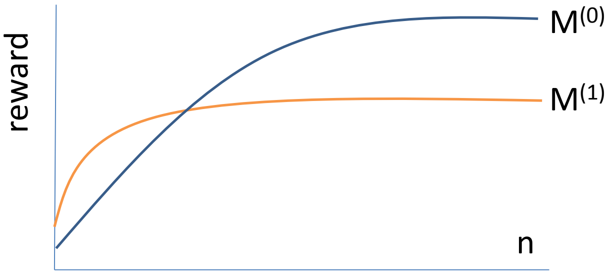
Decision-based model selection

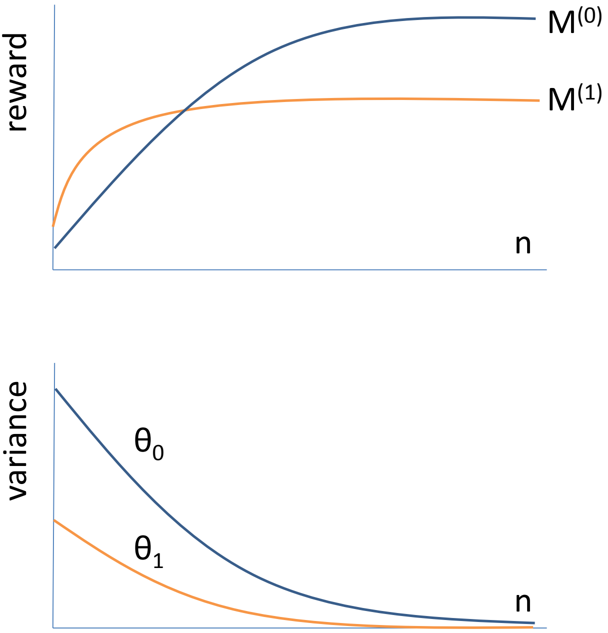
- Write $\theta_0 = \tau^{(0)}(x_1, y_1, \ldots, x_n, y_n)$.
- Replace y_i by $y_i^r \sim F^{(0)}_{x_i,\theta_0}$, $i = 1, \ldots, n$.
- Let $\theta_r = \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r)$.
- Select model

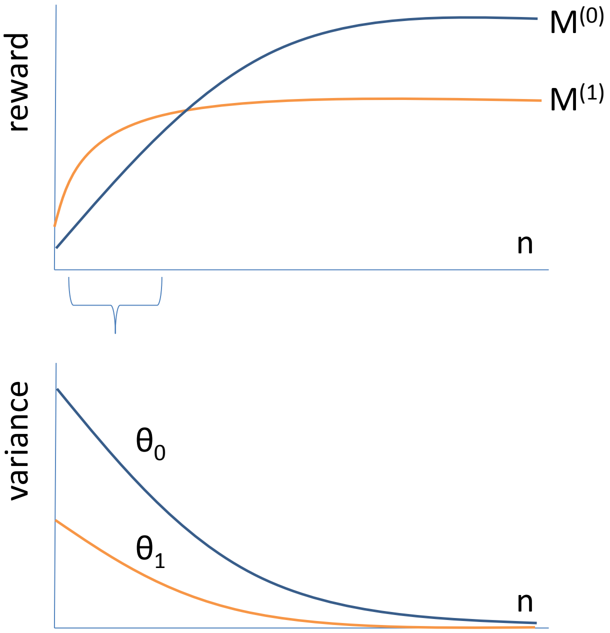
$$\underset{k \in \{0,1,\ldots,K\}}{\arg\max}\; r(x^{(k)}, \theta_r)$$

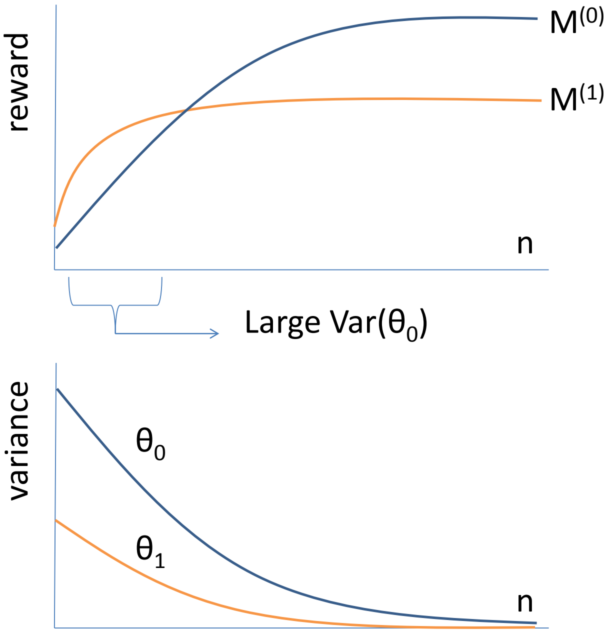
(smallest k in case of tie).

Intuition

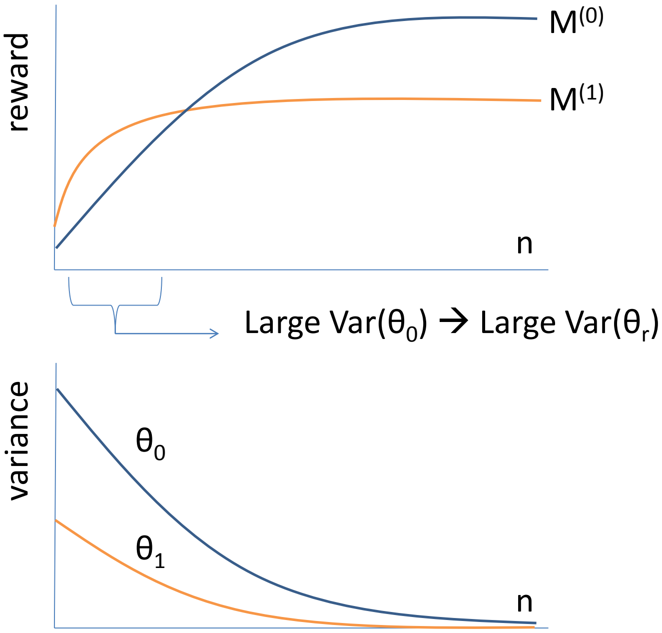

reward
n
$M^{(0)}$
$M^{(1)}$


reward
$M^{(0)}$
$M^{(1)}$
n
variance
θ_0
θ_1
n

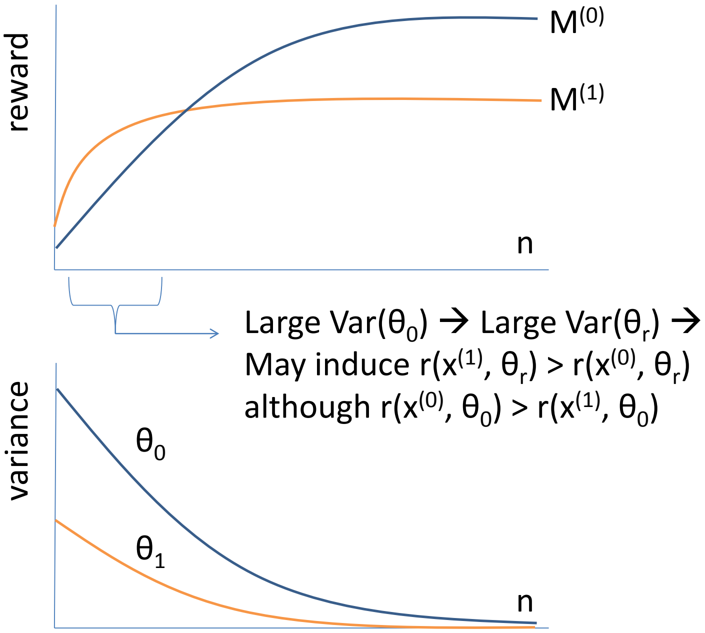

reward
M(0)
M(1)
n
variance
θ0
θ1
n

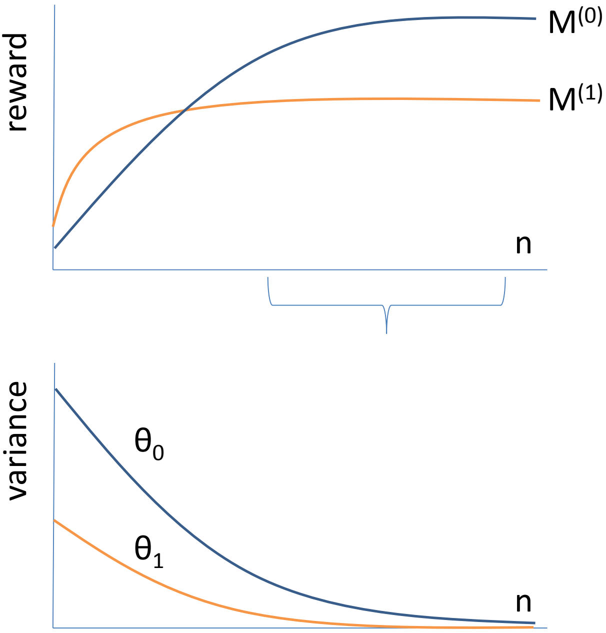


reward
M(0)
M(1)
n
Large Var(θ0)
variance
θ0
θ1
n

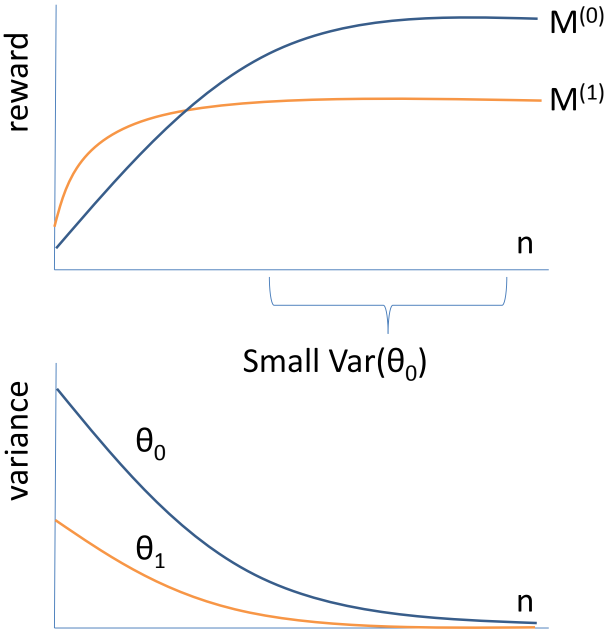


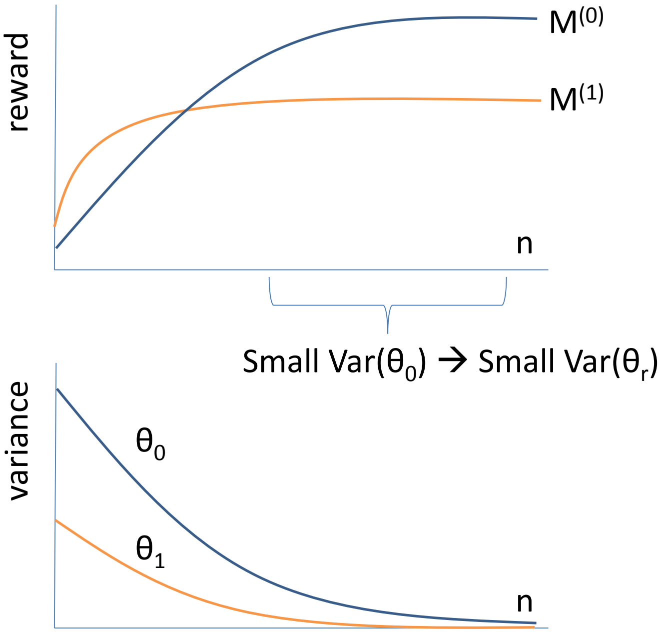
reward
M(0)
M(1)
n
Large Var(θ0) → Large Var(θr)
variance
θ0
θ1
n

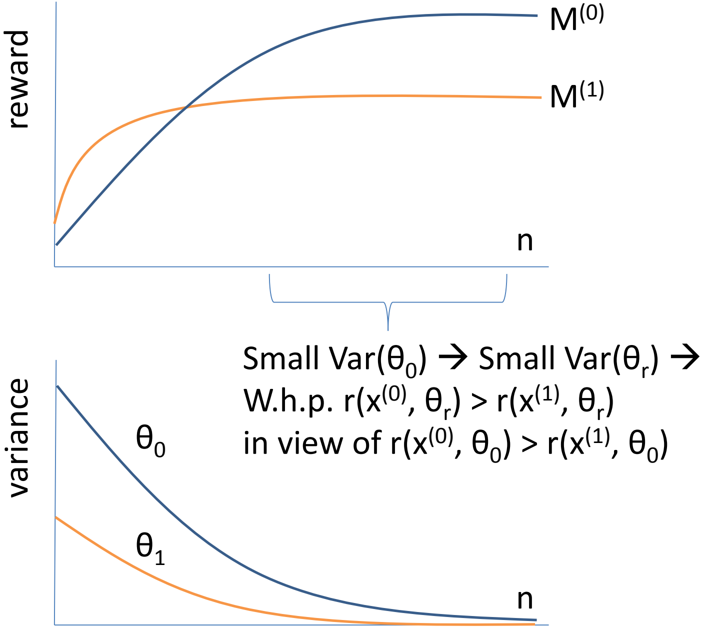

reward
M^(0)
M^(1)
n
Large Var(θ_0) → Large Var(θ_r) →
May induce r(x^(1), θ_r) > r(x^(0), θ_r)
although r(x^(0), θ_0) > r(x^(1), θ_0)
variance
θ_0
θ_1
n

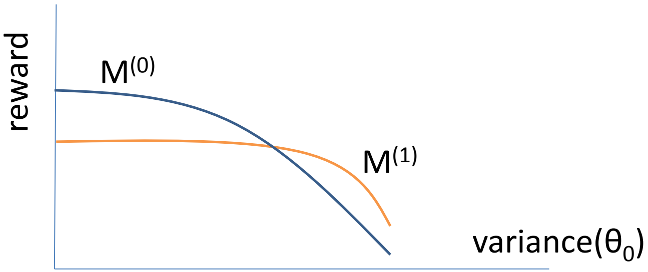


reward
M(0)
M(1)
n
variance
θ0
θ1
n

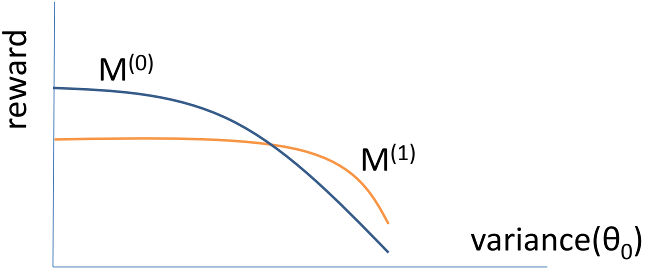

reward
M(0)
M(1)
n
Small Var(θ0)
variance
θ0
θ1
n


reward
M(0)
M(1)
n
Small Var(θ0) → Small Var(θr)
variance
θ0
θ1
n


reward
M(0)
M(1)
n
Small Var(θ0) → Small Var(θr) →
W.h.p. r(x(0), θr) > r(x(1), θr)
in view of r(x(0), θ0) > r(x(1), θ0)
variance
θ0
θ1
n



reward
M(0)
M(1)
variance(θ0)



Gain $r(x^{(1)}, \theta^*) - r(x^{(0)}, \theta^*)$ is increasing (up to a point) in $Var(\theta_0)$

Both probability of selecting $x^{(1)}$ and the resulting gain are increasing (up to a point) in $Var(\theta_0)$

Both probability of selecting $x^{(1)}$
and the resulting gain are
increasing (up to a point) in $Var(\theta_0)$

↓

Probability of selecting $x^{(1)}$
is increasing in the resulting gain

Example: Newsvendor problem

$$\min_{x \geq 0} \; h \int_0^x (x-y)\mathrm{d}\theta(y) + b \int_x^\infty (y-x)\mathrm{d}\theta(y) \quad (h, b > 0)$$

$$\min_{x \geq 0} \; h \int_0^x (x - y) \mathrm{d}\theta(y) + b \int_x^\infty (y - x) \mathrm{d}\theta(y) \quad (h, b > 0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.

$$\min_{x \geq 0} \; h \int_0^x (x-y) \mathrm{d}\theta(y) + b \int_x^\infty (y-x) \mathrm{d}\theta(y) \quad (h, b > 0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1, \ldots, y_n \sim Y$.

$$\min_{x \geq 0} h \int_0^x (x-y) \mathrm{d}\theta(y) + b \int_x^\infty (y-x) \mathrm{d}\theta(y) \quad (h, b > 0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1, \ldots, y_n \sim Y$.

Model $M^{(0)}$: nonparametric.

$$\min_{x\geq 0} h\int_0^x (x-y)\mathrm{d}\theta(y) + b\int_x^\infty (y-x)\mathrm{d}\theta(y) \quad (h,b>0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1,\dots,y_n \sim Y$.

Model $M^{(0)}$: nonparametric.
Estimate θ by empirical distribution function.

$$\min_{x \geq 0} \; h \int_0^x (x-y)\mathrm{d}\theta(y) + b \int_x^\infty (y-x)\mathrm{d}\theta(y) \quad (h, b > 0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1, \ldots, y_n \sim Y$.

Model $M^{(0)}$: nonparametric.
Estimate θ by empirical distribution function.
$\chi^{(0)}(\theta) := \inf\{z \geq 0 : \theta(z) \geq \frac{b}{b+h}\}$.

$$\min_{x \geq 0} h \int_0^x (x - y) \mathrm{d}\theta(y) + b \int_x^\infty (y - x) \mathrm{d}\theta(y) \quad (h, b > 0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1, \ldots, y_n \sim Y$.

Model $M^{(0)}$: nonparametric.
Estimate θ by empirical distribution function.
$\chi^{(0)}(\theta) := \inf\{z \geq 0 : \theta(z) \geq \frac{b}{b+h}\}$.

Model $M^{(1)}$: assume demand is exponentially distributed.

$$\min_{x\geq 0}\; h\int_0^x (x-y)\mathrm{d}\theta(y) + b\int_x^\infty (y-x)\mathrm{d}\theta(y) \quad (h,b>0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1,\dots,y_n \sim Y$.

Model $M^{(0)}$: nonparametric.
Estimate θ by empirical distribution function.
$\chi^{(0)}(\theta) := \inf\{z\geq 0 : \theta(z) \geq \frac{b}{b+h}\}$.

Model $M^{(1)}$: assume demand is exponentially distributed.
Estimate θ by $y \mapsto 1-\exp(-y/\bar{y})$.

$$\min_{x\geq 0} h\int_0^x (x-y)\mathrm{d}\theta(y) + b\int_x^\infty (y-x)\mathrm{d}\theta(y) \quad (h, b > 0)$$

Unknown cdf $\theta(\cdot)$ of the demand Y.
Data: $y_1, \ldots, y_n \sim Y$.

Model $M^{(0)}$: nonparametric.
Estimate θ by empirical distribution function.
$\chi^{(0)}(\theta) := \inf\{z \geq 0 : \theta(z) \geq \frac{b}{b+h}\}$.

Model $M^{(1)}$: assume demand is exponentially distributed.
Estimate θ by $y \mapsto 1 - \exp(-y/\bar{y})$.
$\chi^{(1)}(\bar{y}) := -\bar{y}\log(h/(b+h))$.

Experiment A: $Y \sim$ lognormal with mean m, variance v

$$m \sim U(0,5), \quad v \sim U(0,25).$$

Experiment A: $Y \sim$ lognormal with mean m, variance v
$m \sim U(0,5), \quad v \sim U(0,25).$

Experiment B: $Y \sim$ lognormal with mean m, variance m^2
$m \sim U(0,5).$

Experiment A: $Y \sim$ lognormal with mean m, variance v
$m \sim U(0, 5), \quad v \sim U(0, 25).$

Experiment B: $Y \sim$ lognormal with mean m, variance m^2
$m \sim U(0, 5).$

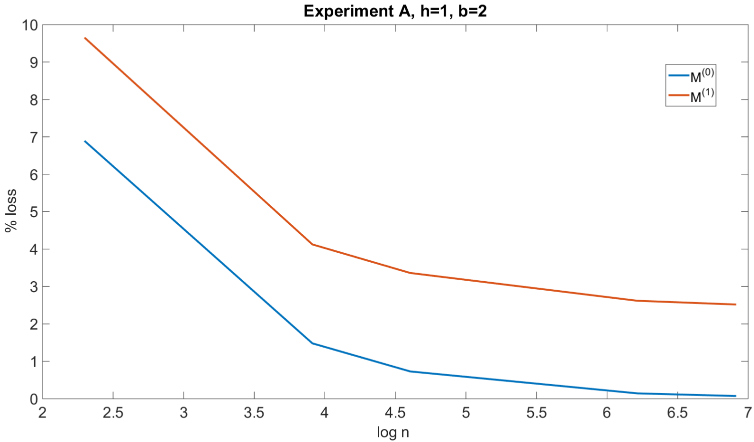
Experiment C: $Y \sim$ exponential with mean m
$m \sim U(0, 5).$

Experiment A: $Y \sim$ lognormal with mean m, variance v
$m \sim U(0, 5), \quad v \sim U(0, 25)$.

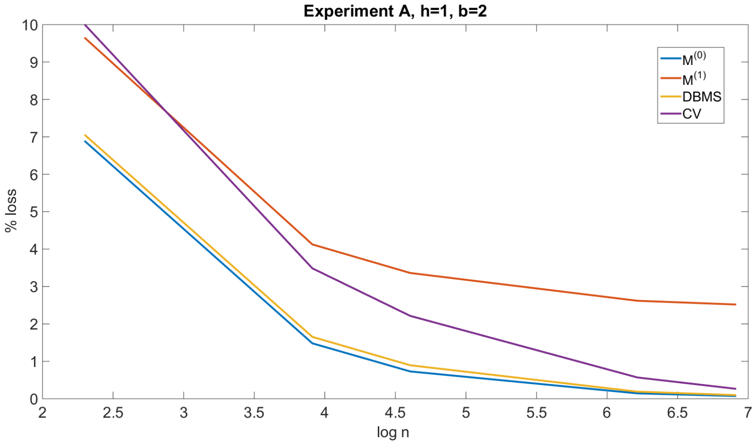
Experiment B: $Y \sim$ lognormal with mean m, variance m^2
$m \sim U(0, 5)$.

Experiment C: $Y \sim$ exponential with mean m
$m \sim U(0, 5)$.

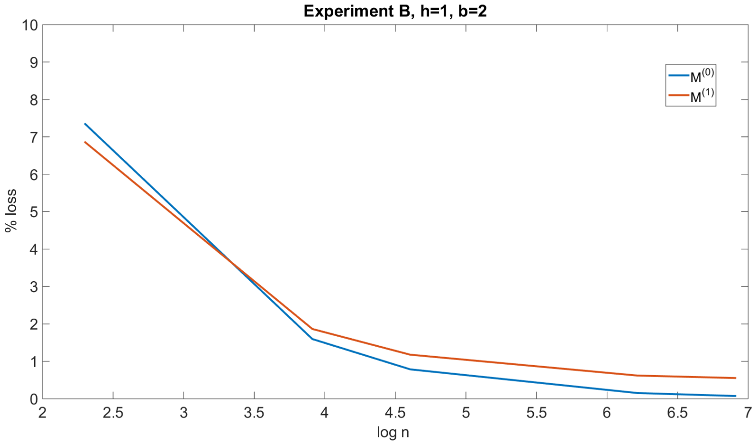
For each $n \in \{10, 50, 100, 500, 1000\}$ run 10000 simulations.
Apply DBMS, CV.



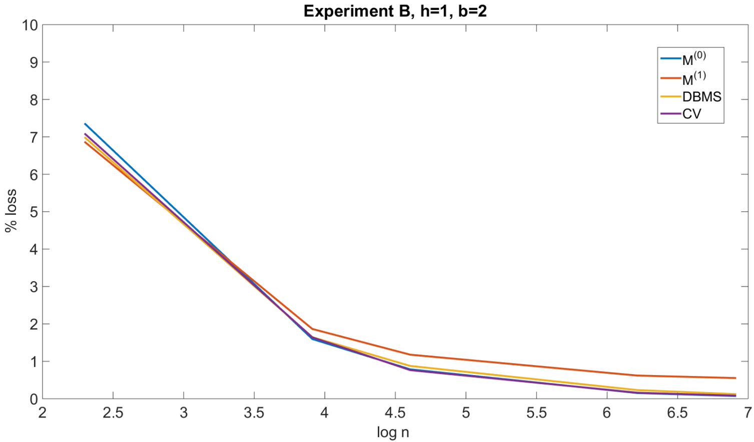
Experiment A, h=1, b=2
% loss
log n
M(0)
M(1)
0
1
2
3
4
5
6
7
8
9
10
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7



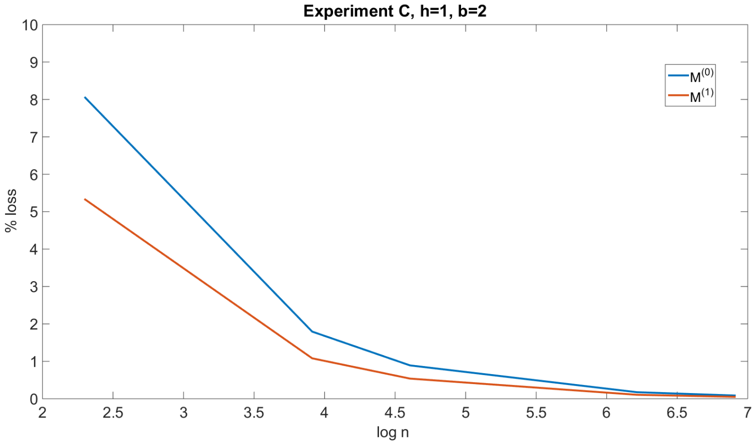
Experiment A, h=1, b=2
% loss
log n
M^(0)
M^(1)
DBMS
CV
0
1
2
3
4
5
6
7
8
9
10
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7



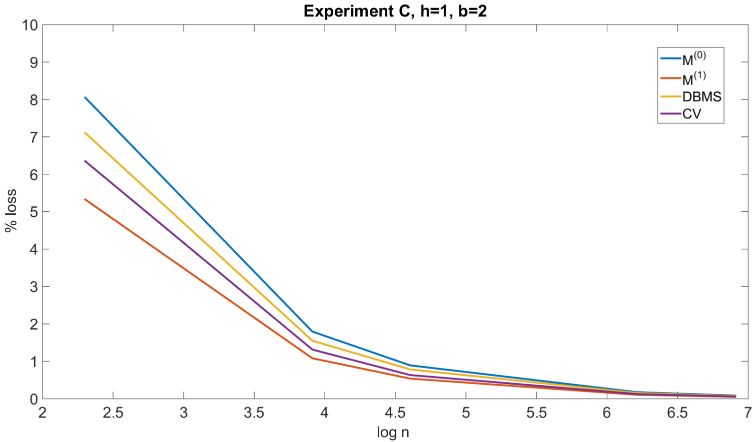
Experiment B, h=1, b=2
% loss
log n
$M^{(0)}$
$M^{(1)}$
0
1
2
3
4
5
6
7
8
9
10
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7



Experiment B, h=1, b=2
10
9
8
7
6
5
4
3
2
1
0
% loss
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7
log n
M(0)
M(1)
DBMS
CV



Experiment C, h=1, b=2
10
9
8
7
6
5
4
3
2
1
0
% loss
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7
log n
M(0)
M(1)



Experiment C, h=1, b=2
10
9
8
7
6
5
4
3
2
1
0
% loss
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7
log n
M(0)
M(1)
DBMS
CV

DBMS can significantly improve upon CV

DBMS can significantly improve upon CV

CV can improve upon DBMS

DBMS can significantly improve upon CV

CV can improve upon DBMS

Both can improve significantly upon a fixed model $M^{(k)}$

DBMS can significantly improve upon CV

CV can improve upon DBMS

Both can improve significantly upon a fixed model $M^{(k)}$

DBMS ♥ $M^{(0)}$ CV ♥ $M^{(1)}$

Summary

Summary

When are

simple & simplified models

preferable to

complex & realistic models?

Summary

When are

simple & simplified models

preferable to

complex & realistic models?

DBMS is a generic and computationally-friendly method to answer this question.

Thanks for your attention

Alternatives

DBMS: $$\underset{k \in \{0,1,\ldots,K\}}{\arg\max}\; r\big(x^{(k)}, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r)\big)$$

Alternatives

DBMS: $\underset{k\in\{0,1,\ldots,K\}}{\arg\max}\ r(x^{(k)}, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r))$

Alternative: $\underset{k\in\{0,1,\ldots,K\}}{\arg\max}\ r(x^{(k)}, \tau^{(k)}(x_1, y_1^r, \ldots, x_n, y_n^r))$

Alternatives

DBMS: $\arg\max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r))$

Alternative: $\arg\max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(k)}(x_1, y_1^r, \ldots, x_n, y_n^r))$

Alternative: $\arg\max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(i_n)}(x_1, y_1^r, \ldots, x_n, y_n^r)),\ i_n \to 0$

Alternatives (cont'd)

DBMS: $$\arg\max_{k \in \{0,1,\ldots,K\}} r\big(x^{(k)}, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r)\big)$$

Alternatives (cont'd)

DBMS: $$\arg\max_{k\in\{0,1,\ldots,K\}} r(x^{(k)}, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r))$$

Alternative: $$\arg\max_{k\in\{0,1,\ldots,K\}} \mathbb{E}[r(x^{(k)}, \tau^{(0)}(x_1, Y_1^r, \ldots, x_n, Y_n^r))]$$

$$Y_i^r \sim F^{(0)}_{x_i,\theta_0}$$

Alternatives (cont'd)

DBMS: $$\underset{k \in \{0,1,\ldots,K\}}{\arg\max}\; r(x^{(k)}, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r))$$

Alternative: $$\underset{k \in \{0,1,\ldots,K\}}{\arg\max}\; \mathbb{E}[r(x^{(k)}, \tau^{(0)}(x_1, Y_1^r, \ldots, x_n, Y_n^r))]$$

$$Y_i^r \sim F_{x_i,\theta_0}^{(0)}$$

Alternative: $$\underset{x \in \mathcal{X}}{\arg\max}\, r(x, \tau^{(0)}(x_1, y_1^r, \ldots, x_n, y_n^r)) \text{ or}$$

$$\underset{x \in \mathcal{X}}{\arg\max}\, \mathbb{E}[r(x, \tau^{(0)}(x_1, Y_1^r, \ldots, x_n, Y_n^r))]$$

Consistency

Consider a sequence of increasing data sets

$$D_n = (x_1, Y_{x_1}, \dots, x_n, Y_{x_n}), \quad n \in \mathbb{N}.$$

Consistency

Consider a sequence of increasing data sets

$$D_n = (x_1, Y_{x_1}, \ldots, x_n, Y_{x_n}), \quad n \in \mathbb{N}.$$

Let $x^{(k)}(n)$ be the decision of model $M^{(k)}$, for data set D_n.

Consistency

Consider a sequence of increasing data sets

$$D_n = (x_1, Y_{x_1}, \ldots, x_n, Y_{x_n}), \quad n \in \mathbb{N}.$$

Let $x^{(k)}(n)$ be the decision of model $M^{(k)}$, for data set D_n.
Let $\theta_r(n)$ be the resampled estimator, for data set D_n.

Consistency

Consider a sequence of increasing data sets

$$D_n = (x_1, Y_{x_1}, \ldots, x_n, Y_{x_n}), \quad n \in \mathbb{N}.$$

Let $x^{(k)}(n)$ be the decision of model $M^{(k)}$, for data set D_n.
Let $\theta_r(n)$ be the resampled estimator, for data set D_n.
Let $x^{(DBMS)}(n)$ be the decision selected by DBMS, for data set D_n.

Consistency

Consider a sequence of increasing data sets

$$D_n = (x_1, Y_{x_1}, \ldots, x_n, Y_{x_n}), \quad n \in \mathbb{N}.$$

Let $x^{(k)}(n)$ be the decision of model $M^{(k)}$, for data set D_n.
Let $\theta_r(n)$ be the resampled estimator, for data set D_n.
Let $x^{(DBMS)}(n)$ be the decision selected by DBMS, for data set D_n.

If $r(\cdot, \cdot)$ is continuous, $x^{(k)}(n) \xrightarrow{P} x^{(k)}(\infty)$ for all k, and $\theta_r(n) \xrightarrow{P} \theta^*$, then

$$|r(x^{(DBMS)}(n), \theta^*) - \max_{k \in \{0,1,\ldots,K\}} r(x^{(k)}(n), \theta^*)| \xrightarrow{P} 0 \text{ as } n \to \infty.$$

Example: Assortment optimization

$$\max_{x \subset \{1,\dots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \dots, m$.

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.
Estimate each $\theta_{i,x}$ by (adjusted) relative frequency.

$$\max_{x\subset\{1,\ldots,m\}} \sum_{i\in x} r_i\theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.
Estimate each $\theta_{i,x}$ by (adjusted) relative frequency.
Optimization: brute force (small m).

$$\max_{x \subset \{1,\dots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \dots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \dots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.
Estimate each $\theta_{i,x}$ by (adjusted) relative frequency.
Optimization: brute force (small m).

Model $M^{(1)}$: multinomial logit model.

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.
Estimate each $\theta_{i,x}$ by (adjusted) relative frequency.
Optimization: brute force (small m).

Model $M^{(1)}$: multinomial logit model.
Assume $\theta_{i,x} = \exp(v_i)/(1 + \sum_{j \in x} \exp(v_j))$
for some $v_1, \ldots, v_m \in \mathbb{R}$ and all i, x.

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.
Estimate each $\theta_{i,x}$ by (adjusted) relative frequency.
Optimization: brute force (small m).

Model $M^{(1)}$: multinomial logit model.
Assume $\theta_{i,x} = \exp(v_i)/(1 + \sum_{j \in x} \exp(v_j))$
for some $v_1, \ldots, v_m \in \mathbb{R}$ and all i, x.
Estimate $v_1, \ldots, v_m$ by maximum likelihood estimation.

$$\max_{x \subset \{1,\ldots,m\}} \sum_{i \in x} r_i \theta_{i,x}$$

r_i revenue of product $i = 1, \ldots, m$.
$\theta_{i,x}$ probability that consumer selects i from assortment x.

Data $x_1, y_1, \ldots, x_n, y_n$;

Model $M^{(0)}$: nonparametric.
Estimate each $\theta_{i,x}$ by (adjusted) relative frequency.
Optimization: brute force (small m).

Model $M^{(1)}$: multinomial logit model.
Assume $\theta_{i,x} = \exp(v_i)/(1 + \sum_{j \in x} \exp(v_j))$
for some $v_1, \ldots, v_m \in \mathbb{R}$ and all i, x.
Estimate $v_1, \ldots, v_m$ by maximum likelihood estimation.
Optimization: brute force (efficient algorithm).

Experiment A: Draw each $(\theta_{i,x} : i \in x \cup \{0\})$ from the simplex of appropriate dimension.

Experiment A: Draw each $(\theta_{i,x} : i \in x \cup \{0\})$ from the simplex of appropriate dimension.

Experiment B: Parsimonious Generalized Attraction Model

$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j) + \eta \sum_{j \notin x} \exp(v_j)}, \quad \eta, v_1, \ldots, v_m \sim U(0,1).$$

Experiment A: Draw each $(\theta_{i,x} : i \in x \cup \{0\})$ from the simplex of appropriate dimension.

Experiment B: Parsimonious Generalized Attraction Model

$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j) + \eta \sum_{j \notin x} \exp(v_j)}, \quad \eta, v_1, \ldots, v_m \sim U(0,1).$$

Experiment C: Multinomial Logit Model

$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j)}, \quad v_1, \ldots, v_m \sim U(0,1).$$

Experiment A: Draw each $(\theta_{i,x} : i \in x \cup \{0\})$ from the simplex of appropriate dimension.

Experiment B: Parsimonious Generalized Attraction Model

$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j) + \eta \sum_{j \notin x} \exp(v_j)}, \quad \eta, v_1, \ldots, v_m \sim U(0,1).$$

Experiment C: Multinomial Logit Model

$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j)}, \quad v_1, \ldots, v_m \sim U(0,1).$$

Set $r_i = 100i/m$, $i = 1, \ldots, m$.

Experiment A: Draw each $(\theta_{i,x} : i \in x \cup \{0\})$ from the simplex of appropriate dimension.

Experiment B: Parsimonious Generalized Attraction Model

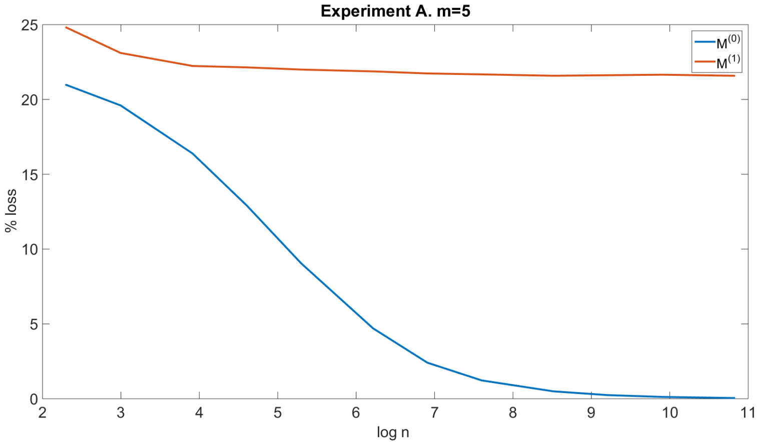
$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j) + \eta \sum_{j \notin x} \exp(v_j)}, \quad \eta, v_1, \ldots, v_m \sim U(0,1).$$

Experiment C: Multinomial Logit Model

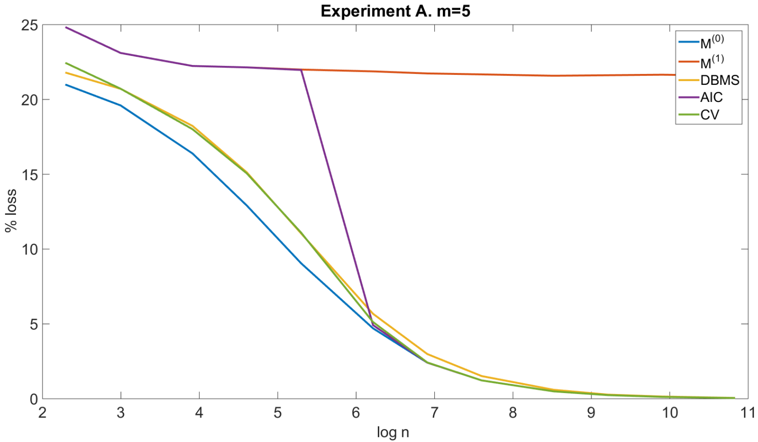
$$\theta_{i,x} = \frac{\exp(v_i)}{1 + \sum_{j \in x} \exp(v_j)}, \quad v_1, \ldots, v_m \sim U(0,1).$$

Set $r_i = 100i/m$, $i = 1, \ldots, m$.

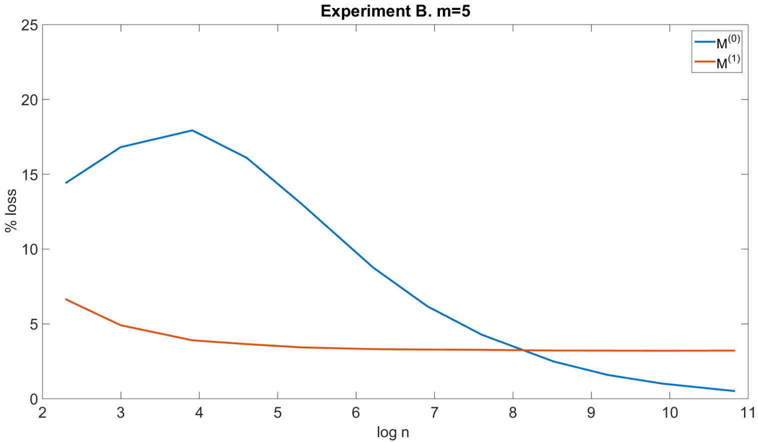
For each $n \in \{10, 20, 50, 100, 200, 500, ..., 50000\}$ run 10000 simulations. Apply DBMS, AIC, CV.



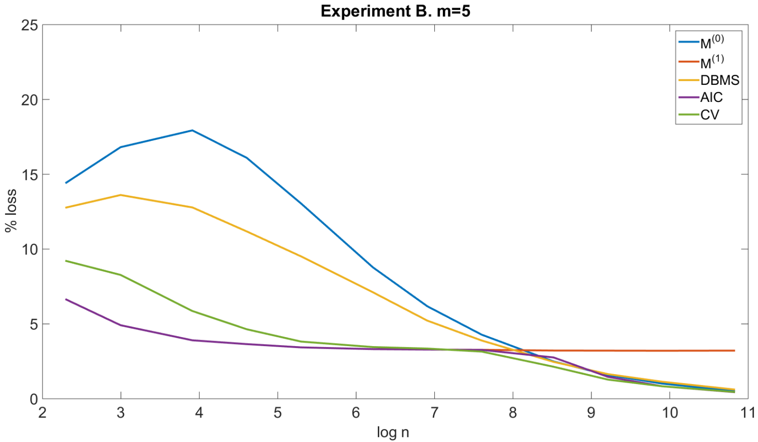
Experiment A. m=5
25
20
15
10
5
0
% loss
2
3
4
5
6
7
8
9
10
11
log n
M(0)
M(1)

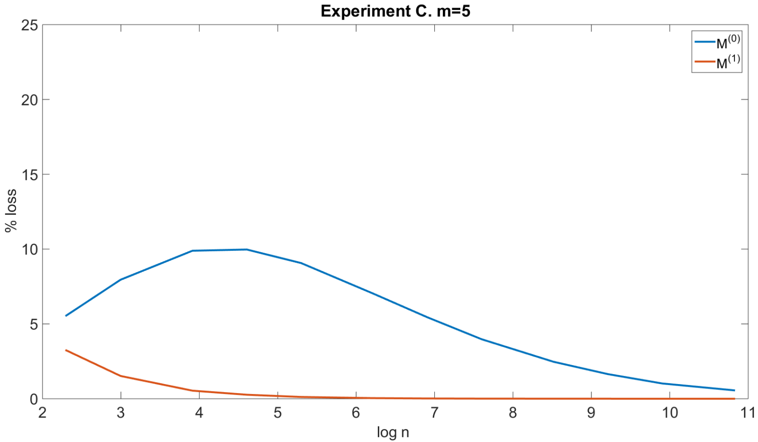


Experiment A. m=5
% loss
log n
0
5
10
15
20
25
2
3
4
5
6
7
8
9
10
11
M(0)
M(1)
DBMS
AIC
CV

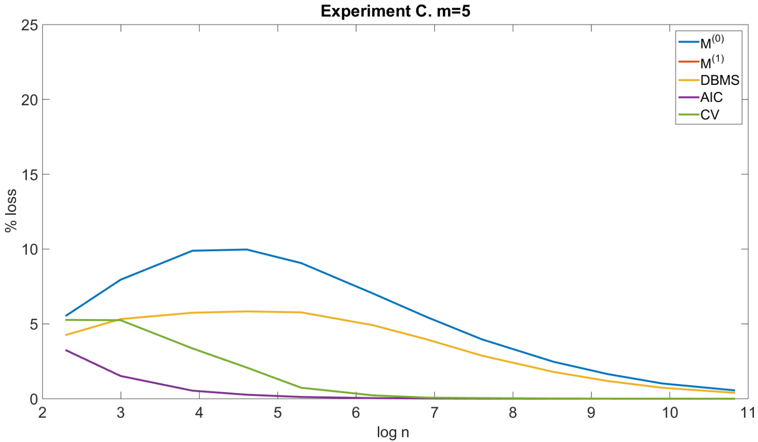


Experiment B. m=5
% loss
log n
M(0)
M(1)
0
5
10
15
20
25
2
3
4
5
6
7
8
9
10
11


Experiment B. m=5
25
20
15
10
5
0
% loss
2
3
4
5
6
7
8
9
10
11
log n
M(0)
M(1)
DBMS
AIC
CV



Experiment C. m=5
% loss
log n
M^(0)
M^(1)
25
20
15
10
5
0
2
3
4
5
6
7
8
9
10
11


Experiment C. m=5
25
20
15
10
5
0
% loss
2
3
4
5
6
7
8
9
10
11
log n
M(0)
M(1)
DBMS
AIC
CV

DBMS and CV can significantly outperform AIC

DBMS and CV can significantly outperform AIC

AIC can outperform DBMS and CV

DBMS and CV can significantly outperform AIC

AIC can outperform DBMS and CV

All three can improve significantly upon a fixed model $M^{(k)}$

DBMS and CV can significantly outperform AIC

AIC can outperform DBMS and CV

All three can improve significantly upon a fixed model $M^{(k)}$

DBMS ♥ $M^{(0)}$ CV ♥ $M^{(1)}$ AIC ♥♥ $M^{(1)}$