Optimal Probabilistic Forecasts for Counts

Brendan McCabe, Gael Martin and David Harris

Optimal Probabilistic Forecasts for Counts

Brendan McCabe, Gael Martin and David Harris

- Focus on **low** count time series

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$

Optimal Probabilistic Forecasts for Counts

Brendan McCabe, Gael Martin and David Harris

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$
- Particular interest in X_t as :

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$
- Particular interest in X_t as :
  - a queue

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$
- Particular interest in X_t as :
  - a queue
  - a stock (inventory)

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$
- Particular interest in X_t as :
  - a queue
  - a stock (inventory)
  - a birth and death process

Optimal Probabilistic Forecasts for Counts

Brendan McCabe, Gael Martin and David Harris

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$
- Particular interest in X_t as :
  - a queue
  - a stock (inventory)
  - a birth and death process
  - a branching process

- Focus on **low** count time series
- $\Rightarrow X_t = \{0, 1, 2...\}; \quad t = 1, 2, \ldots, T$
- Particular interest in X_t as :
    - a queue
    - a stock (inventory)
    - a birth and death process
    - a branching process
- Wish to produce 'optimal' probabilistic forecasts of X_t

- Enormous number of applications.....

- Enormous number of applications.....
- $\Rightarrow$ wide applicability

- Enormous number of applications.....
- $\Rightarrow$ wide applicability
- E.g. no. of 'iceberg' stock market order book entries

- Enormous number of applications.....
- $\Rightarrow$ wide applicability
- E.g. no. of 'iceberg' stock market order book entries
- $\Rightarrow$ Only a portion of the volume of the order

- Enormous number of applications.....
- $\Rightarrow$ wide applicability
- E.g. no. of 'iceberg' stock market order book entries
- $\Rightarrow$ Only a portion of the volume of the order
- or the 'tip of the iceberg', is revealed in the order book

- Enormous number of applications.....
- $\Rightarrow$ wide applicability
- E.g. no. of 'iceberg' stock market order book entries
- $\Rightarrow$ Only a portion of the volume of the order
- or the 'tip of the iceberg', is revealed in the order book
- $\Rightarrow$ 'hidden liquidity'

- Enormous number of applications.....
- $\Rightarrow$ wide applicability
- E.g. no. of 'iceberg' stock market order book entries
- $\Rightarrow$ Only a portion of the volume of the order
- or the 'tip of the iceberg', is revealed in the order book
- $\Rightarrow$ 'hidden liquidity'
- $\Rightarrow$ affects trading behaviour (Frey and Sandas, 2008)

Probabilistic forecasts

Probabilistic forecasts

- Continous approximation

Probabilistic forecasts

- Continous approximation

Probabilistic forecasts

- Continous approximation X

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}
- Defined only on the (non-negative) integer support

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}
- Defined only on the (non-negative) integer support
- Quantities of interest are:

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}
- Defined only on the (non-negative) integer support
- Quantities of interest are:

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}
- Defined only on the (non-negative) integer support
- Quantities of interest are:

$$f_i = P\left[X_{T+m} = i|\mathbf{x}\right], \ i = 0, 1, 2, \ldots$$

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}
- Defined only on the (non-negative) integer support
- Quantities of interest are:

$$
\begin{aligned}
f_i &= P\left[X_{T+m} = i|\mathbf{x}\right], \ i = 0, 1, 2, \ldots \\
\widehat{f}_i &= \widehat{P}\left[X_{T+m} = i|\mathbf{x}\right], \ i = 0, 1, 2, \ldots
\end{aligned}
$$

Probabilistic forecasts

- Continous approximation X
- Models and methods for discrete data
- Want predictions that are consistent with the discrete sample space
- $\Rightarrow$ focus on estimating the **predictive distribution of** X_{t+m}
- Defined only on the (non-negative) integer support
- Quantities of interest are:

$$
\begin{aligned}
f_i &= P\left[X_{T+m} = i|\mathbf{x}\right], \; i = 0, 1, 2, \dots \\
\widehat{f}_i &= \widehat{P}\left[X_{T+m} = i|\mathbf{x}\right], \; i = 0, 1, 2, \dots
\end{aligned}
$$

Optimal forecasts within model class

- Our aim:

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{ \widehat{f_i} \right\}$ via non-parametric MLE

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{\widehat{f}_i\right\}$ via non-parametric MLE
  - $\Rightarrow \left\{\widehat{f}_i\right\}$ **optimal** for any dgp (within class)

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ via non-parametric MLE
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ **optimal** for any dgp (within class)
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ appropriate choice

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ via non-parametric MLE
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ **optimal** for any dgp (within class)
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ appropriate choice
- Contrast with existing forecasting-evaluation literature:

Optimal forecasts within model class

- Our aim:
    - Define **broad** class of count model
    - appropriate for particular data types
    - $\Rightarrow \left\{ \widehat{f}_i \right\}$ via non-parametric MLE
    - $\Rightarrow \left\{ \widehat{f}_i \right\}$ **optimal** for any dgp (within class)
    - $\Rightarrow \left\{ \widehat{f}_i \right\}$ appropriate choice
- Contrast with existing forecasting-evaluation literature:
    - predictions treated as 'primitives'

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ via non-parametric MLE
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ **optimal** for any dgp (within class)
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ appropriate choice
- Contrast with existing forecasting-evaluation literature:
  - predictions treated as 'primitives'
  - model and inferential procedure (if any) not relevant

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{\widehat{f}_i\right\}$ via non-parametric MLE
  - $\Rightarrow \left\{\widehat{f}_i\right\}$ **optimal** for any dgp (within class)
  - $\Rightarrow \left\{\widehat{f}_i\right\}$ appropriate choice
- Contrast with existing forecasting-evaluation literature:
  - predictions treated as 'primitives'
  - model and inferential procedure (if any) not relevant
  - predictions only assessed via out-of-sample performance

Optimal forecasts within model class

- Our aim:
  - Define **broad** class of count model
  - appropriate for particular data types
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ via non-parametric MLE
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ **optimal** for any dgp (within class)
  - $\Rightarrow \left\{ \widehat{f}_i \right\}$ appropriate choice
- Contrast with existing forecasting-evaluation literature:
  - predictions treated as 'primitives'
  - model and inferential procedure (if any) not relevant
  - predictions only assessed via out-of-sample performance
- Could combine both approaches.....

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
-
$$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$
- $\alpha_k \circ X_{t-k}$ on $\{0, 1, 2, ...\}$; $k = 1, 2, \ldots, p$

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$
- $\alpha_k \circ X_{t-k}$ on $\{0, 1, 2, ...\}$; $k = 1, 2, \ldots, p$
- $\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k}$

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$
- $\alpha_k \circ X_{t-k}$ on $\{0, 1, 2, ...\}$; $k = 1, 2, \ldots, p$
- $\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k}$
- with $B_{1,k}, B_{2,k}, \ldots, B_{X_{t-k},k}$ *iid* Bernoulli:

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$
- $\alpha_k \circ X_{t-k}$ on $\{0, 1, 2, ...\}$; $k = 1, 2, \ldots, p$
- $\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k}$
- with $B_{1,k}, B_{2,k}, \ldots, B_{X_{t-k},k}$ *iid* Bernoulli:
- $P(B_{i,k} = 1) = \alpha_k$

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$
- $\alpha_k \circ X_{t-k}$ on $\{0, 1, 2, ...\}$; $k = 1, 2, \ldots, p$
- $\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k}$
- with $B_{1,k}, B_{2,k}, \ldots, B_{X_{t-k},k}$ *iid* Bernoulli:
- $P(B_{i,k} = 1) = \alpha_k$
- '$\circ$' binomial thinning

INAR(p)

- **Integer-valued autoregressive** models of Al-Osh and Alzaid (1987), McKenzie (1988), Du and Li (1991):
- $$X_t = \underbrace{\alpha_1 \circ X_{t-1}} + .. + \underbrace{\alpha_k \circ X_{t-k}} + .. + \underbrace{\alpha_p \circ X_{t-p}} + \underbrace{\varepsilon_t}$$
- ε_t *iid* on $\{0, 1, 2, ...\}$
- $\alpha_k \circ X_{t-k}$ on $\{0, 1, 2, ...\}$; $k = 1, 2, \dots, p$
- $\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k}$
- with $B_{1,k}, B_{2,k}, \dots, B_{X_{t-k},k}$ *iid* Bernoulli:
- $P(B_{i,k} = 1) = \alpha_k$
- '$\circ$' binomial thinning
- $\Rightarrow INAR(p)$ a branching process with immigration

INAR(1)

- When $p = 1$, X_t behaves like a **queue:**

$$X_t = \underbrace{\alpha_1 \circ X_{t-1}}_{\textit{survivors}} + \underbrace{\varepsilon_t}_{\textit{arrivals}}$$

INAR(1)

- When $p = 1$, X_t behaves like a **queue:**

$$X_t = \underbrace{\alpha_1 \circ X_{t-1}}_{survivors} + \underbrace{\varepsilon_t}_{arrivals}$$

- or a **birth and death** process

INAR(1)

- When $p = 1$, X_t behaves like a **queue:**

$$X_t = \underbrace{\alpha_1 \circ X_{t-1}}_{survivors} + \underbrace{\varepsilon_t}_{arrivals}$$

- or a **birth and death** process
- $\varepsilon_t =$ the births

INAR(1)

- When $p = 1$, X_t behaves like a **queue:**

$$X_t = \underbrace{\alpha_1 \circ X_{t-1}}_{survivors} + \underbrace{\varepsilon_t}_{arrivals}$$

- or a **birth and death** process
- ε_t = the births
- $\alpha_1 \circ X_{t-1}$ = the survivors (non-deaths)

INAR(1)

- When $p = 1$, X_t behaves like a **queue:**

$$X_t = \underbrace{\alpha_1 \circ X_{t-1}}_{survivors} + \underbrace{\varepsilon_t}_{arrivals}$$

- or a **birth and death** process
- ε_t = the births
- $\alpha_1 \circ X_{t-1}$ = the survivors (non-deaths)
- $INAR(p)$ a **broad class**

INAR(1)

- When $p = 1$, X_t behaves like a **queue:**

$$X_t = \underbrace{\alpha_1 \circ X_{t-1}}_{survivors} + \underbrace{\varepsilon_t}_{arrivals}$$

- or a **birth and death** process
- ε_t = the births
- $\alpha_1 \circ X_{t-1}$ = the survivors (non-deaths)
- $INAR(p)$ a **broad class**
- Many references in paper.......

Nonparametric prediction in the INAR(p) model

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$;

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$;

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$; e.g. $G = Poisson$

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$; e.g. $G = Poisson$
- $\Rightarrow$ MLE of $\{f_i\}$

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$; e.g. $G = Poisson$
- $\Rightarrow$ MLE of $\{f_i\}$
- **Non-parametric** MLE (NPMLE) imposes no structure on $\{g_r\}$

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$; e.g. $G = Poisson$
- $\Rightarrow$ MLE of $\{f_i\}$
- **Non-parametric** MLE (NPMLE) imposes no structure on $\{g_r\}$
- (other than $0 \leq g_r \leq 1, \sum_{r=0}^{\infty} g_r = 1$)

Nonparametric prediction in the INAR(p) model

- ε_t *iid* with distribution G
- $G = \{g_r\}$ is an infinite sequence of probabilities on the set $\mathbb{Z} = \{0, 1, 2, ...\}$
- MLE imposes parametric structure on $\{g_r\}$; e.g. $G = Poisson$
- $\Rightarrow$ MLE of $\{f_i\}$
- **Non-parametric** MLE (NPMLE) imposes no structure on $\{g_r\}$
- (other than $0 \leq g_r \leq 1, \sum_{r=0}^{\infty} g_r = 1$)
- $\Rightarrow$ NPMLE of $\{f_i\}$

Nonparametric prediction in the INAR(p) model

Nonparametric prediction in the INAR(p) model

- MLE of $\{f_i\}$ optimal only under **correct** distributional assumption

Nonparametric prediction in the INAR(p) model

- MLE of $\{f_i\}$ optimal only under **correct** distributional assumption
- NPMLE of $\{f_i\}$ shown to be **optimal** under **any** distributional assumption for ε_t

Nonparametric prediction in the INAR(p) model

- MLE of $\{f_i\}$ optimal only under **correct** distributional assumption
- NPMLE of $\{f_i\}$ shown to be **optimal** under **any** distributional assumption for ε_t
- Need to show:

Nonparametric prediction in the INAR(p) model

- MLE of $\{f_i\}$ optimal only under **correct** distributional assumption
- NPMLE of $\{f_i\}$ shown to be **optimal** under **any** distributional assumption for ε_t
- Need to show:

1. Optimality of NPMLE of $\theta = (\alpha_1, ..., \alpha_p, \{g_r\})$

Nonparametric prediction in the INAR(p) model

- MLE of $\{f_i\}$ optimal only under **correct** distributional assumption
- NPMLE of $\{f_i\}$ shown to be **optimal** under **any** distributional assumption for ε_t
- Need to show:

1. Optimality of NPMLE of $\theta = (\alpha_1, ..., \alpha_p, \{g_r\})$
2. 'Smoothness' of map between θ and $\{f_i\} \Rightarrow$ Optimality of NPMLE of $\{f_i\}$

Nonparametric prediction in the INAR(p) model

- MLE of $\{f_i\}$ optimal only under **correct** distributional assumption
- NPMLE of $\{f_i\}$ shown to be **optimal** under **any** distributional assumption for ε_t
- Need to show:

1. Optimality of NPMLE of $\theta = (\alpha_1, ..., \alpha_p, \{g_r\})$
2. 'Smoothness' of map between θ and $\{f_i\} \Rightarrow$ Optimality of NPMLE of $\{f_i\}$

- $\{g_r\}$ (and hence θ) and $\{f_i\}$ are of **infinite** dimension

Implementation of NPMLE

Implementation of NPMLE

- Consider $X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t$

Implementation of NPMLE

- Consider $X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t$
- Conditional on x_1 :

$$\begin{aligned}
\log L(\theta) &= \sum_{t=2}^{T} \log \left\{ \Pr(X_t = x_t | X_{t-1} = x_{t-1}) \right\} \\
&= \sum_{t=2}^{T} \log \left\{ \sum_{r=\max(0,\Delta x_t)}^{x_t} p^B_{x_t - r} g_r \right\} \\
p^B_{x_t - r} &= Bin(\alpha \circ X_{t-1} = x_t - r | X_{t-1} = x_{t-1})
\end{aligned}$$

Implementation of NPMLE

- Consider $X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t$
- Conditional on x_1 :

$$\begin{aligned}
\log L(\theta) &= \sum_{t=2}^{T} \log \left\{ \Pr(X_t = x_t | X_{t-1} = x_{t-1}) \right\} \\
&= \sum_{t=2}^{T} \log \left\{ \sum_{r=\max(0,\Delta x_t)}^{x_t} p^B_{x_t - r} g_r \right\} \\
p^B_{x_t - r} &= Bin(\alpha \circ X_{t-1} = x_t - r | X_{t-1} = x_{t-1})
\end{aligned}$$

- Conditional binomials **mixed** over arrivals

Implementation of NPMLE

- Consider $X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t$
- Conditional on x_1 :

$$
\begin{aligned}
\log L(\theta) &= \sum_{t=2}^{T} \log \left\{ \Pr(X_t = x_t | X_{t-1} = x_{t-1}) \right\} \\
&= \sum_{t=2}^{T} \log \left\{ \sum_{r=\max(0,\Delta x_t)}^{x_t} p^B_{x_t - r} g_r \right\} \\
p^B_{x_t - r} &= Bin(\alpha \circ X_{t-1} = x_t - r | X_{t-1} = x_{t-1})
\end{aligned}
$$

- Conditional binomials **mixed** over arrivals
- Estimate $\{g_r\}$ and α_1 via (constrained) ML

Implementation of NPMLE

- Consider $X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t$
- Conditional on x_1 :

$$\begin{aligned}
\log L(\theta) &= \sum_{t=2}^{T} \log\left\{\Pr(X_t = x_t | X_{t-1} = x_{t-1})\right\} \\
&= \sum_{t=2}^{T} \log\left\{ \sum_{r=\max(0,\Delta x_t)}^{x_t} p^B_{x_t - r} g_r \right\} \\
p^B_{x_t - r} &= Bin(\alpha \circ X_{t-1} = x_t - r | X_{t-1} = x_{t-1})
\end{aligned}$$

- Conditional binomials **mixed** over arrivals
- Estimate $\{g_r\}$ and α_1 via (constrained) ML
- $\Rightarrow$ NPMLE: $\hat{\theta} = (\widehat{\alpha}_1, \{\hat{g}_r\})$

Optimality of NPMLE

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood
- $\{\hat{g}_r\}$ contains only a finite number of non-zero values in finite samples

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood
- $\{\hat{g}_r\}$ contains only a finite number of non-zero values in finite samples
- $\Rightarrow \{\hat{g}_r\}$ (and $\widehat{\theta}$) infinite as $T \rightarrow \infty$

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood
- $\{\hat{g}_r\}$ contains only a finite number of non-zero values in finite samples
- $\Rightarrow \{\hat{g}_r\}$ (and $\widehat{\theta}$) infinite as $T \longrightarrow \infty$
- Asymptotic theory needs to accommodate this

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood
- $\{\hat{g}_r\}$ contains only a finite number of non-zero values in finite samples
- $\Rightarrow \{\hat{g}_r\}$ (and $\widehat{\theta}$) infinite as $T \rightarrow \infty$
- Asymptotic theory needs to accommodate this
- Drost et al (2008) $\Rightarrow$

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood
- $\{\hat{g}_r\}$ contains only a finite number of non-zero values in finite samples
- $\Rightarrow \{\hat{g}_r\}$ (and $\widehat{\theta}$) infinite as $T \longrightarrow \infty$
- Asymptotic theory needs to accommodate this
- Drost et al (2008) $\Rightarrow$
- asy. Gaussianity and (non-parametric) asy. efficiency of $\hat{\theta}$

Optimality of NPMLE

- **Formally:** maximizing an **empirical** likelihood
- $\{\hat{g}_r\}$ contains only a finite number of non-zero values in finite samples
- $\Rightarrow \{\hat{g}_r\}$ (and $\widehat{\theta}$) infinite as $T \to \infty$
- Asymptotic theory needs to accommodate this
- Drost et al (2008) $\Rightarrow$
- asy. Gaussianity and (non-parametric) asy. efficiency of $\hat{\theta}$
- $\Rightarrow \hat{\theta}$ optimal in this sense

Optimality of forecasts

- $\theta \Rightarrow \{f_i(\theta)\}$

Optimality of forecasts

- $\theta \Rightarrow \{f_i(\theta)\}$
- $\Rightarrow \left\{\widehat{f_i}(\hat{\theta})\right\}$ asy. Gaussian and (non-parametric) asy. efficient

Optimality of forecasts

- $\theta \Rightarrow \{f_i(\theta)\}$
- $\Rightarrow \left\{\widehat{f}_i(\hat{\theta})\right\}$ asy. Gaussian and (non-parametric) asy. efficient
- Involves showing that the map is (Frechet) differentiable; i.e. that the derivative $\dot{F}$ is a **bounded, linear** operator with

$$\left\| F(\theta + h) - F(\theta) - \dot{F}(h) \right\|_{\ell^1} = o\left(\|h\|_{\mathbb{H}}\right)$$

Optimality of forecasts

- Theorems 1 and 2, plus proofs

Optimality of forecasts

- Theorems 1 and 2, plus proofs
- $\Rightarrow$ linear operations on (asy) Gaussian variables are Gaussian

Optimality of forecasts

- Theorems 1 and 2, plus proofs
- $\Rightarrow$ linear operations on (asy) Gaussian variables are Gaussian
- 'delta' rule $\Rightarrow$ NPMLE of $\{f_i\}$ asy. Gaussian and (non-parametric) asy. efficient

Optimality of forecasts

- Theorems 1 and 2, plus proofs
- $\Rightarrow$ linear operations on (asy) Gaussian variables are Gaussian
- 'delta' rule $\Rightarrow$ NPMLE of $\{f_i\}$ asy. Gaussian and (non-parametric) asy. efficient
- Also performs well in finite samples

Optimality of forecasts

- Theorems 1 and 2, plus proofs
- $\Rightarrow$ linear operations on (asy) Gaussian variables are Gaussian
- 'delta' rule $\Rightarrow$ NPMLE of $\{f_i\}$ asy. Gaussian and (non-parametric) asy. efficient
- Also performs well in finite samples
- Especially in tail ($\equiv$ rare occurrences of high counts)

Measurement of sampling error

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f_i}\}$?

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f_i}\}$?
- Need to impose $\sum_i \widehat{f_i} = 1$

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\}$

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\}$

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f_i}\}$?
- Need to impose $\sum_i \widehat{f_i} = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f_i}\} \Rightarrow \{\widehat{f_{s,i}}\}$, $s = 1, 2, .., N_s$;

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f_i}\}$?
- Need to impose $\sum_i \widehat{f_i} = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f_i}\} \Rightarrow \{\widehat{f_{s,i}}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f_{s,i}} = 1$)

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\} \Rightarrow \{\widehat{f}_{s,i}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f}_{s,i} = 1$)
  2. Calculate the "distance" between the single **empirical estimate**, $\{\widehat{f}_i\}$, and its subsampled counterpart using a metric: e.g. $d = \sqrt{T} \sum_{i=0}^{K} \left| \hat{f}_i - \widehat{f}_{s,i} \right|$

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\} \Rightarrow \{\widehat{f}_{s,i}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f}_{s,i} = 1$)
  2. Calculate the "distance" between the single **empirical estimate**, $\{\widehat{f}_i\}$, and its subsampled counterpart using a metric: e.g. $d = \sqrt{T} \sum_{i=0}^{K} \left| \hat{f}_i - \widehat{f}_{s,i} \right|$

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\} \Rightarrow \{\widehat{f}_{s,i}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f}_{s,i} = 1$)
  2. Calculate the "distance" between the single **empirical estimate**, $\{\widehat{f}_i\}$, and its subsampled counterpart using a metric: e.g. $d = \sqrt{T} \sum_{i=0}^{K} \left| \hat{f}_i - \widehat{f}_{s,i} \right|$
  3. Ranking (in ascending order) of **metric** d

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\} \Rightarrow \{\widehat{f}_{s,i}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f}_{s,i} = 1$)
  2. Calculate the "distance" between the single **empirical estimate**, $\{\widehat{f}_i\}$, and its subsampled counterpart using a metric: e.g. $d = \sqrt{T} \sum_{i=0}^{K} \left| \hat{f}_i - \widehat{f}_{s,i} \right|$
  3. Ranking (in ascending order) of **metric** d

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f_i}\}$?
- Need to impose $\sum_i \widehat{f_i} = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f_i}\} \Rightarrow \{\widehat{f_{s,i}}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f_{s,i}} = 1$)
  2. Calculate the "distance" between the single **empirical estimate**,$\{\widehat{f_i}\}$, and its subsampled counterpart using a metric: e.g. $d = \sqrt{T} \sum_{i=0}^{K} \left| \hat{f}_i - \widehat{f_{s,i}} \right|$
  3. Ranking (in ascending order) of **metric** $d \Rightarrow$ ranking of the subsampled distributions

Measurement of sampling error

- How to measure sampling variation in $\{\widehat{f}_i\}$?
- Need to impose $\sum_i \widehat{f}_i = 1$
- Use **subsampling** method to:
  1. Take draws from (an approximation to) the sampling distribution of $\{\widehat{f}_i\} \Rightarrow \{\widehat{f}_{s,i}\}$, $s = 1, 2, .., N_s$; (Each sub-sampled distribution is proper: $\sum_i \widehat{f}_{s,i} = 1$)
  2. Calculate the "distance" between the single **empirical estimate**, $\{\widehat{f}_i\}$, and its subsampled counterpart using a metric: e.g. $d = \sqrt{T} \sum_{i=0}^{K} \left| \hat{f}_i - \widehat{f}_{s,i} \right|$
  3. Ranking (in ascending order) of **metric** $d \Rightarrow$ ranking of the subsampled distributions

  - Subsample estimator of sampling distribution of $\{\widehat{f}_i\}$ **consistent**

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)

‘Iceberg’ orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- ‘Iceberg’ asks in the order book (up to and including the fifth best order only)

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
    - and the number of orders remaining from the previous ten minute period, waiting for execution:

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
    - and the number of orders remaining from the previous ten minute period, waiting for execution:

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
    - and the number of orders remaining from the previous ten minute period, waiting for execution: $\alpha_1 \circ X_{t-1}$

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
  - and the number of orders remaining from the previous ten minute period, waiting for execution: $\alpha_1 \circ X_{t-1}$
  - the number of new iceberg orders placed in the book (or 'arrivals'):

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
  - and the number of orders remaining from the previous ten minute period, waiting for execution: $\alpha_1 \circ X_{t-1}$
  - the number of new iceberg orders placed in the book (or 'arrivals'):

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
    - and the number of orders remaining from the previous ten minute period, waiting for execution: $\alpha_1 \circ X_{t-1}$
    - the number of new iceberg orders placed in the book (or 'arrivals'): ε_t

'Iceberg' orders

- Stock: Deutsche Telekom (traded on the Deutsche Borse, 2004)
- 'Iceberg' asks in the order book (up to and including the fifth best order only)
- Counted every 10 minutes
- Over any 10 minute time period t, the number of iceberg orders, X_t, is the sum of:
  - and the number of orders remaining from the previous ten minute period, waiting for execution: $\alpha_1 \circ X_{t-1}$
  - the number of new iceberg orders placed in the book (or 'arrivals'): ε_t
- $\Rightarrow INAR(1)$

Prediction of ‘iceberg’ orders

- Use $T = 500$ sample to predict X_{T+1} :

Prediction of 'iceberg' orders

- Use $T = 500$ sample to predict X_{T+1} :
- Plot **empirical** $\widehat{f}_i = \text{Prob}(X_{T+1} = i|\mathbf{x})$; $i = 0, 1, 2, ...$

Prediction of 'iceberg' orders

- Use $T = 500$ sample to predict X_{T+1} :
- Plot **empirical** $\widehat{f_i} = \text{Prob}(X_{T+1} = i|\mathbf{x})$; $i = 0, 1, 2, ...$
- Plot 5 **extreme** subsampled $\{\widehat{f_i}\}'s$

Prediction of 'iceberg' orders

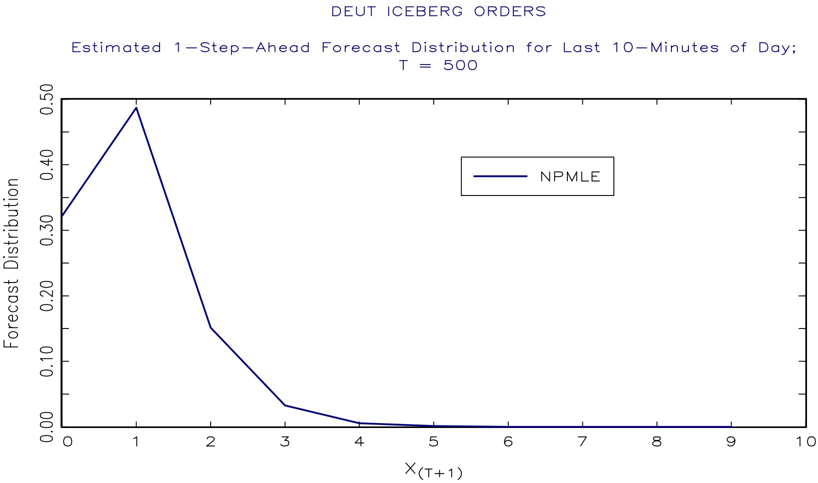
- Use $T = 500$ sample to predict X_{T+1} :
- Plot **empirical** $\widehat{f_i} = \text{Prob}(X_{T+1} = i|\mathbf{x})$; $i = 0, 1, 2, ...$
- Plot 5 **extreme** subsampled $\{\widehat{f_i}\}'s$
- at $95th$ percentiles of metric and two distributons either side

Prediction of 'iceberg' orders

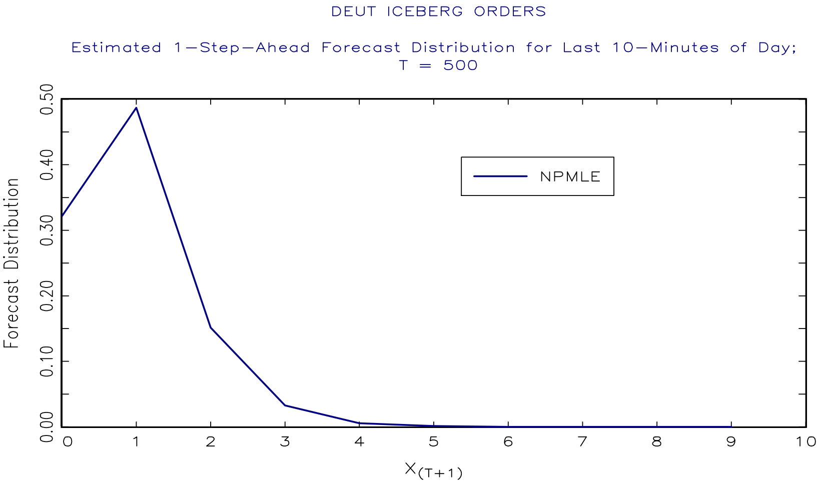
- Use $T = 500$ sample to predict X_{T+1} :
- Plot **empirical** $\widehat{f_i} = \text{Prob}(X_{T+1} = i|\mathbf{x})$; $i = 0, 1, 2, ...$
- Plot 5 **extreme** subsampled $\{\widehat{f_i}\}'s$
- at $95th$ percentiles of metric and two distributons either side
- What do extreme distributional estimates look like?

Prediction of ‘iceberg’ orders

- Use $T = 500$ sample to predict X_{T+1} :
- Plot **empirical** $\widehat{f_i} = \text{Prob}(X_{T+1} = i|\mathbf{x})$; $i = 0, 1, 2, ...$
- Plot 5 **extreme** subsampled $\{\widehat{f_i}\}'s$
- at $95th$ percentiles of metric and two distributons either side
- What do extreme distributional estimates look like?
- How different could our probabilitistic predictions be?

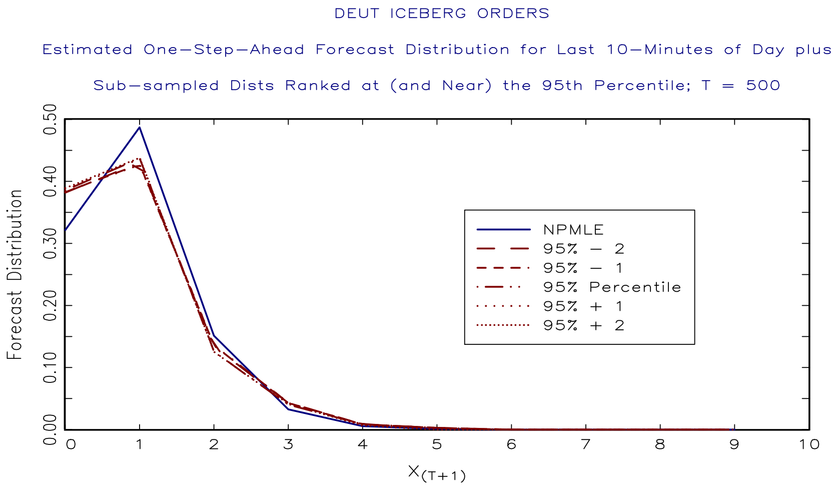


- $\mathrm{Prob}(X_{T+1} \geq 1) = 78\%$

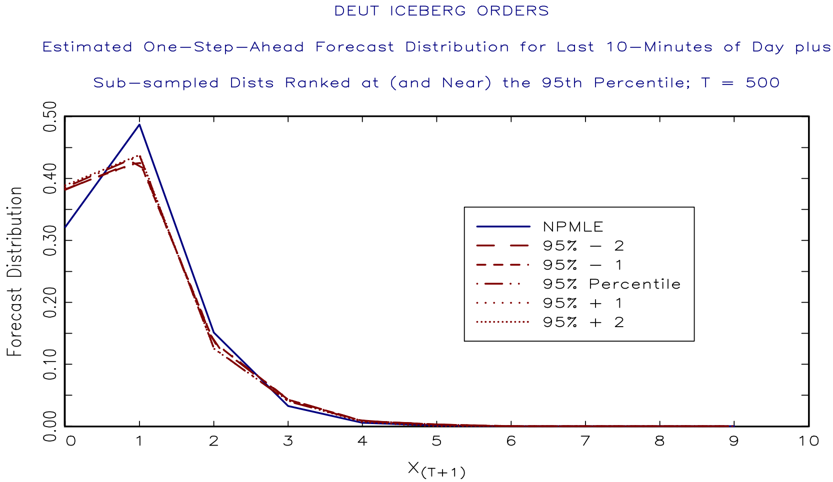


- $\text{Prob}(X_{T+1} \geq 1) = 78\%$
- $\Rightarrow$ high prob. of some hidden liquidity

- Extreme estimates?

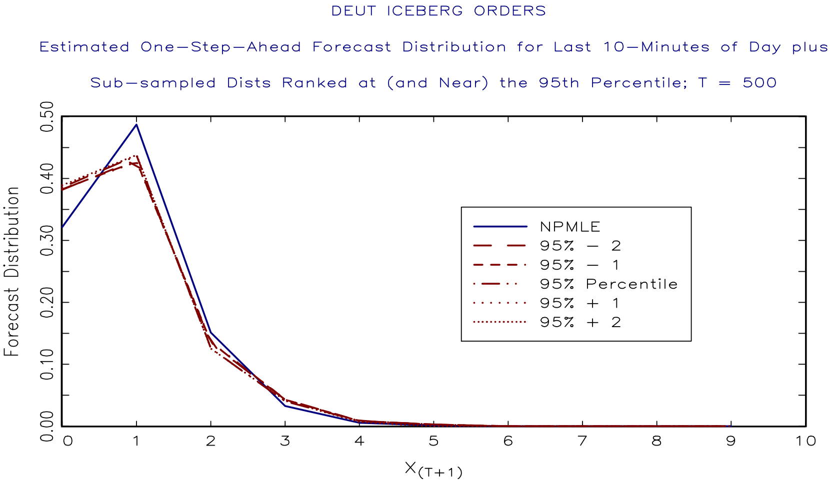


- Extreme estimates?



- $93rd$ - $97th \Rightarrow \text{Prob}(X_{T+1} \geq 1)$ lower

- Extreme estimates?



- $93rd$ - $97th \Rightarrow \text{Prob}(X_{T+1} \geq 1)$ lower
- $\Rightarrow$ sampling variability shifts prob. mass across support

Enough for 20 minutes........