# Sharp oracle inequalities and universality of the AIC and FPE

Georg Köstenberger
joint work with Moritz Jirak

17.09.2025

Department of Statistics and Operations Research,
University of Vienna

# Structure of the talk

**Part 1:** Why does the AIC work?

**Part 2:** Can we compare the AIC of two models?

# Why does the AIC work?

# Motivation: Prediction Problem

**Problem:** Given $X_1, \ldots, X_n$, predict $X_{n+1}$.

**"Optimal" solution:**

$$\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) = \operatorname*{argmin}_{Z \in \mathcal{M}(\sigma(X_1,\ldots,X_n))} \mathbb{E}(X_{n+1} - Z)^2.$$

**Classical approach:** Fit linear models, i.e., predict $X_{n+1}$ using

$$X_{n+1} \approx \sum_{j=1}^{k} \widehat{a}_j(k) X_{n+1-j}.$$

# Motivation: Prediction Problem

**Problem:** Given $X_1, \ldots, X_n$, predict $X_{n+1}$.

**"Optimal" solution:**

$$\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) = \operatorname*{argmin}_{Z \in \mathcal{M}(\sigma(X_1,\ldots,X_n))} \mathbb{E}(X_{n+1} - Z)^2.$$

**Classical approach:** Fit linear models, i.e., predict $X_{n+1}$ using

$$X_{n+1} \approx \sum_{j=1}^{k} \widehat{a}_j(k) X_{n+1-j}.$$

**Question:** How do we pick $k$ and $\widehat{a}(k)$?

# Motivation: Prediction Problem

**Problem:** Given $X_1, \ldots, X_n$, predict $X_{n+1}$.

**"Optimal" solution:**

$$\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) = \operatorname*{argmin}_{Z \in \mathcal{M}(\sigma(X_1,\ldots,X_n))} \mathbb{E}(X_{n+1} - Z)^2.$$

**Classical approach:** Fit linear models, i.e., predict $X_{n+1}$ using

$$X_{n+1} \approx \sum_{j=1}^{k} \widehat{a}_j(k) X_{n+1-j}.$$

**Question:** How do we pick $k$ and $\widehat{a}(k)$? Does this work?

# Motivation: Prediction Problem

**Problem:** Given $X_1, \ldots, X_n$, predict $X_{n+1}$.

**"Optimal" solution:**

$$\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) = \operatorname*{argmin}_{Z \in \mathcal{M}(\sigma(X_1,\ldots,X_n))} \mathbb{E}(X_{n+1} - Z)^2.$$

**Classical approach:** Fit linear models, i.e., predict $X_{n+1}$ using

$$X_{n+1} \approx \sum_{j=1}^{k} \widehat{a}_j(k) X_{n+1-j}.$$

**Question:** How do we pick $k$ and $\widehat{a}(k)$? Does this work? What does "work" mean in this context?

# Two ways to measure success

1. **Asymptotic efficiency:**

$$\frac{\text{Model selection-based prediction error}}{\text{Optimal Oracle-based prediction error}} \xrightarrow[n\to\infty]{\mathbb{P}} \text{Efficiency} \in [1, \infty].$$

2. **Sharp finite-sample oracle inequalities:**

$$\begin{aligned}&\text{Model selection-based prediction error}\\ &\leq \text{Optimal Oracle-based prediction error} \times \big(1 + o(1)\big)\end{aligned}$$

with high probability.

# Model selection via AIC

1. Choose $K_n \in \{1, \ldots, n-1\}$, and estimate $\widehat{a}(k)$ for $k = 1, \ldots, K_n$.
2. Choose the $k = \widehat{k}_n \in \{1, \ldots, K_n\}$ for which

$$\mathrm{AIC}(k) = n \log(\widehat{\sigma}_k^2) + 2k$$

   is minimal, where

$$\widehat{\sigma}_k^2 = \frac{1}{n - K_n} \sum_{t=K_n+1}^{n} \left( X_t - \sum_{j=1}^{k} \widehat{a}_j(k) X_{t-j} \right)^2.$$

3. Predict $X_{n+1}$ using

$$X_{n+1} \approx \sum_{j=1}^{\widehat{k}_n} \widehat{a}_j(\widehat{k}_n) X_{n+1-j}.$$

# Model selection via AIC

**Problems:**

1. Classical theory only covers linear processes? What about other dynamics such as GARCH, Markov Chains, SDEs, etc?
2. Finite sample behavior?
3. Assumptions are currently not falsifiable.

# Table of Contents: Part 1

1. Setup and Oracle.
2. Model selection.
3. State of the art.
4. Contribution.

# Setup and Oracle

# Setup

We are given a sample $X_1, \ldots, X_n$ from a stationary process $X = (X_t)_{t \in \mathbb{Z}}$.

For $t \in \mathbb{Z}$, let $V = \overline{\langle X_s \mid s < t \rangle} \subseteq L^2$. Then

$$P_V(X_t) = \sum_{j=1}^{\infty} a_j X_{t-j}.$$

The vector $a = (a_j)_{j \geq 1}$ represents the *best possible* linear model.

**Goal:** Approximate $a$.

# Setup

For $t, k \in \mathbb{N}$, we set $V_k = \langle X_{t-1}, \ldots, X_{t-k} \rangle \subseteq L^2$. We have

$$P_{V_k}(X_t) = \sum_{j=1}^{k} a_j(k) X_{t-j}.$$

The vector $a(k) = (a_1(k), \ldots a_k(k))^T$ represents the best linear model of dimension $k$.

**Idea:** Approximate $a$ by $a(k)$, and estimate $a(k)$ from the data.

# Setup

**Question:** How do we estimate $a(k)$?

By Yule-Walker theory, we have

$$a(k) = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \, \mathbb{E}\left( X_t - \sum_{j=1}^{k} b_j X_{t-j} \right)^2 = R(k)^{-1} r(k),$$

where $R(k) = (R_{ij})_{i,j=1}^{k}$, $r(k) = (R_{0j})_{j=1}^{k}$ and $R_{ij} = \mathbb{E}(X_i X_j)$.

# Setup

**Question:** How do we estimate $a(k)$?

By Yule-Walker theory, we have

$$a(k) = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \, \mathbb{E}\left( X_t - \sum_{j=1}^{k} b_j X_{t-j} \right)^2 = R(k)^{-1} r(k),$$

where $R(k) = (R_{ij})_{i,j=1}^{k}$, $r(k) = (R_{0j})_{j=1}^{k}$ and $R_{ij} = \mathbb{E}(X_i X_j)$.

Use plug-in estimation!

# Setup: Estimating $a(k)$

For $k = 1, \ldots, K_n$, we set

$$\widehat{a}(k) = \widehat{R}(k)^{-1}\widehat{r}(k),$$

where $\widehat{R}(k) = (\widehat{R}_{ij})_{i,j=1}^{k}$, $\widehat{r}(k) = (\widehat{R}_{0,j})_{j=1}^{k}$ and

$$\widehat{R}_{ij} = \frac{1}{n - K_n} \sum_{t=K_n+1}^{n} X_{t-i} X_{t-j}.$$

# Recap

**Sharp finite-sample oracle inequalities:**

$$\begin{aligned}&\text{Model selection-based prediction error}\\ &\leq \text{Optimal Oracle-based prediction error} \times \big(1 + o(1)\big)\end{aligned}$$

with high probability.

We have an *estimator* $\widehat{a}(k)$ of $a$.

**Question:** Prediction error? Oracle?

# Prediction error and oracle

**Question:** How close is $\widehat{a}(k)$ to $a$? What does "close" mean?

We measure in the *intrinsic units* of the process $X_t$, i.e., we use the norm

$$\|z\|_R^2 = \sum_{i,j=1}^{\infty} z_i z_j R_{ij},$$

for $z \in \ell^2$.

Our prediction error is

$$Q_n(k) = \|a - \widehat{a}(k)\|_R^2.$$

# Prediction error and oracle

We have the following Bias-Covariance tradeoff

$$Q_n(k) = \|a - \widehat{a}(k)\|_R^2 = \|a - a(k)\|_R^2 + \|a(k) - \widehat{a}(k)\|_R^2.$$

The oracle is given by

$$L_n(k) = \|a - a(k)\|_R^2 + \sigma^2 \frac{k}{n - K_n},$$

where $\sigma^2 = \mathbb{E}\big(X_0 - \sum_{j=1}^{\infty} a_j X_{-j}\big)^2$, and the **oracle model order** is any

$$k_n^* \in \operatorname*{argmin}_{1 \le k \le K_n} L_n(k).$$

# Prediction error and oracle

**Question:** What does it mean to be an *oracle*?

**Theorem (informal)**
*If the problem is well-posed (e.g. $K_n \to \infty$), than for any sequence $\tilde{k}_n$ of $\sigma(X_1, \ldots, X_n)$ measurable functions, and any $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}\left( \frac{Q_n(\tilde{k}_n)}{L_n(k_n^*)} > 1 - \varepsilon \right) = 1.$$

# Model selection

# AIC vs. BIC

| | True model among candidates? | |
|---|---|---|
| | Yes | No |
| AIC | may not select true model | AIC selects approximation of the true model with asymptotically optimal bias-variance tradeoff |
| BIC | $\mathbb{P}(\text{BIC selects true model}) \to 1$ | may yield suboptimal approximation |

# AIC and FPE

**1969:** Final Prediction Error (FPE)

$$FPE(k) = \frac{n+k}{n-k}\widehat{\sigma}_k^2,$$

where

$$\widehat{\sigma}_k^2 = \frac{1}{n-K_n}\sum_{m=K_n+1}^{n}\left(X_m - \sum_{j=1}^{k}\widehat{a}_j(k)X_{m-j}\right)^2$$

**1973/4:** Akaike Information Criterion (AIC)

$$AIC(k) = n\log(\widehat{\sigma}_k^2) + 2k.$$

# AIC and FPE

**1969:** Final Prediction Error (FPE)

$$FPE(k) = \frac{n+k}{n-k}\widehat{\sigma}_k^2,$$

where

$$\widehat{\sigma}_k^2 = \frac{1}{n-K_n}\sum_{m=K_n+1}^{n}\left(X_m - \sum_{j=1}^{k}\widehat{a}_j(k)X_{m-j}\right)^2$$

**1973/4:** Akaike Information Criterion (AIC)

$$AIC(k) = n\log(\widehat{\sigma}_k^2) + 2k.$$

**Question:** What guarantees do we have?

# Classical literature

# Classical literature

**1980:** Justification of FPE, AIC (and other criteria) in a prediction setting by R. Shibata in terms of **asymptotic efficiency**. Requires the innovations $e_t$ in

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + e_t$$

to be i.i.d. Gaussian.

**1995-2001:** A. Karagrigoriou & S. Lee relax assumptions on $e_t$ to i.i.d. with finite 8th moments.

**2003,2005:** C.-K. Ing & C.-Z. Wei introduce *same-realization* setting (still require i.i.d. innovations $e_t$).

**2006:** E. J. Candès (among others) popularizes usage of *oracle inequalities* (see also [Barron, Birgé, and Massart, PTRF, 1999]).

# Classical literature: shortcomings

1. Innovations $e_t$ need to be i.i.d. This excludes many frequently used models (e.g. GARCH, Markov processes, SDEs, etc.).

2. Assumptions are imposed on the unobservable innovations $e_t$ (rather than $X_t$), and are thus impossible to check in practice.

3. No finite-sample oracle inequalities. Everything is purely asymptotic.

# Contribution

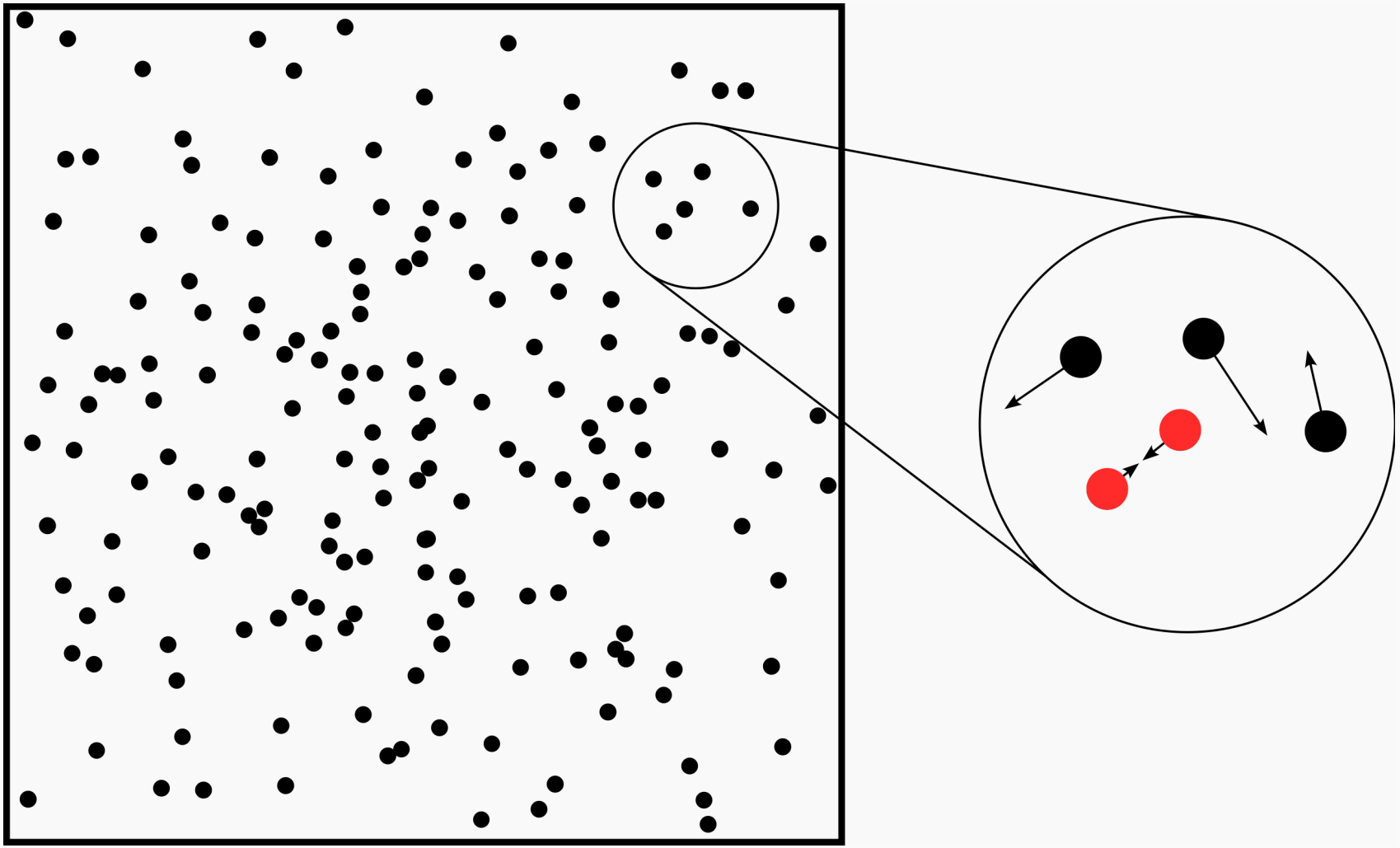
# Assumptions

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process. Then

1. $\mathbb{E}(X_t) = 0$ and $X_t \in L^q$, for $q > 8$.
2. $X_t$ does not degenerate to a finite order autoregressive process.
3. The spectral density $f_X$ is (uniformly) bounded away from zero.
4. $K_n \in \{1, \ldots, n-1\}$ is a divergent sequence of integers, and there is $\kappa > 0$ such that $K_n^{2+\kappa}/n$ is bounded.
5. $X_t$ is weakly (physically) dependent.

# Physical dependence: intuition

**Goal:** Measure temperature of gas.



# Physical Dependence

Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be an $E$-valued i.i.d. sequence, $g_t : E^\infty \to \mathbb{R}$ measurable.

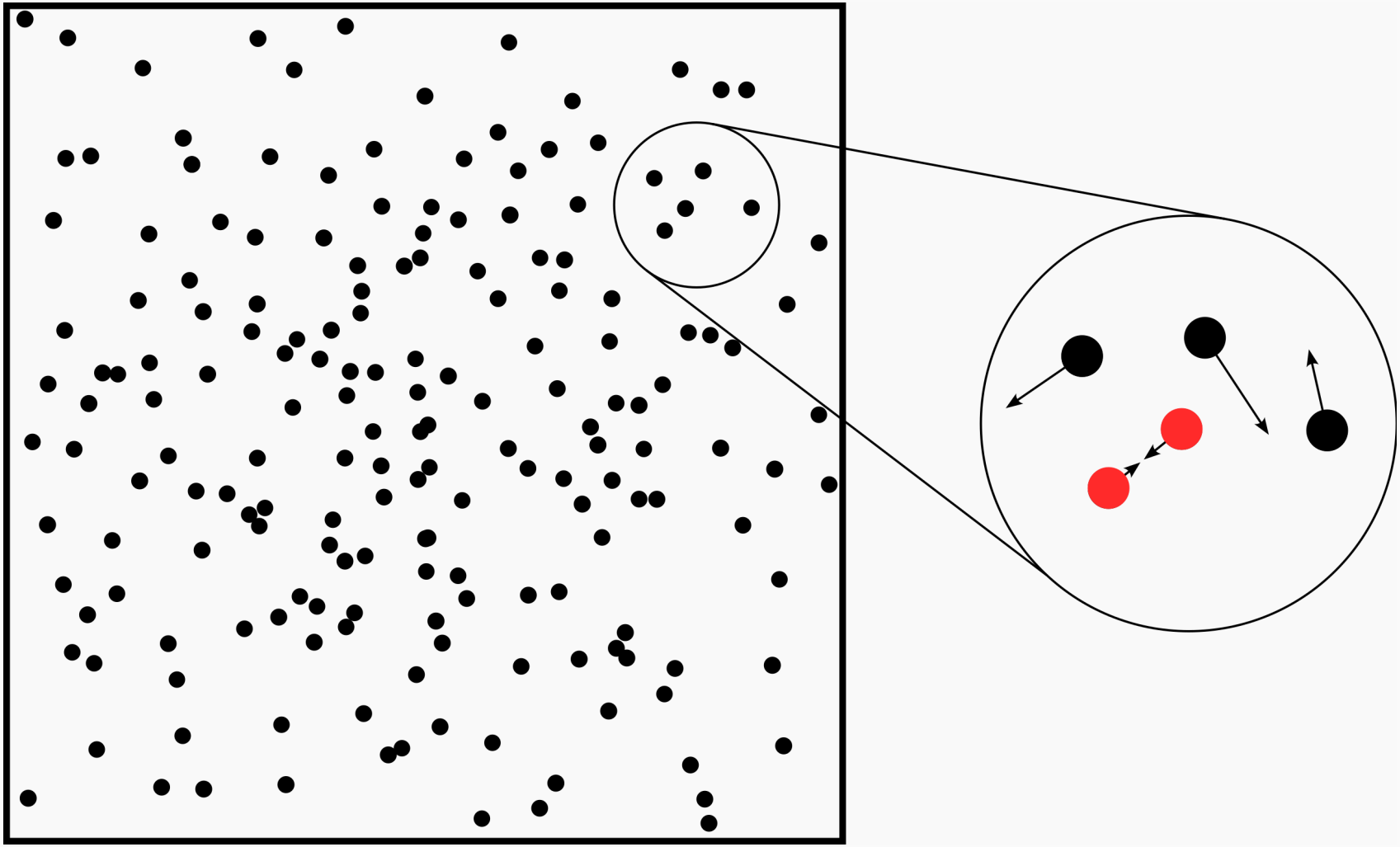
We assume

$$X_t = g_t(\varepsilon_t, \varepsilon_{t-1}, \dots).$$

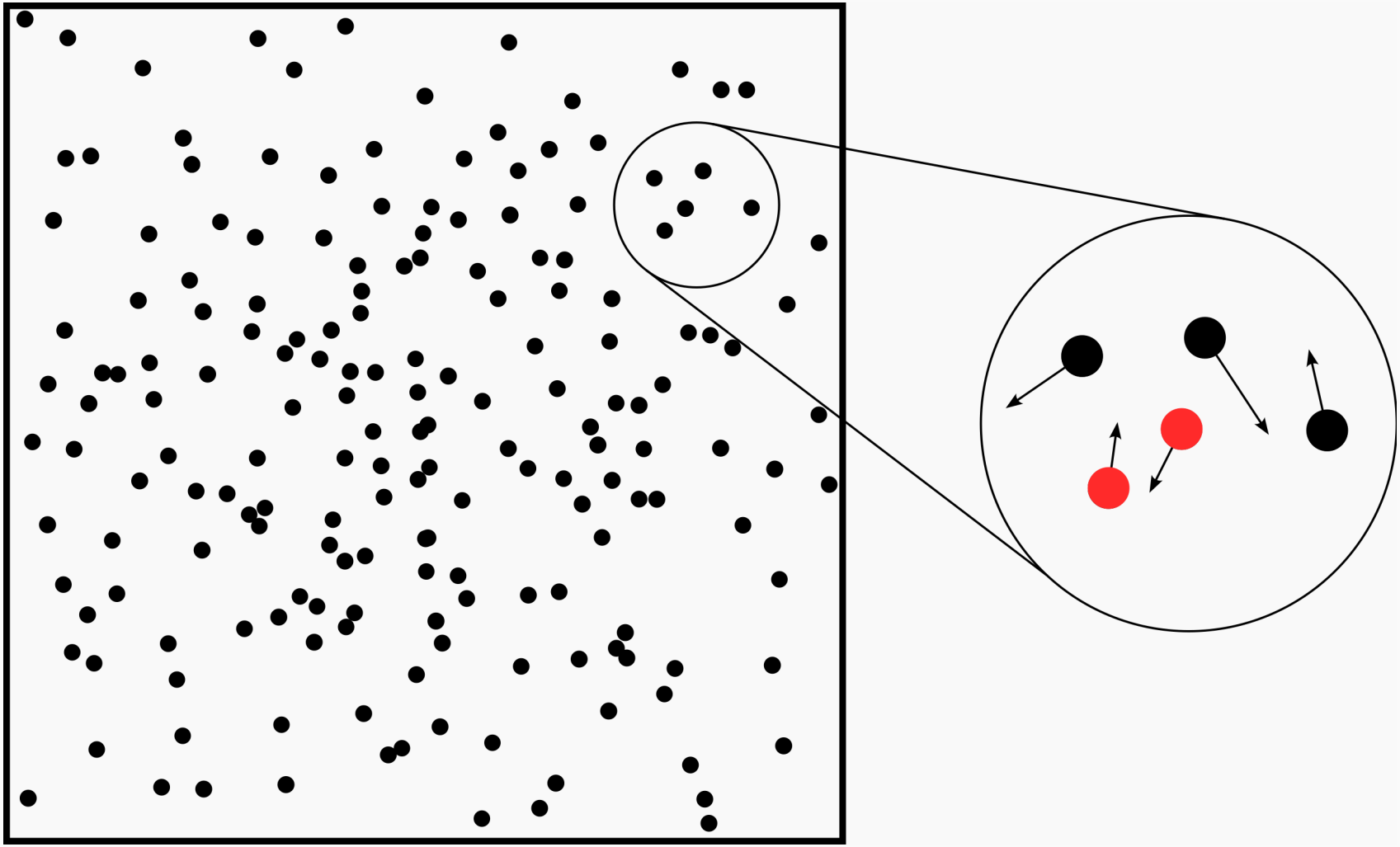
**Remark.** Starting from a general stationary process $X_t$, we may end up with $e_t = g_t(\varepsilon_t, \varepsilon_{t-1}, \dots)$, but we can restrict ourselves to the case

$$X_t = g(\varepsilon_t, \varepsilon_{t-1}, \dots).$$

# Physical Dependence



# Physical Dependence



# Physical Dependence

To measure the degree of dependence among the $X_t$'s, let $(\delta_t)_{t\in\mathbb{Z}}$ be an i.i.d. copy of $(\varepsilon_t)_{t\in\mathbb{Z}}$ and set

$$X_t' = g(\varepsilon_t, \ldots, \varepsilon_1, \delta_0, \varepsilon_{-1}, \ldots).$$

The quantity

$$D_q(\alpha) = \sum_{t=1}^{\infty} t^{\alpha} \|X_t - X_t'\|_q$$

measures the rate at which the process $X$ *forgets* its past.

# Assumptions revisited

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process. Then

1. $\mathbb{E}(X_t) = 0$ and $X_t \in L^q$, for $q > 8$.
2. $X_t$ does not degenerate to a finite order autoregressive process.
3. The spectral density $f_X$ is (uniformly) bounded away from zero.
4. $K_n \in \{1, \ldots, n-1\}$ is a divergent sequence of integers, and there is $\kappa > 0$ such that $K_n^{2+\kappa}/n$ is bounded.
5. $X_t$ is weakly (physically) dependent, with $\alpha \geq 5/2$.

# Contribution

**Theorem**

*Under the above assumptions, any sequence $\widehat{k}_n$ of minimizers of AIC, FPE, Shibata's Criterion (and more) satisfies*

$$\mathbb{P}\left(\left|\frac{Q_n(\widehat{k}_n)}{L_n(k_n^*)} - 1\right| \leq 8(k_n^*)^{-\delta}\right) \geq 1 - C(k_n^*)^{-\gamma},$$

*for some $C, \delta, \gamma > 0$, and $k_n^* \to \infty$.*

# Examples

Examples include:

1. random walks on the regular group,
2. functionals of iterated random systems,
3. functionals of (augmented) GARCH models of any order,
4. functionals of (Banach space valued) linear processes,
5. solutions to many SDEs,
6. (in)finite memory Markov chains, and many more...

# Simulations

The innovations $e_t$ follow a GARCH(0.25,0.25) model with standard Gaussian innovations, that is,

$$e_t = \varepsilon_t L_t,$$

where $\varepsilon_t$ is a sequence of i.i.d. standard Gaussian random variables, and

$$L_t^2 = \frac{1}{10} + \frac{1}{4} L_{t-1}^2 + \frac{1}{4} e_{t-1}^2.$$

The process we consider is given by

$$X_t^{(p)} = e_t + \sum_{j=1}^{\infty} j^{-p} e_{t-j}, \quad p > 1.$$

# Simulations

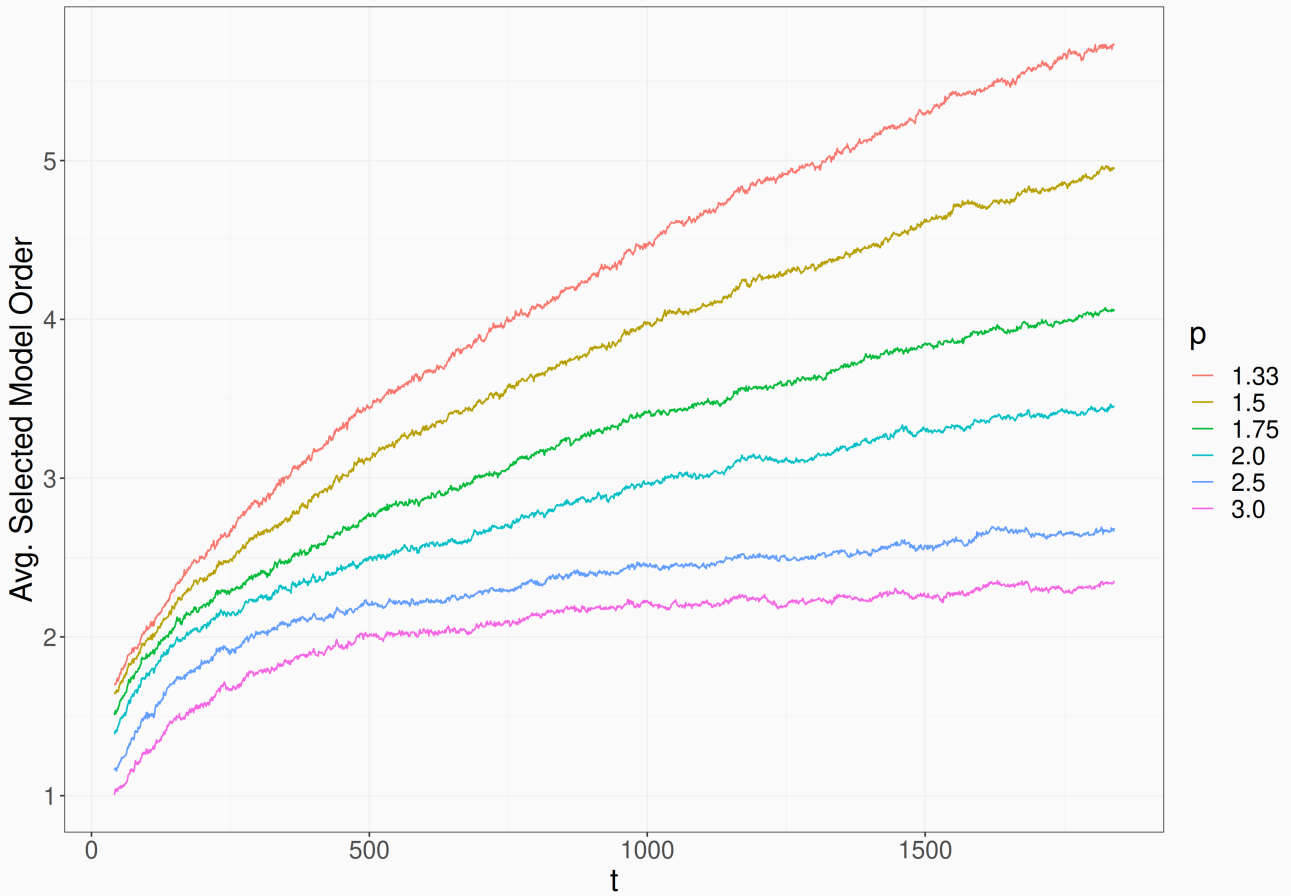



**Figure 1:** Average model order selected by AIC for $(X_s^{(p)})_{s=1}^t$ over 4000 runs, for various values of $p$ and $t = 41, \ldots, 1840$.

# Can we compare the AIC of two models?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

Only $X_1, \ldots, X_n$?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

Only $X_1, \ldots, X_n$? Inflation?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

Only $X_1, \ldots, X_n$? Inflation? GDP growth?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

Only $X_1, \ldots, X_n$? Inflation? GDP growth? Interest rates?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

Only $X_1, \ldots, X_n$? Inflation? GDP growth? Interest rates? Or a combination of all of them?

# Motivation: Compare AIC of two models?

**Goal:** Predict return $X_{n+1}$ of U.S. Treasury bonds $X_1, \ldots, X_n$.

**Problem:** What do we use to predict $X_{n+1}$?

Only $X_1, \ldots, X_n$? Inflation? GDP growth? Interest rates? Or a combination of all of them?

**Question:** Can we use AIC to decide this question?

# Motivation: Compare AIC of two models?

**Goal:** Predict next data point $X_{n+1}$ from a time series $(X_t)_{t\in\mathbb{Z}}$.

**Data:** In addition to $X_1, \ldots, X_n$, we are given two (or more) time series $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_n$, which we can use to predict $X_{n+1}$.

**Question:** Should we use $Y_1, \ldots, Y_n$ or $Z_1, \ldots, Z_n$ to predict $X_{n+1}$? Which of the two approximations

$$X_t \approx \sum_{j=1}^{k_Y} a_j(k_Y) Y_{t-j} \quad \text{or} \quad X_t \approx \sum_{j=1}^{k_Z} b_j(k_Z) Z_{t-j}$$

is *better*?

# Motivation: Compare AIC of two models?

**Part 1:** Use AIC to select model order.

**Part 2:** Use AIC to select the model & model order.

# General strategy

We are given $X_1, \ldots, X_n$, and $Y_1^{(m)}, \ldots, Y_n^{(m)}$, for $m = 1, \ldots, M_n$, and we want to predict $X_{n+1}$.

1. Predict $X_{n+1}$ using the $Y^{(m)}$'s, for every $m = 1, \ldots, M_n$.
2. Compare all the AIC scores of those models.
3. Pick the $\widehat{m}_n \in \{1, \ldots, M_n\}$ and $\widehat{k}_n \in \{1, \ldots, K_n\}$ with the smallest AIC score.
4. Predict $X_{n+1}$ via

$$X_{n+1} \approx \sum_{j=1}^{\widehat{k}_n} a^{(\widehat{m}_n)}(\widehat{k}_n) Y_{n+1-j}^{(\widehat{m}_n)}.$$

# Setup

We are given (jointly stationary) random samples $Y_1^{(m)}, \ldots, Y_n^{(m)}$ for $m = 1, \ldots, M_n$ in addition to $X_1, \ldots, X_n$, and are interested in predicting $X_{n+1}$.

For every $m = 1, \ldots, M_n$, and $t \in \mathbb{Z}$ let $V^{(m)} = \overline{\langle Y_s^{(m)} \mid s < t \rangle}$.

$$P_{V^{(m)}}(X_t) = \sum_{j=1}^{\infty} a_j^{(m)} Y_{t-j}^{(m)}.$$

The vector $a^{(m)} = (a_j^{(m)})_{j \geq 1}$ gives the best linear model for $X_t$ based on $Y_t^{(m)}$.

# Setup

We are given (jointly stationary) random samples $Y_1^{(m)}, \ldots, Y_n^{(m)}$ for $m = 1, \ldots, M_n$ in addition to $X_1, \ldots, X_n$, and are interested in predicting $X_{n+1}$.

For every $m = 1, \ldots, M_n$, and $t \in \mathbb{Z}$ let $V^{(m)} = \overline{\langle Y_s^{(m)} \mid s < t \rangle}$.

$$P_{V^{(m)}}(X_t) = \sum_{j=1}^{\infty} a_j^{(m)} Y_{t-j}^{(m)}.$$

The vector $a^{(m)} = (a_j^{(m)})_{j \geq 1}$ gives the best linear model for $X_t$ based on $Y_t^{(m)}$.

**Idea:** Approximate $a^{(m)} = (a_j^{(m)})_{j \geq 1}$ and choose the model with the smallest prediction error (in $L^2$).

# Setup

For $m = 1, \ldots, M_n$, $k = 1, \ldots, K_n$, and $t \in \mathbb{Z}$, set $V_k^{(m)} = \langle Y_{t-j}^{(m)} \mid j = 1, \ldots, k \rangle \subseteq L^2$.

$$P_{V_k^{(m)}}(X_t) = \sum_{j=1}^{k} a_j^{(m)}(k) Y_{t-j}^{(m)},$$

where $a^{(m)}(k)$ is the best $k$-dimensional model for $X_t$ given $Y_{t-1}^{(m)}, \ldots, Y_{t-k}^{(m)}$.

# Setup

For $m = 1, \ldots, M_n$, $k = 1, \ldots, K_n$, and $t \in \mathbb{Z}$, set $V_k^{(m)} = \langle Y_{t-j}^{(m)} \mid j = 1, \ldots, k \rangle \subseteq L^2$.

$$P_{V_k^{(m)}}(X_t) = \sum_{j=1}^{k} a_j^{(m)}(k) Y_{t-j}^{(m)},$$

where $a^{(m)}(k)$ is the best $k$-dimensional model for $X_t$ given $Y_{t-1}^{(m)}, \ldots, Y_{t-k}^{(m)}$.

**Idea:** Approximate $a^{(m)}$ by $a^{(m)}(k)$ and estimate $a^{(m)}(k)$ from the data.

# Setup

We can use Yule-Walker theory for regression to compute $a^{(m)}(k)$:

$$a^{(m)}(k) = R^{(m)}(k)^{-1} r^{(m)}(k),$$

where

$$R^{(m)}(k) = \left(\mathbb{E}\left(Y_i^{(m)} Y_j^{(m)}\right)\right)_{i,j=1}^{k},$$
$$r^{(m)}(k) = r^{(m)}(k) = \left(\mathbb{E}\left(X_k Y_{k-1}^{(m)}\right), \ldots, \mathbb{E}\left(X_k Y_1^{(m)}\right)\right)^T.$$

# Setup

For $k = 1, \ldots, K_n$, we estimate $R^{(m)}(k)$ and $r^{(m)}(k)$ via $\widehat{R}^{(m)}(k) = (\widehat{R}^{(m)}_{ij})_{i,j=1}^{k}$, and $\widehat{r}^{(m)}(k) = (\widehat{r}^{(m)}_{j})_{j=1}^{k}$, where

$$\widehat{R}^{(m)}_{ij} = \frac{1}{n - K_n} \sum_{t=K_n+1}^{n} Y^{(m)}_{t-i} Y^{(m)}_{t-j}, \quad i, j = 1, \ldots, K_n$$

and

$$\widehat{r}^{(m)}_{j} = \frac{1}{n - K_n} \sum_{t=K_n+1}^{n} X_t Y^{(m)}_{t-j}, \quad j = 1, \ldots, K_n.$$

# Setup: Estimate

Estimate $a^{(m)}(k)$ via

$$\widehat{a}^{(m)}(k) = \widehat{R}^{(m)}(k)^{-1}\widehat{r}^{(m)}(k).$$

We set

$$\begin{aligned} Q^{(m)}(k) &= \|a^{(m)} - \widehat{a}^{(m)}(k)\|_{R,m}^2 \\ &= \|a^{(m)} - a^{(m)}(k)\|_{R,m}^2 + \|a^{(m)}(k) - \widehat{a}^{(m)}(k)\|_{R,m}^2. \end{aligned}$$

where

$$\|z\|_{R,m}^2 = \sum_{i,j=1}^{\infty} z_i z_j R_{ij}^{(m)}.$$

# Setup: Oracle

We define

$$L_n^{(m)}(k) = \|a^{(m)} - a^{(m)}(k)\|_{R,m}^2 + \frac{k}{n - K_n}\sigma_m^2,$$

where

$$\sigma_m^2 = \mathbb{E}\left(X_t - \sum_{j=1}^{\infty} a_j^{(m)} Y_{t-j}^{(m)}\right)^2.$$

The oracle model $m_n^*$ and the oracle model order $k_n^*$ are given by

$$(m_n^*, k_n^*) \in \underset{\substack{1 \le k \le K_n \\ 1 \le m \le M_n}}{\operatorname{argmin}} L_n^{(m)}(k).$$

# Model selection

Define

$$\widehat{\sigma}_m^2(k) = \frac{1}{n - K_n} \sum_{t=K_n+1}^{n} \left( X_t - \sum_{j=1}^{k} \widehat{a}_j^{(m)}(k) Y_{t-j}^{(m)} \right)^2,$$

and set

$$\mathrm{AIC}(m, k) = n \log(\widehat{\sigma}_m^2(k)) + 2k,$$
$$\mathrm{FPE}(m, k) = \frac{n + k}{n - k} \widehat{\sigma}_m^2(k).$$

Our estimators of $m_n^*$ and $k_n^*$ are given by

$$(\widehat{m}_n, \widehat{k}_n) \in \underset{\substack{1 \le k \le K_n \\ 1 \le m \le M_n}}{\operatorname{argmin}} \mathrm{AIC}(m, k).$$

# Model selection

Define

$$\widehat{\sigma}^2_m(k) = \frac{1}{n - K_n} \sum_{t=K_n+1}^{n} \left( X_t - \sum_{j=1}^{k} \widehat{a}_j^{(m)}(k) Y_{t-j}^{(m)} \right)^2,$$

and set

$$\mathrm{AIC}(m,k) = n \log(\widehat{\sigma}^2_m(k)) + 2k,$$
$$\mathrm{FPE}(m,k) = \frac{n+k}{n-k} \widehat{\sigma}^2_m(k).$$

Our estimators of $m_n^*$ and $k_n^*$ are given by

$$(\widehat{m}_n, \widehat{k}_n) \in \underset{\substack{1 \le k \le K_n \\ 1 \le m \le M_n}}{\operatorname{argmin}} \mathrm{AIC}(m,k).$$

**Question:** Does this work?

# Assumptions

Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be an i.i.d. sequence, and assume that there are functions $g$, and $g^{(m)}$ such that

$$X_t = g(\varepsilon_t, \varepsilon_{t-1}, \dots), \quad \text{and}$$
$$Y_t^{(m)} = g^{(m)}(\varepsilon_t, \varepsilon_{t-1}, \dots),$$

for $m \geq 1$ and $t \in \mathbb{Z}$. For $\alpha \geq 0$ we define

$$D_q^X(\alpha) = \sum_{t=1}^{\infty} t^\alpha \|X_t - X_t'\|_q,$$
$$D_q^Y(\alpha) = \sup_{m\geq 1}\left\{\|Y^{(m)}\|_q + \sum_{t=1}^{\infty} t^\alpha \|Y_t^{(m)} - (Y_t^{(m)})'\|_q\right\}.$$

Recall: $(\delta_t)$ i.i.d. copy of $(\varepsilon_t)$, $X_t' = g(\varepsilon_t, \dots, \varepsilon_1, \delta_0, \varepsilon_{-1}, \dots)$ and $(Y_t^{(m)})' = g^{(m)}(\varepsilon_t, \dots, \varepsilon_1, \delta_0, \varepsilon_{-1}, \dots)$.

# Assumptions

Let $(X_t, Y_t^{(m)})_{t\in\mathbb{Z}}$ be a centered, stationary process for all $m \geq 1$. For $q > 8$ and $\alpha \geq 5/2$ we assume:

1. $X_t, Y_t^{(m)} \in L^q$ for all $m \geq 1$.
2. $X_t, Y_t^{(m)}$ are jointly phys. dep. with $D_q^X(\alpha), D_q^Y(\alpha) < \infty$.
3. There is a $c_s > 0$ such that the spectral densities $f^{(m)}$ of $Y_t^{(m)}$ satisfy $f^{(m)} \geq c_s$ for all $m \geq 1$.
4. The sequence $a^{(m)}$ is not eventually zero for any $m \geq 1$.
5. The sequences $K_n \in \{1, \ldots, n-1\}$ and $M_n \in \mathbb{N}$ are divergent, and there is $\kappa > 0$ such that $K_n^{2+\kappa}/n$ is bounded, $M_n/K_n^{2\alpha} \to 0$, and $K_n M_n/n \to 0$.
6. There is a $\psi > 0$, and a $n_0 \geq 1$, such that for all $n \geq n_0$

$$\sigma^2_{m_n^*} \leq \inf_{\substack{1\leq m\leq M_n \\ m\neq m_n^*}} \sigma_m^2 - n^{-1/2}\log^{1/2+\psi}(n).$$

# Contribution

**Theorem**

*Given the previous assumptions, any sequence of minizers* $(\widehat{m}_n, \widehat{k}_n)$ *of AIC, FPE, Shibata's Criterion (and more) satisfies*

$$\mathbb{P}\left(\left|\frac{Q^{(\widehat{m}_n)}(\widehat{k}_n)}{L_n^{(m_n^*)}(k_n^*)} - 1\right| \leq 8(k_n^*)^{-\delta}\right) \geq 1 - C(k_n^*)^{-\gamma},$$

*for some* $C, \delta, \gamma > 0$.

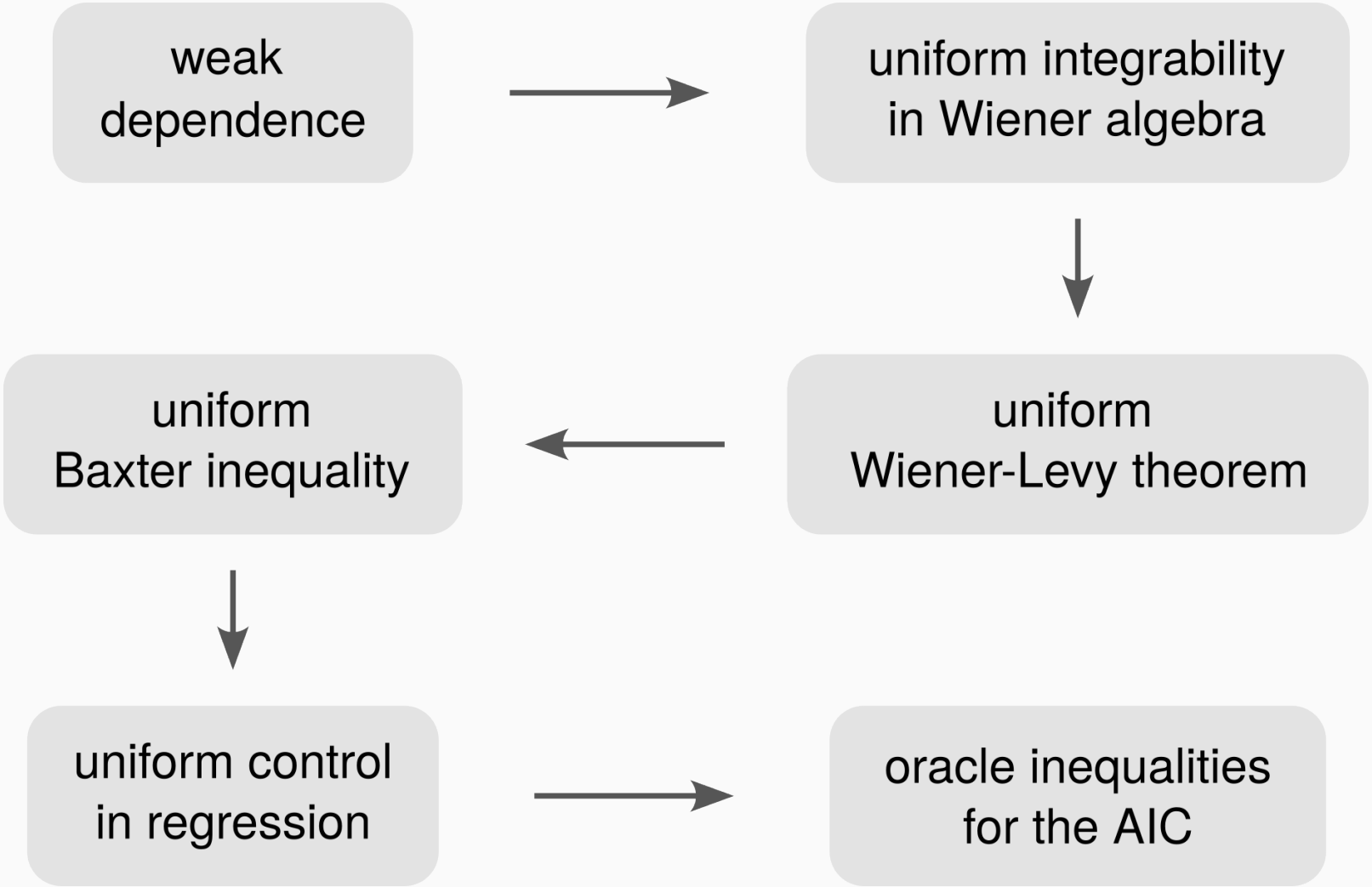
# Questions?


- H. Akaike, *Fitting Autoregressive Models for Prediction.* Ann. Inst. Stat. Math. 21, 243–247 (1969).
- H. Akaike, *Statistical predictor identification.* Ann. Inst. Stat. Math. 22, 203–217 (1970).
- H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle.* Proceeding of the Second International Symposium on Information Theory, 267-281 (1973).
- H. Akaike, *A New Look at the Statistical Model Identification*, IEEE Trans. Automat. Control AC19-6, 716-723 (1974).
- G. Schwarz, *Estimating the Dimension of a Model.* Ann. Statist., 6(2), 461-464 (1978).

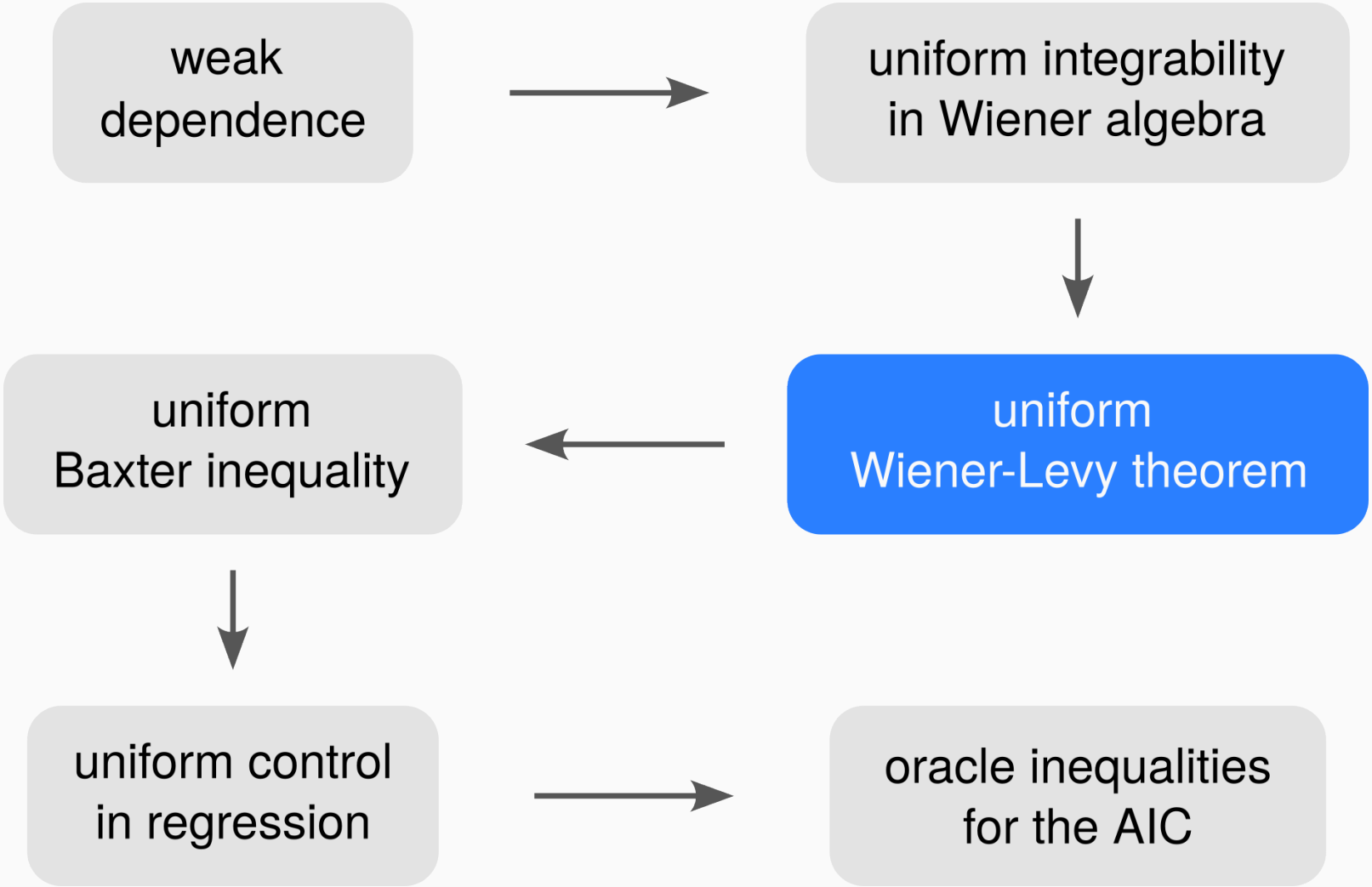
# Questions?


- R. Shibata. *Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process.* Ann. Statist., 8(1), 147–164, 1980.
- S. Lee and A. Karagrigoriou. *An asymptotically optimal selection of the order of a linear process.* Sankhyā Ser. A, (1961- 2002), 63, 01 2001. 73
- C.-K. Ing and C.-Z. Wei. *Order selection for same-realization predictions in autoregressive processes*. Ann. Statist., 33(5), 2423–2474, 2005.
- C.-K. Ing. *Accumulated Prediction Errors, Information Criteria and Optimal Forecasting for Autoregressive Time Series*. Ann. Statist., 35(3), 1238–1277, 2007.
- A. Barron, L. Birgé, and P. Massart. *Risk bounds for model selection via penalization* Probab. Theory Related Fields, 113(3):301–413, 1999.
- E. J. Candès. *Modern statistical estimation via oracle inequalities.* Acta Numer., 15:257–325, 2006

# The argument



# The argument



# Wiener-Levy Theorem

**Theorem**
*Let* $f(t) = \sum_{h\in\mathbb{Z}} f_h e^{iht} : [0, 2\pi] \to \mathbb{C}$ *such that* $\sum_{h\in\mathbb{Z}} |f_h| < \infty$, *and* $H$ *is an analytic (not necessarily single-valued) function which is regular at every point of* $\mathrm{im}(f)$, *then* $H \circ f$ *has an absolutely convergent Fourier series.*

# Uniform integrability in the Wiener algebra

Let $\lambda \geq 0$, and set

$$\mathcal{W}_\lambda = \left\{ f(t) = \sum_{h\in\mathbb{Z}} f_h e^{iht} \mid \sum_{h\in\mathbb{Z}} |h|^\lambda |f_h| < \infty \right\}.$$

A set $F \subseteq \mathcal{W}_\lambda$ is called *uniformly integrable* (in $\mathcal{W}_\lambda$), if

$$\lim_{K\to\infty} \sup_{f\in F} \sum_{|k|>K} |k|^\lambda |f_k| = 0.$$

**Question:** Do analytic functions preserve uniform integrability?

# Uniform Wiener-Levy theorem

**Theorem**
*Let $\lambda \geq 0$, and $F \subseteq \mathcal{W}_\lambda$ be uniformly integrable. If $H$ is an analytic (not necessarily single-valued) function which is regular at every point of $\bigcup_{f \in F} \operatorname{im}(f)$, then*

$$H(F) = \{H \circ f \mid f \in F\}$$

*is uniformly integrable in $\mathcal{W}_\lambda$.*

# Uniform Baxter inequalities

**Theorem**

*Let $(X_t, Y_t^{(m)})_{t\in\mathbb{Z}}$, be centered, jointly stationary processes for $m \in \mathcal{M}$. If the family of spectral densities $f^{(m)}$ of the $Y_t^{(m)}$'s is uniformly integrable in $\mathcal{W}_0$, and*

$$\inf_{\substack{m\in\mathcal{M}\\ t\in[0,2\pi]}} f^{(m)}(t) > 0,$$

*then there is a constant $C > 0$ and a $k_0 > 0$, such that for all $k \geq k_0$, $m \in \mathcal{M}$, and all non-decreasing functions $g : \mathbb{N} \to (0, \infty)$,*

$$\sum_{j=1}^{k} g(j)|a_j^{(m)}(k) - a_j^{(m)}| \leq Cg(k) \sum_{j=k+1}^{\infty} |a_j^{(m)}|,$$

*where $a^{(m)}$ and $a^{(m)}(k)$ are the coefficients of the best $\infty/k$-dimensional linear model for $X_t$ based on $Y_t^{(m)}$.*