Robust Multiclass Queuing Theory for Wait Time Estimation in Resource Allocation Systems

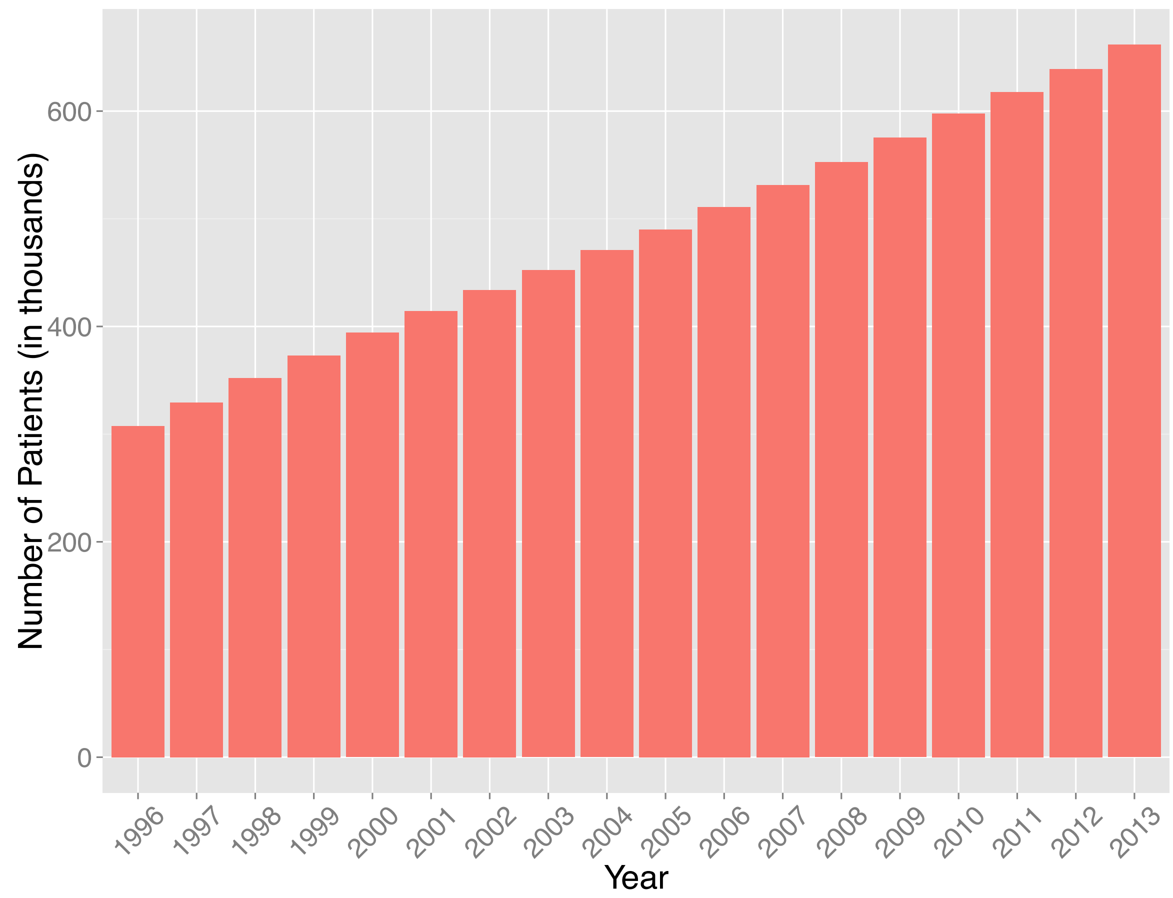
Banff Workshop on Distributionally Robust Optimization

Phebe Vayanos
Assistant Professor of Industrial & Systems Engineering and Computer Science
Viterbi School of Engineering, USC

(joint with Bandi and Trichakis, forthcoming in *Management Science*)

End-Stage Renal Disease

source: https://www.usrds.org



- terminal disease affecting >600,000 patients in U.S.
- dialysis vs. kidney transplant (preferred)
- living donors vs. deceased donors

Organ Shortage

- 100k patients waiting
- 36k additions per year
- 19k transplants/year
  - 13.4k (70%) from deceased donors
  - 5.6k (30%) from living donors

Organ Shortage

3-yr trend

- 100k patients waiting +20%
- 36k additions per year
- 19k transplants/year
  - 13.4k (70%) from deceased donors +20%
  - 5.6k (30%) from living donors -2%

U.S. Kidney Allocation System

patient information

UNOS DONATE LIFE

U.S. Kidney Allocation System

patient information

deceased donor information

UNOS DONATE LIFE

1	John
2	Maria
3	Alexander
4	Lea
...	...

U.S. Kidney Allocation System

patient information

deceased donor information

UNOS DONATE LIFE

1	John
2	Maria
3	Alexander
4	Lea
...	...

- medical compatibility: blood group, weight, etc.
- geographic proximity (24-36 hours to transplant)
- *point based:* wait time, blood antigens: **~FCFS**

Wait Time Estimation

Patient X of blood type O is listed in a given geographic region. He is currently ranked 50th. How long until he receives an offer of a particular quality?

Wait Time Estimation

Patient X of blood type O is listed in a given geographic region. He is currently ranked 50th. How long until he receives an offer of a particular quality?

- important for:
  - dialysis management
  - planning of daily life activities
  - accept/reject decisions

Challenges

Challenges

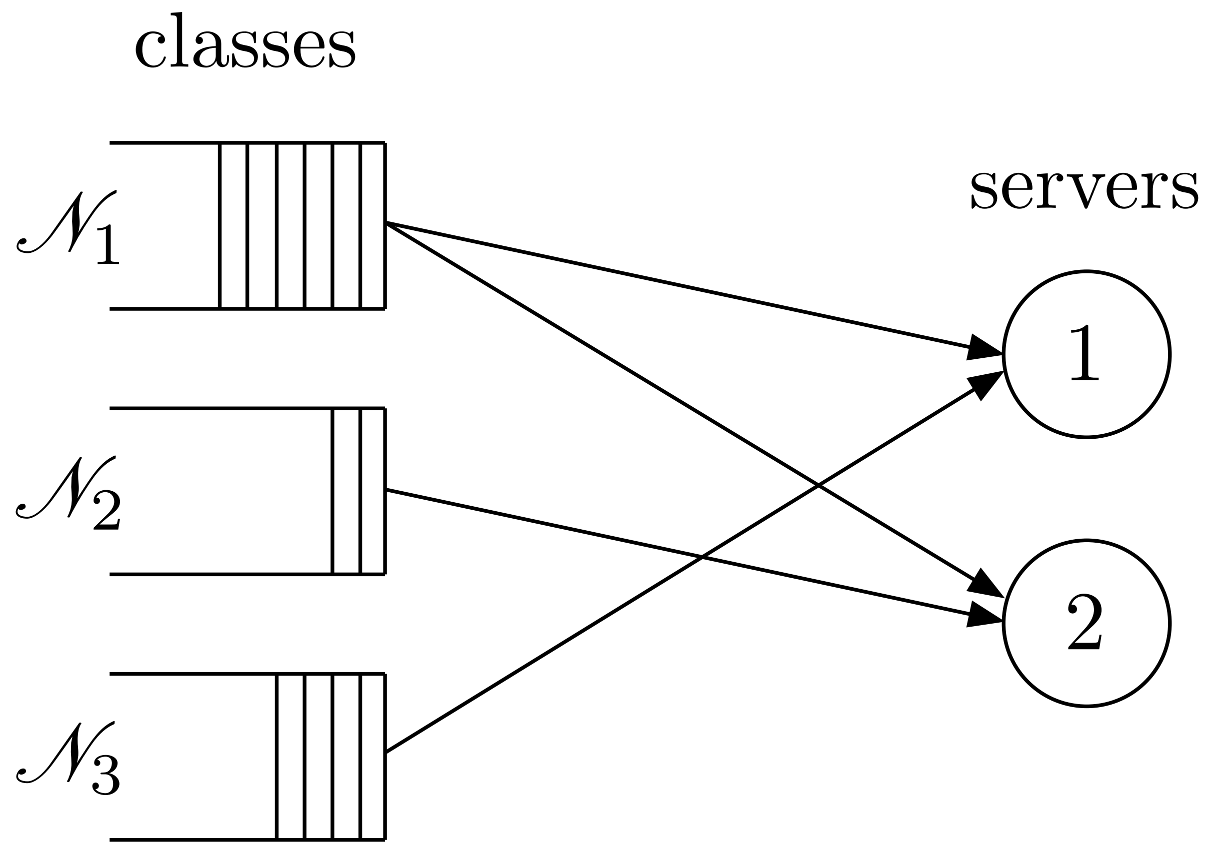
- predicting accept/decline decisions is already hard
  - Kim *et al* 15: use all available historical data, build series of prediction models (log. reg., SVM, CART, RF); error rates vary 22-47%

Challenges

- predicting accept/decline decisions is already hard
  - Kim *et al* 15: use all available historical data, build series of prediction models (log. reg., SVM, CART, RF); error rates vary 22-47%

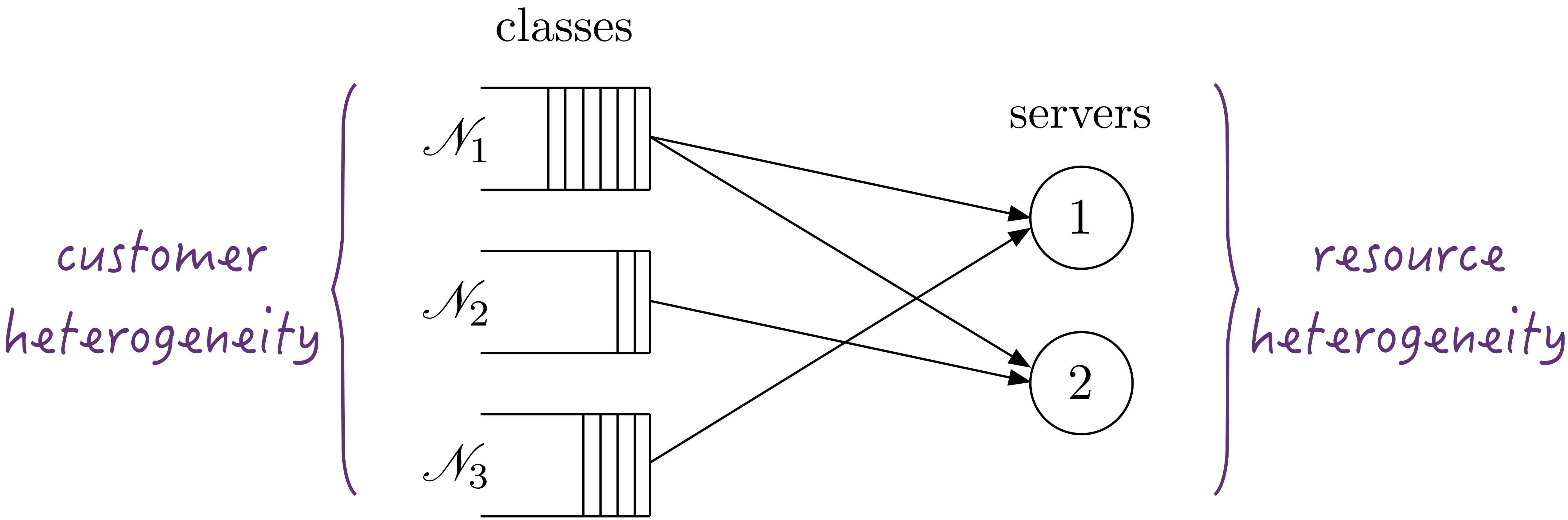
- in practice:
  - incomplete information: other patients' preferences
  - unstable/ non-stationary system

Resource Allocation System



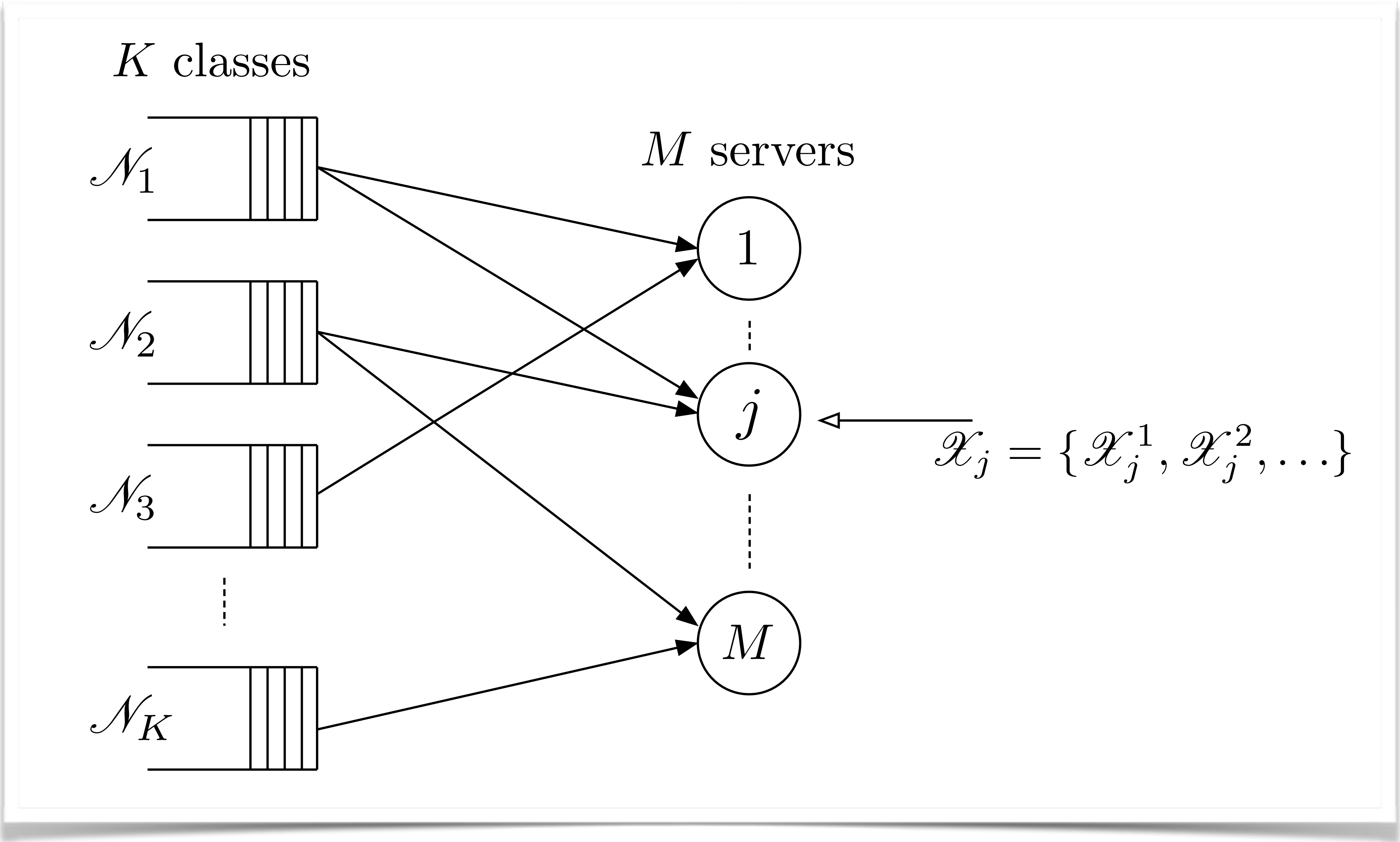
- multiclass, multiserver (MCMS) queuing system
    - servers: resource types
    - customer classes/queues: preferences

Resource Allocation System

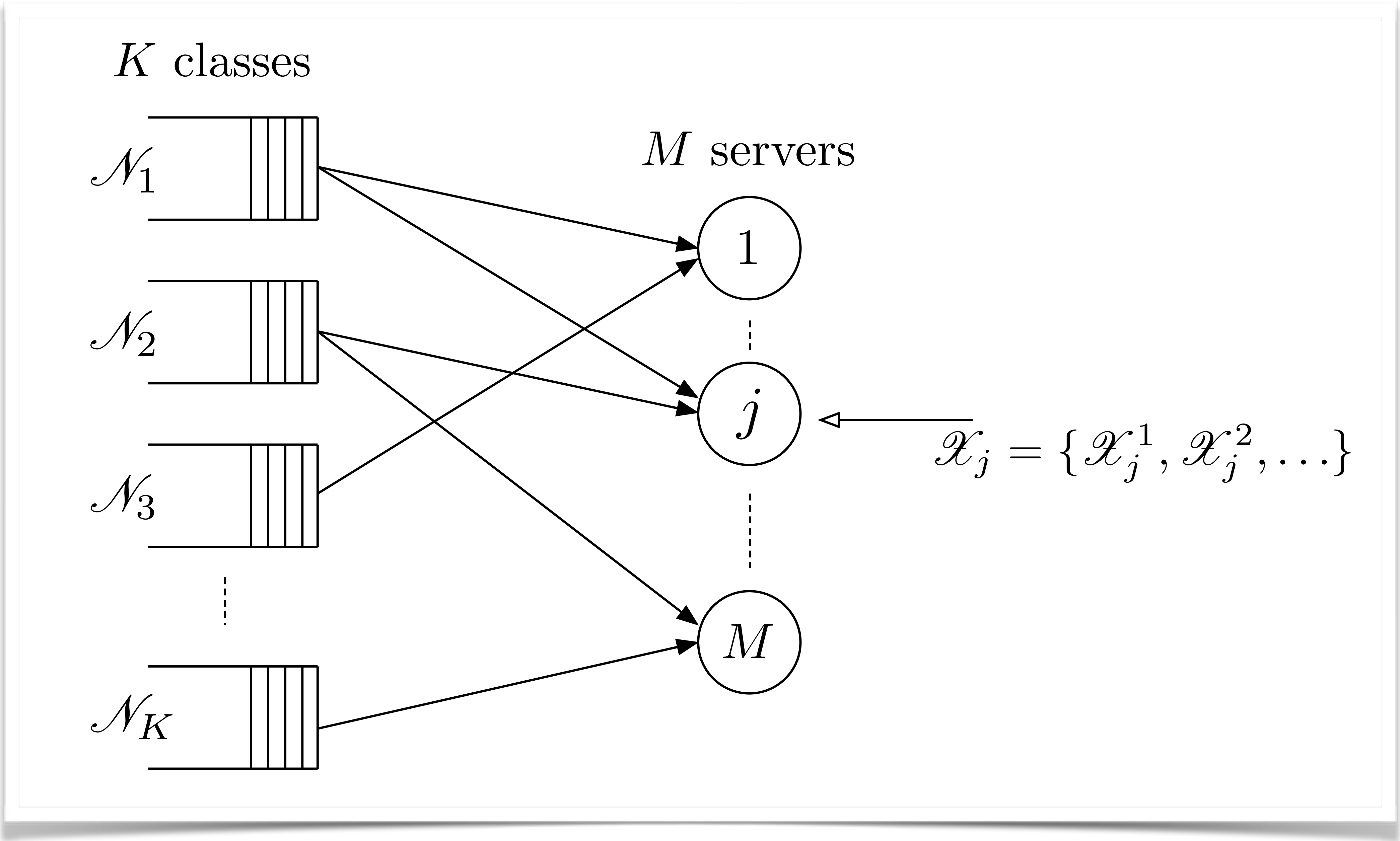


- multiclass, multiserver (MCMS) queuing system
  - servers: resource types
  - customer classes/queues: preferences

MCMS under FCFS

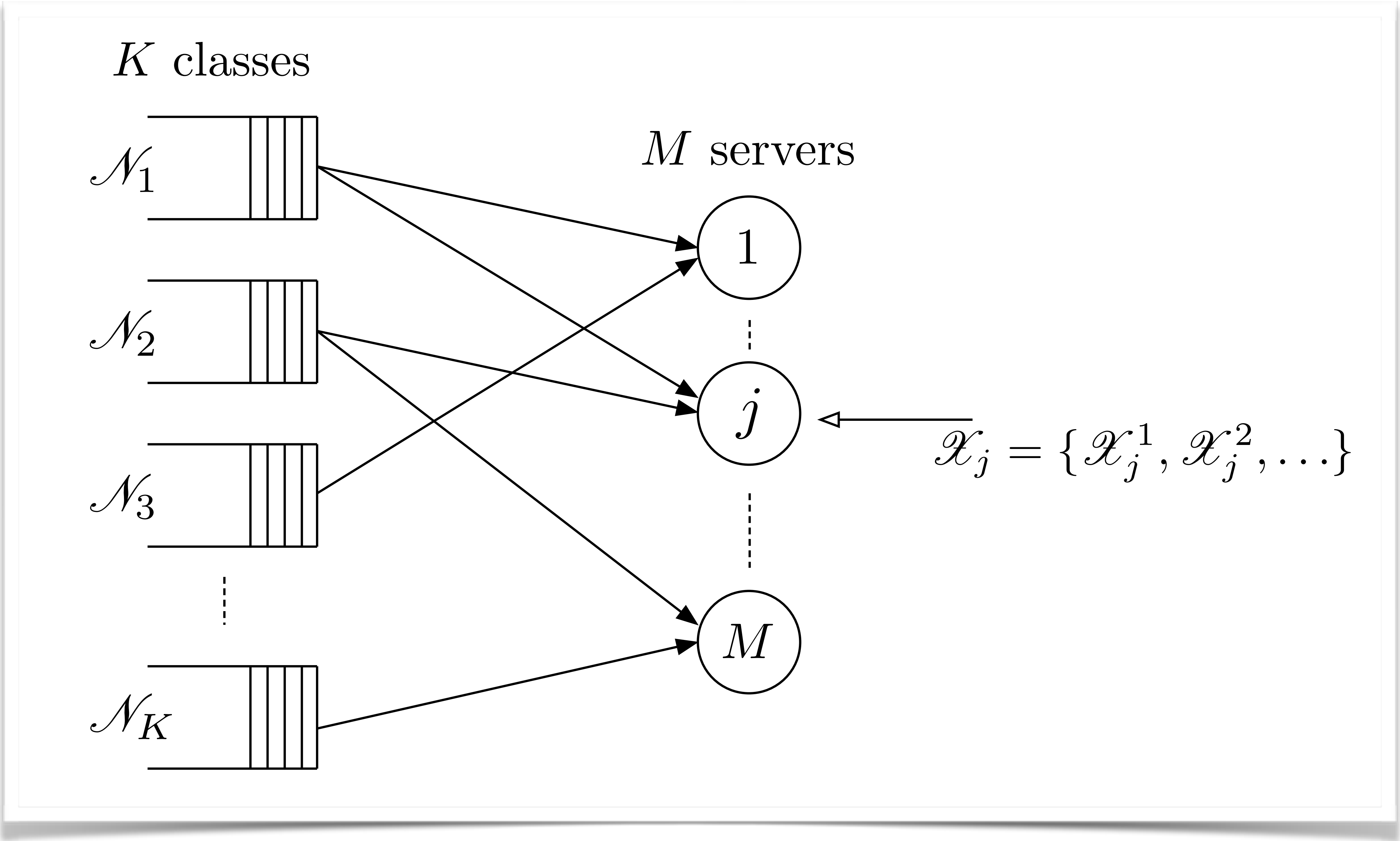


MCMS under FCFS



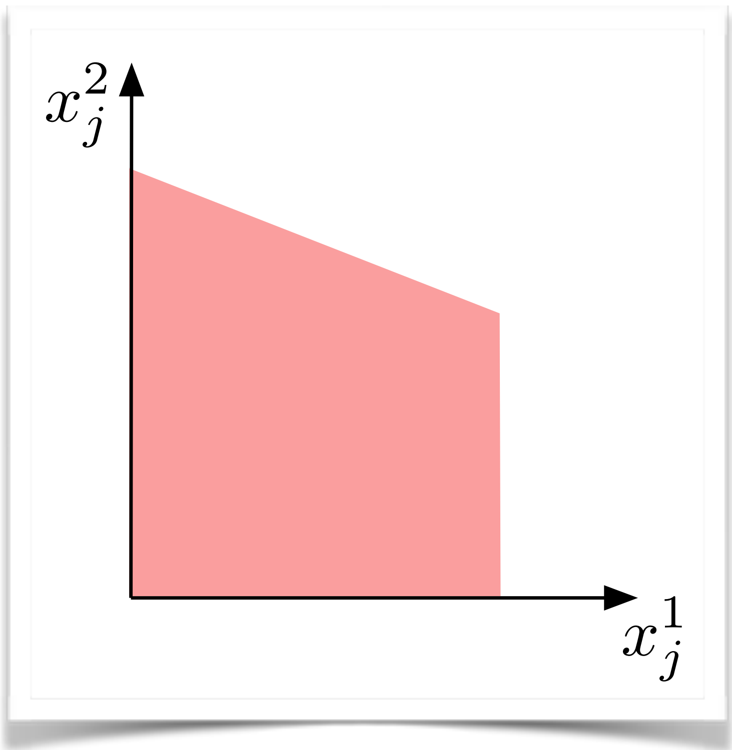
- $\sigma(\nu)$ arrival order of customer $\nu \in \{1, \ldots, \sum_i \mathcal{N}_i\}$

MCMS under FCFS



- $\sigma(\nu)$ arrival order of customer $\nu \in \{1, \ldots, \sum_i \mathscr{N}_i\}$
- $\mathscr{W}_i(\mathscr{N}_1, \ldots, \mathscr{N}_K, \sigma, \mathscr{X}_1, \ldots, \mathscr{X}_M)$ clearing time of queue i

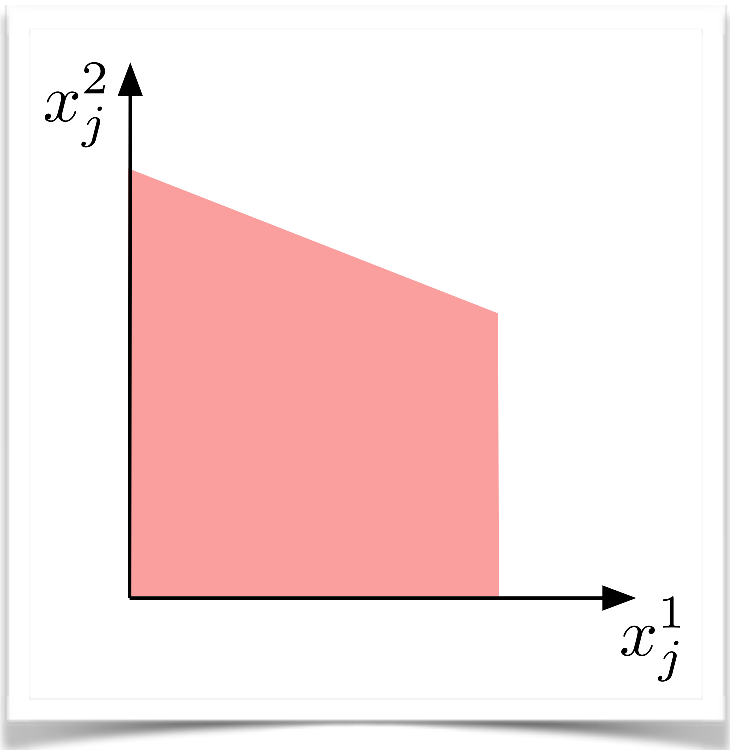
Model of Uncertainty

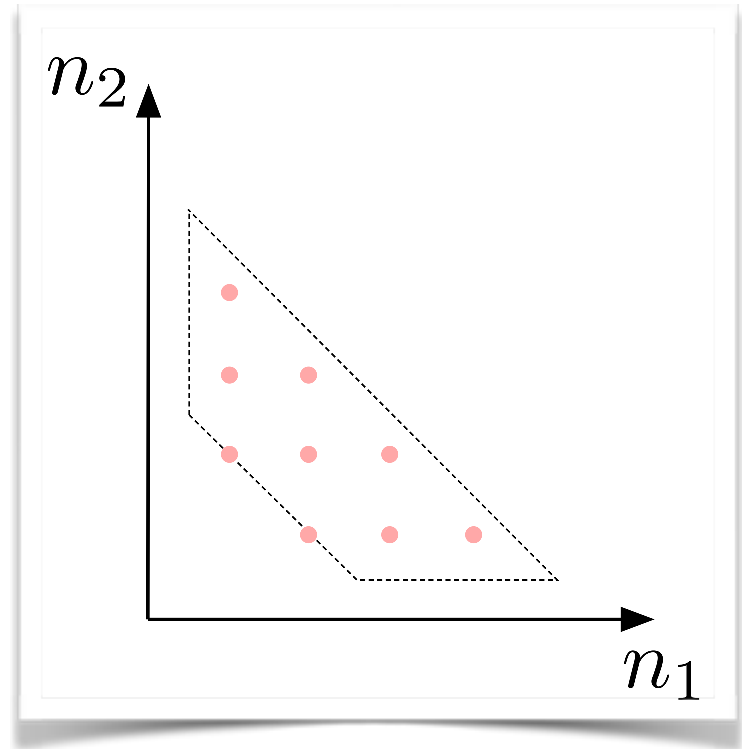


- service times:

$$\mathbb{X}_j = \left\{ x_j \in \mathbb{R}^{\bar{\ell}_j} \; : \; \sum_{k=1}^{\ell} x_j^k \leq \frac{\ell}{\mu_j} + \Gamma_j^{\mathbb{X}} (\ell)^{1/\alpha_j}, \, \ell = 1, \ldots, \bar{\ell}_j \right\}$$

Model of Uncertainty



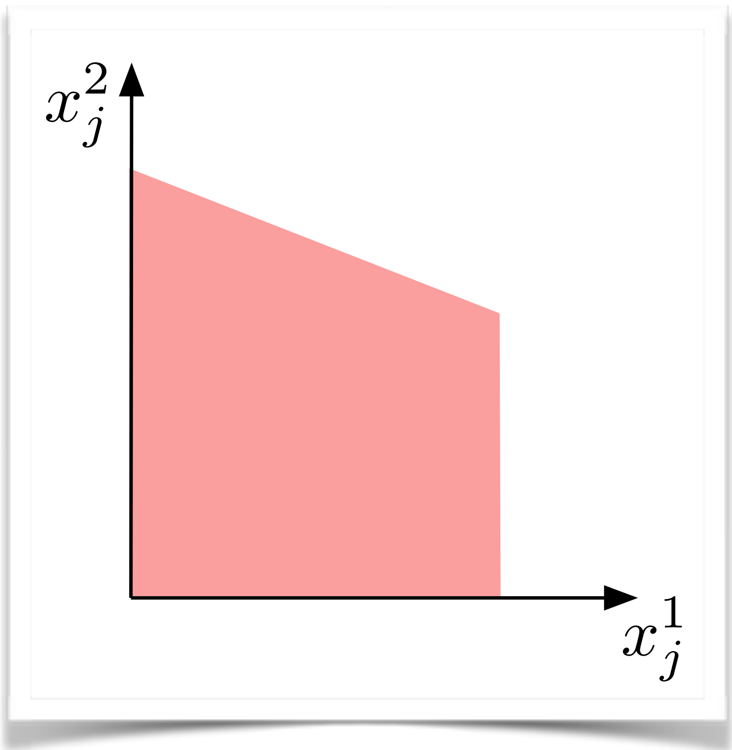


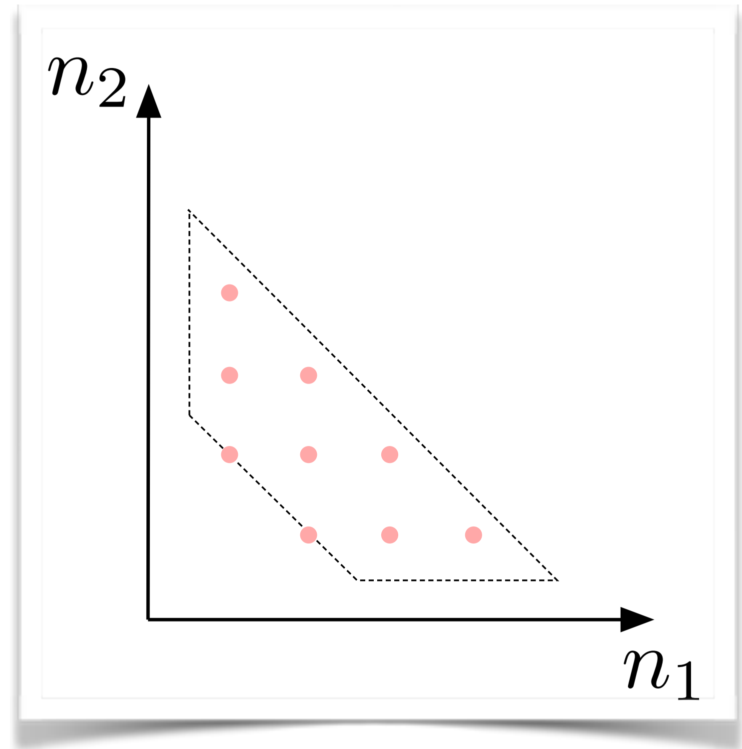
- service times:

$$\mathbb{X}_j = \left\{ x_j \in \mathbb{R}^{\bar{\ell}_j} \; : \; \sum_{k=1}^{\ell} x_j^k \leq \frac{\ell}{\mu_j} + \Gamma_j^{\mathbb{X}} (\ell)^{1/\alpha_j}, \, \ell = 1, \ldots, \bar{\ell}_j \right\}$$

- population vector $n \in \mathbb{P} \cap \mathbb{N}^K$

Model of Uncertainty





- service times:

$$\mathbb{X}_j = \left\{ x_j \in \mathbb{R}^{\bar{\ell}_j} \ : \ \sum_{k=1}^{\ell} x_j^k \leq \frac{\ell}{\mu_j} + \Gamma_j^{\mathbb{X}} (\ell)^{1/\alpha_j}, \ \ell = 1, \ldots, \bar{\ell}_j \right\}$$

- population vector $n \in \mathbb{P} \cap \mathbb{N}^K$
- arrival order $\sigma \in \Sigma(n)$

Robust Wait Times

$$
\begin{array}{lll}
W_i: & \text{maximize} & \mathscr{W}_i(n_1, \ldots, n_K, \sigma, x_1, \ldots, x_M) \\
& \text{subject to} & n \in \mathbb{P} \cap \mathbb{N}^K \\
& & \sigma \in \Sigma(n) \\
& & x_j \in \mathbb{X}_j, \quad j = 1, \ldots, M
\end{array}
$$

Robust Wait Times

$$
\begin{aligned}
W_i: \quad & \text{maximize} \quad && \mathscr{W}_i(n_1, \ldots, n_K, \sigma, x_1, \ldots, x_M) \\
& \text{subject to} \quad && n \in \mathbb{P} \cap \mathbb{N}^K \\
& && \sigma \in \Sigma(n) \\
& && x_j \in \mathbb{X}_j, \quad j = 1, \ldots, M
\end{aligned}
$$

- robust wait time estimation problem is NP-hard

Robust Wait Times

$$
\begin{array}{lrl}
W_i : & \text{maximize} & \mathscr{W}_i(n_1, \ldots, n_K, \sigma, x_1, \ldots, x_M) \\
& \text{subject to} & n \in \mathbb{P} \cap \mathbb{N}^K \\
& & \sigma \in \Sigma(n) \\
& & x_j \in \mathbb{X}_j, \quad j = 1, \ldots, M
\end{array}
$$

- robust wait time estimation problem is NP-hard
- no tractable expression for $\mathscr{W}_i$
  - Lindley equations break down

Robust Wait Times

$$
\begin{array}{lll}
W_i: & \text{maximize} & \mathscr{W}_i(n_1, \ldots, n_K, \sigma, x_1, \ldots, x_M) \\
& \text{subject to} & n \in \mathbb{P} \cap \mathbb{N}^K \\
& & \sigma \in \Sigma(n) \\
& & x_j \in \mathbb{X}_j, \quad j = 1, \ldots, M
\end{array}
$$

- robust wait time estimation problem is NP-hard
- no tractable expression for $\mathscr{W}_i$
  - Lindley equations break down
- key idea: model assignment of servers to customers
  - y_{kj}^{ℓ}: ℓth service from server j assigned to class k

Robust Wait Times

assignment-style formulation

$$
\begin{array}{ll}
\text{maximize} & w_i \\
\text{subject to} & \displaystyle\sum_k y_{kj}^\ell \le 1, \quad \sum_{\ell,j} y_{kj}^\ell \le n_k \\
& \displaystyle\sum_{k'} y_{k'j}^\ell \ge f_{kj}^\ell \\
& w_k \le c_j^\ell + \bar{\zeta} f_{kj}^\ell \\
& w_k \ge c_j^\ell - \bar{\zeta}\left(1 - y_{kj}^\ell\right) \\
& (c, n) \in \text{uncertainty sets}, \ (y, f) \text{ binary}
\end{array}
$$

Performance: Accuracy

- estimation error vs simulation

statistics	avg.	95-%ile	97-%ile	99-%ile
avg. abs. rel. error	6.52%	2.64%	2.55%	3.41%

Performance: Accuracy

- estimation error vs simulation

statistics	avg.	95-%ile	97-%ile	99-%ile
avg. abs. rel. error	6.52%	2.64%	2.55%	3.41%

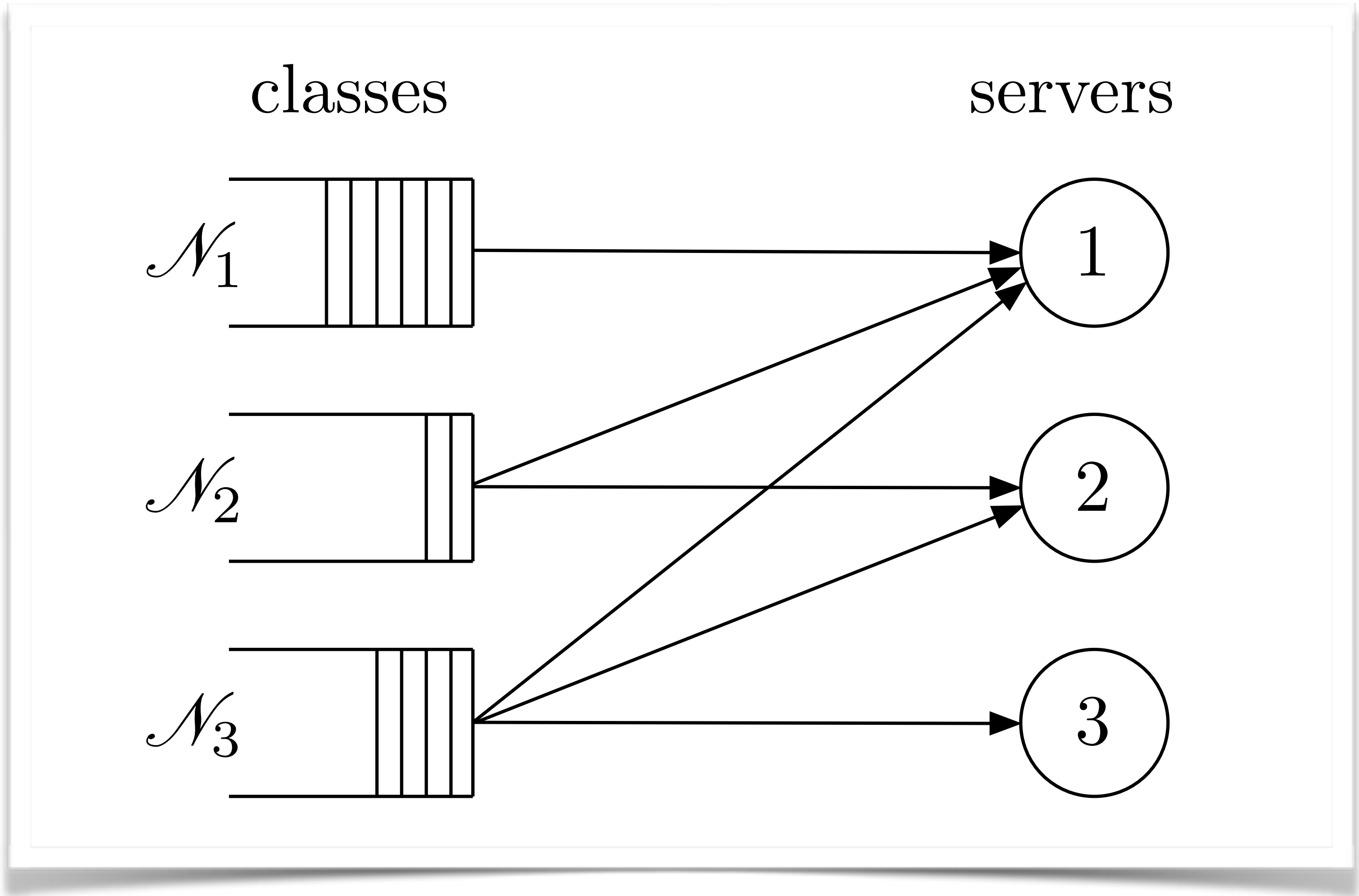
- estimation error when true distribution $\neq$ assumed

avg. queue population	5	100	500
simulation avg. abs. rel. error	21%	15%	12%
our avg. abs. rel. error	13%	9%	7.5%

Hierarchical MCMS

- hierarchy across resource types
  - e.g., different quality service level
    - radiation therapy
    - organ quality
- server j provides jth ranked service
- induces "threshold-type" customer preferences

Hierarchical MCMS



- nested structure enables to strengthen formulations
- robust wait time for service of any rank W_K
- problem remains NP-hard

Wait Time for Service in HMCMS

Lemma. For HMCMS systems:

- $\mathscr{W}_K$ increasing in completion times
- completion times can be fixed to their worst-case values:

$$c_j^\ell = x_j^1 + \ldots + x_j^\ell = \frac{\ell}{\mu_j} + \Gamma_j^{\mathbb{X}}(\ell)^{1/\alpha_j}$$

Wait Time for Service in HMCMS

Lemma. For HMCMS systems:

- $\mathscr{W}_K$ increasing in completion times
- completion times can be fixed to their worst-case values:

$$c_j^\ell = x_j^1 + \ldots + x_j^\ell = \frac{\ell}{\mu_j} + \Gamma_j^{\mathbb{X}}(\ell)^{1/\alpha_j}$$

- assignment of servers to customers
  - y_{kj}^ℓ : ℓth service from server j assigned to class k
  - c_j^ℓ : time ℓth service from server j starts

Wait Time for Service in HMCMS

Lemma. For HMCMS systems:

- $\mathscr{W}_K$ increasing in completion times
- completion times can be fixed to their worst-case values:

$$c_j^\ell = x_j^1 + \ldots + x_j^\ell = \frac{\ell}{\mu_j} + \Gamma_j^{\mathbb{X}}(\ell)^{1/\alpha_j}$$

- assignment of servers to customers
  - y_{kj}^ℓ : ℓth service from server j assigned to class k
  - ~~c_j^ℓ : time ℓth service from server j starts~~

$\Rightarrow$ *drastic variable reduction*

Scalable Heuristic

Scalable Heuristic

- view so far: individual assignments y_{kj}^{ℓ}
  - scales with n

Scalable Heuristic

- view so far: individual assignments y_{kj}^{ℓ}
  - scales with n
- alternative view:
  - aggregate assignments m_j
  - independent of n

Scalable Heuristic

- view so far: individual assignments y_{kj}^{ℓ}
  - scales with n
- alternative view:
  - aggregate assignments m_j
  - independent of n

$$
\widehat{W}_K \qquad
\begin{array}{ll}
\text{maximize} & w \\
\text{subject to} & w \leq \dfrac{m_j}{\mu_j} + \Gamma_j^{\mathbb{X}} (m_j)^{1/\alpha_j} \\
& \displaystyle\sum_{k=j}^{K} m_k \leq \sum_{k=j}^{K} n_k + K - j \\
& n \in \mathbb{P}
\end{array}
$$

Scalable Heuristic

- view so far: individual assignments y_{kj}^{ℓ}
  - scales with n
- alternative view:
  - aggregate assignments m_j
  - independent of n

$$\widehat{W}_K \quad \begin{array}{ll} \text{maximize} & w \\ \text{subject to} & w \leq \dfrac{m_j}{\mu_j} + \Gamma_j^{\mathbb{X}} (m_j)^{1/\alpha_j} \\ & \displaystyle\sum_{k=j}^{K} m_k \leq \sum_{k=j}^{K} n_k + K - j \\ & n \in \mathbb{P} \end{array} \Bigg\} \text{ SOCP!}$$

Approximation Guarantee

- W_K exact robust wait time
- $\widehat{W}_K$ approximation

let

$$\chi = \max_j \left\{ \frac{1}{\mu_j} + \Gamma_j^{\mathbb{X}} \right\}$$

for a hierarchical MCMS system,

$$W_K \leq \widehat{W}_K \leq W_K + 2\chi$$

Approximation Guarantee

- W_K exact robust wait time
- $\widehat{W}_K$ approximation

let

$$\chi = \max_j \left\{ \frac{1}{\mu_j} + \Gamma_j^{\mathbb{X}} \right\}$$

for a hierarchical MCMS system,

$$W_K \leq \widehat{W}_K \leq W_K + 2\chi$$

$\Rightarrow$ *approximation becomes tighter as n increases*

Heuristic: Performance

- computation times for different HMCMS instances

	general MIP	simpler MIP	SOCP
100 customers	1 sec	0.8 sec	0.8 sec
1,000 customers	< 1 min	< 1/2 min	1.2 sec
10,000 customers	6 min	2 min	5.4 sec
100,000 customers	40 min	10 min	< 1 min

Heuristic: Performance

- computation times for different HMCMS instances

	general MIP	simpler MIP	SOCP
100 customers	1 sec	0.8 sec	0.8 sec
1,000 customers	< 1 min	< 1/2 min	1.2 sec
10,000 customers	6 min	2 min	5.4 sec
100,000 customers	40 min	10 min	< 1 min

- heuristic approximation errors

50 customers	1.9%
100 customers	0.85%
1,000 customers	0.08%

Application to the KAS

Patient X of blood type O is listed in a given geographic region. He is currently ranked 50th. How long until he receives an offer of a particular quality?

Application to the KAS

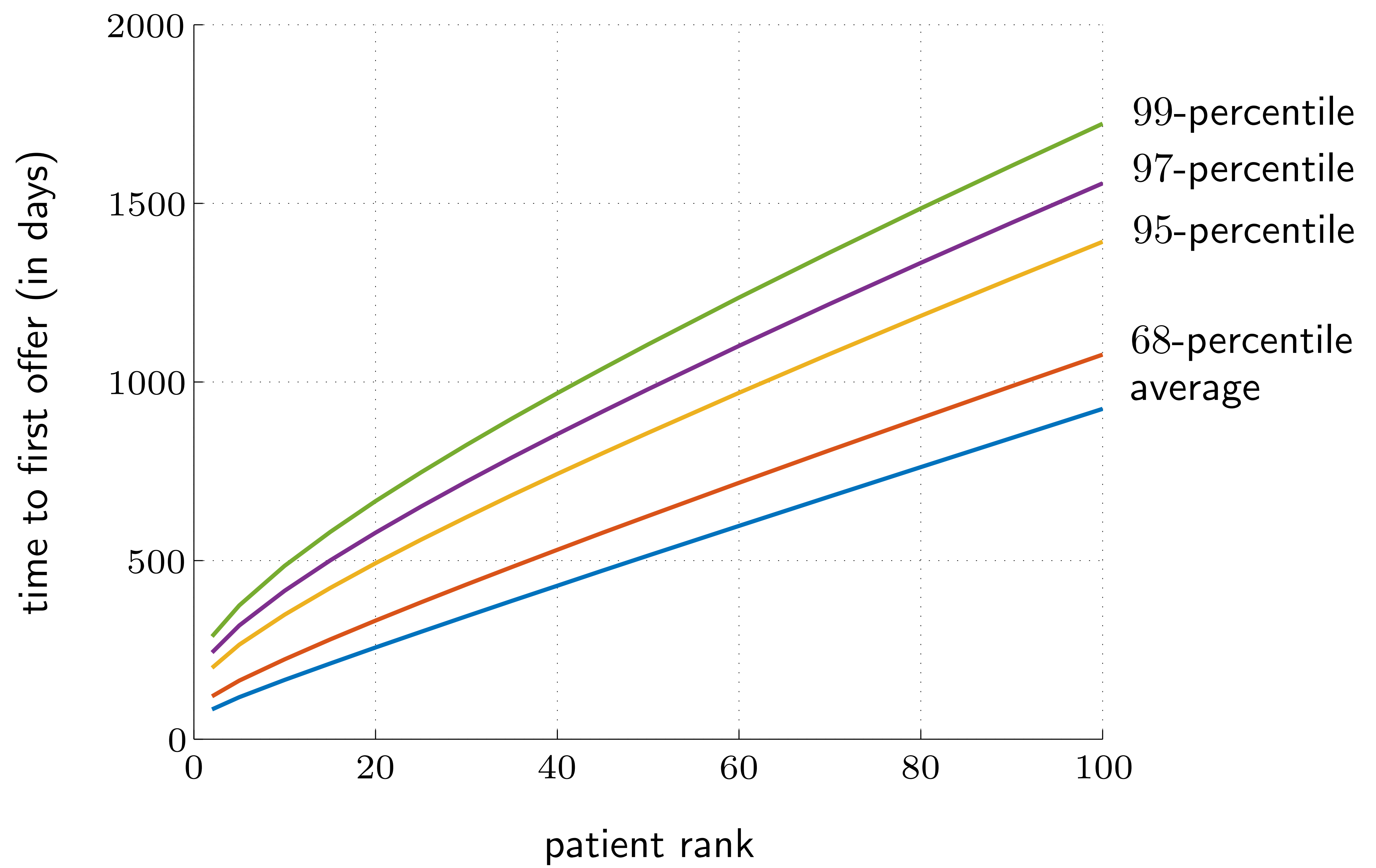
Patient X of blood type O is listed in a given geographic region. He is currently ranked 50th. How long until he receives an offer of a particular quality?

- PADV-OP1 Gift of Life Donor Program
- threshold type decisions
- model as HMCMS

Available Data

- well accepted kidney quality metric: KDPI
- historical kidney procurement rates (for each quality)
- historical patient accept/decline decisions
- 2007-2010 training set
- 2010-2013 testing set

Out-of-Sample Performance



Out-of-Sample Performance

- relative prediction errors
  - 14.96% for avg. and 11.73% for 68-percentile
- delay history estimator:
  - uses personalized info unavailable in practice
  - cannot estimate wait times for high ranks
- relative prediction errors of delay history estimator:
  - 16.76% for avg. and 14.65% for 68-percentile

Summary

- modeling framework for MCMS systems under incomplete information
- MIP formulation
  - more structure: provably tight scalable heuristic
- FCFS, class priority
- application to U.S. kidney allocation system

Thank you!

Model Calibration

- cluster kidneys in quality levels
- service time uncertainty sets: for each quality level j

$$\mathbb{X}_j = \left\{ x_j \in \mathbb{R}^{\bar{\ell}_j} \ : \ \sum_{k=1}^{\ell} x_j^k \leq \frac{\ell}{\mu_j} + \Gamma \sigma_j \sqrt{\ell}, \ \ell = 1, \ldots, \bar{\ell}_j \right\}$$

- μ_j (σ_j) historical procurement rate (std)

Model Calibration: Queue Populations

Model Calibration: Queue Populations

- patient is type i, observes rank r and historical accept/decline decisions

Model Calibration: Queue Populations

- patient is type i, observes rank r and historical accept/decline decisions
- q_k probability of a patient being type k

Model Calibration: Queue Populations

- patient is type i, observes rank r and historical accept/decline decisions
- q_k probability of a patient being type k
- fit q_k to maximize likelihood of observed decisions

Model Calibration: Queue Populations

- patient is type i, observes rank r and historical accept/decline decisions
- q_k probability of a patient being type k
- fit q_k to maximize likelihood of observed decisions
- CLT-based approach:

$$\sum_{\nu=1}^{r-1} \mathscr{L}_\nu \leq (r-1)\mu_{\mathscr{L}} + \Gamma \sigma_{\mathscr{L}} \sqrt{r-1}$$

class of νth patient

$$\sum_{i=1}^{K} i q_i \qquad \sum_{i=1}^{K} i^2 q_i - \mu_{\mathscr{L}}^2$$

Model Calibration: Queue Populations

- patient is type i, observes rank r and historical accept/decline decisions
- q_k probability of a patient being type k
- fit q_k to maximize likelihood of observed decisions
- CLT-based approach:

$$\mathbb{P} = \left\{ n \in \mathbb{R}^K : \sum_{i=1}^{K} i n_i - k \leq (r-1)\mu_{\mathscr{L}} + \Gamma \sigma_{\mathscr{L}} \sqrt{r-1} \right\}$$

Extensions

motivation:

- recent policy change: top 20% of healthier patients have priority for top 20% kidneys

modeling implications:

- alternative priority rule: <u>class priority</u>
- customer arrivals

all of our results can be extended!!

HMCMS under Class Priority

HMCMS under Class Priority

- model arrival times similar to service times

HMCMS under Class Priority

- model arrival times similar to service times
- v_k^ℓ arrival time of ℓth customer in class k

$$\sum y_{kj}^\ell \leq n_k \quad \text{becomes} \quad \sum y_{kj}^\ell \leq n_k + v_k^\ell$$

HMCMS under Class Priority

- model arrival times similar to service times
- v_k^ℓ arrival time of ℓth customer in class k

$$\sum y_{kj}^\ell \le n_k \quad \textbf{becomes} \quad \sum y_{kj}^\ell \le n_k + v_k^\ell$$

- constraints on assignments y_{kj}^ℓ reflect priority rules

HMCMS under Class Priority

- model arrival times similar to service times
- v_k^ℓ arrival time of ℓth customer in class k

$$\sum y_{kj}^\ell \leq n_k \quad \text{becomes} \quad \sum y_{kj}^\ell \leq n_k + v_k^\ell$$

- constraints on assignments y_{kj}^ℓ reflect priority rules
- e.g., if servers indexed in descending priority

$$y_{kj}^\ell \leq 1 - f_{k'}^\ell, \quad k' < k$$

HMCMS under Class Priority

- model arrival times similar to service times
- v_k^ℓ arrival time of ℓth customer in class k

$$\sum y_{kj}^\ell \leq n_k \quad \textbf{becomes} \quad \sum y_{kj}^\ell \leq n_k + v_k^\ell$$

- constraints on assignments y_{kj}^ℓ reflect priority rules
- e.g., if servers indexed in descending priority

$$y_{kj}^\ell \leq 1 - f_{k'}^\ell, \quad k' < k$$

- all our results can be extended