

Seminar

1.1 Motivation

Ziel des Seminars ist es die Theorie des Ahnenlinienprozesses und des Kingman Koaleszenten von einem klassischen Wright Fischer Modell zu erweitern, in dem auch Dormanzen in Betracht gezogen werden. Die Ausarbeitung beruht auf der wissenschaftlichen Publikation *Coalescent theory of seed Bank models* [2]. Das Hauptresultat zeigt, dass der Ahnenlinienprozess von einem m-Schritt Seedbank Modell Zeitbeschleunigt gegen einen langsameren Kingman-Koaleszenten konvergiert.

1.2 Klassische Kingman Koaleszent

Bevor wir mit Verallgemeinerungen vom Wright Fischer Modell beginnen, führen wir zuerst die klassische Theorie des Kingman Koaleszenten ein. Anstatt, wie in der klassischen Theorie die viel allgemeineren Canningsmodelle zu betrachten, beschränken wir uns im folgenden Abschnitt nur auf das Wright Fischer Modell. Wir beginnen mit der Notation, welche wir von [1] übernehmen.

Definition 1 (Ahnenprozess des Wright Fischer Modells). Für $N \in \mathbb{N}$ sei

$$A_{r,i}^N, r \in \mathbb{N}, i \in [N]$$

eine Familie von unabhängigen $[N]$ -wertigen gleichverteilten Zufallsvariablen auf $(\Omega, \mathcal{F}, \mathbb{P})$. Für jedes $r \in \mathbb{N}$ heißt

$$A_r^N := (A_{r,1}^N, \dots, A_{r,N}^N) \in [N]^N$$

die Ahnenkonfiguration in Generation r . Der resultierende $[N]^N$ -wertige stochastische Prozess

$$\{A_r^N\}_{r \in \mathbb{N}}$$

heißt Ahnenprozess des Wright-Fischer-Modells.

Wir interpretieren die Definition als: jedes Individuum in der derzeitigen Generation sucht sich gleichverteilt ein Vorfahren aus der vorherigen Generation aus und erbt die Gene davon (wir nehmen an, die Individuen sind Haploid).

Definition 2 (Ahnenlinienprozess einer Stichprobe). Sei A_r^N der Ahnenprozess eines Wright-Fischer-Modells mit Generationen $r \in \mathbb{Z}$. Der Ahnenlinienprozess $\{\Pi_k^{n,N}\}_{k \geq 0}$ einer Stichprobe der Größe $n \in [N]$, startend in Generation $k = 0$, ist eine $P^{(n)}$ -wertige Markovkette, wobei für jedes $k \geq 0$ die Blöcke von $\Pi_k^{n,N}$ durch die Äquivalenzrelation

$$i \sim_k j : \iff A_{0,i}^{(N,\mu)}[k] = A_{0,j}^{(N,\mu)}[k]$$

gegeben sind. Damit sind die Übergänge des Ahnlinienprozesses der Stichprobe (Vorwärts in der Zeit) durch die Dynamik des Ahnenprozesses rückwärts in der Zeit gegeben.

add smth about why next def and theorem is relevant

Definition 3. (Kingman, 1982) Sei $n \in \mathbb{N}$. Der Kingman n -Koaleszent ist die zeitstetige $\mathcal{P}^{(n)}$ -wertige Markov-Kette $\{\Pi_t^n, t \geq 0\}$ mit Q -Matrix gegeben durch

$$q_{\xi,\eta} = \begin{cases} 1 & \text{falls } \xi \prec \eta \\ -\binom{|\xi|}{2} & \text{falls } \xi = \eta \\ 0 & \text{sonst} \end{cases} \quad (1.1)$$

Das Haupttheorem besagt

Satz 1.1. Sei $\{\Pi_k^{(N,\nu)}\}_{k \geq 0}$ der Ahnenlinienprozess einer Stichprobe der Größe n in einem Wright-Fischer-Modell mit Populationsgröße N . Dann gilt für $N \rightarrow \infty$

$$\{\Pi_{\lfloor tN \rfloor}^{n,(N,\nu)}\}_{t \geq 0} \rightarrow \{\Pi_t^n\}_{t \geq 0}$$

im Sinne der endlich-dimensionalen Verteilungen (und in Verteilung auf dem Raum der càdlàg-Pfade auf $P^{(n)}$)

1.3 m-Schritt Seed-Bank Modell

Nun haben wir die Werkzeuge um das Wright Fischer Modell zu verallgemeinern. Im Vergleich zum klassischen Wright Fischer Modell, wo sich jedes Individuum nur aus der vorherigen Generation ein Vorfahren aussucht, kann im m-Schritt Seed-Bank Modell ein Vorfahren aus einem der letzten m Generationen ausgesucht werden. Genauer gesagt, ein Individuum der nächsten Generation sucht sich zuerst mit Wahrscheinlichkeit $b_i, i \in \{1..m\}$ die Generation aus **irgendwie vorherige noch spezifizieren**, von welcher er den Vorfahren aussucht, und anschließend wählt er sich aus der ausgewählten Generation ein Individuum gleichverteilt aus. Wir beginnen mit der formalen Definition.

Definition 4 (Ahnenprozess des m-Schritt Seed-Bank Modells). Für $N \in \mathbb{N}$ sei

$$A_{r,i}^N, r \in \mathbb{N}, i \in [N]$$

eine Familie von unabhängigen $E_{N,m}$ -wertigen Zufallsvariablen auf $(\Omega, \mathcal{F}, \mathbb{P})$. Hier bezeichnet

$$E_{N,m} := [N] \times [m] = \{(i, j) | i \in \{1, 2, \dots, N\}, j \in \{1, 2, \dots, m\}\}$$

die Menge der Tupeln wo die erste Komponente das Wright-Fischer Modell Analogon ist, und die zweite Komponente der Zahl der Generation in der Vergangenheit, aus der geerbt wurde, entspricht. Wir fordern zusätzlich zu der Wright-Fischer Annahme, dass die erste Komponente unabhängig und gleichverteilt auf $[N]$ ist, dass auch die zweite Komponente unabhängig, aber verteilt ist zu der Verteilung $\mathbb{P}(X = i) = b_i$ *fix*. Für jedes $r \in \mathbb{N}$ heißt

$$A_r^N := (A_{r,1}^N, \dots, A_{r,N}^N) \in (E_{N,m})^N$$

die Ahnenkonfiguration in Generation r . Der resultierende $(E_{N,m})^N$ -wertige stochastische Prozess

$$\{A_r^N\}_{r \in \mathbb{N}}$$

heißt Ahnenprozess des m -Schritt Seed-Bank Modells.

Diesmal beruht das Kingman-ähnliche Resultat nicht auf dem Ahnenlinienprozess selber, sondern betrachtet die Anzahl Ahnenlinien, welche durch den Ahnenprozess gegeben sind. Dafür beschränken wir die Stichprobe nicht nur auf eine Generation, sondern betrachten sogenannte m -Fenster, und definieren:

Definition 5 (m -Fenster). Mit einem m -Fenster bezeichnen wir die geordnete Sammlung von m Generationen $(r, r+1, \dots, r+(m-1))$, $r, m \in \mathbb{N}$

Definition 6 (Ahnenlinienprozess einer Stichprobe). Sei A_r^N der Ahnenprozess eines m -Schritt Seed-Bank Modells mit Generationen $r \in \mathbb{Z}$. Sei $S_n := \{(x_i, y_i) : i \in \{1, \dots, n\}\} \subset [N] \times \{0, \dots, m-1\}$ eine beliebige Stichprobe von Individuen im m -Fenster der Generation 0. Dann definieren wir $X^n(k) := (X_1^n(k), X_2^n(k), \dots, X_m^n(k))$, wobei

$$X_i^n(k) := |\{A_{i,x_i}^N[k] : (x_i, i) \subset S_n\}|$$

wrong definition als den Vektor der Anzahl von übrig gebliebenen disjunkten Individuen im m -Fenster von vor k -Generationen und nennen X_i^n den Konfigurationsprozess.

Nun wollen wir unser m -Fenster in jeder Generation neu verschieben und die Anzahl der Übrig gebliebener Individuen aufzählen. Dafür definieren wir:

Definition 7. Für einen Konfigurationsprozess $X_i^n(k)$ aus 6, definieren wir

$$|\Pi_k^{n,N}| := \sum_{i=1}^m X_i^n(k)$$

als Gesamtanzahl von disjunkten Vorfahren einer n -Stichprobe von vor k Generationen

Beginnend mit einer Stichprobe der Größe n im ersten m -Fenster beruht die Konvergenz auf der Zufallsvariable $|\Pi_k^{n,N}|$. Dafür führen wir zuerst den langsameren Kingman-Koaleszenten ein.

add smth about why next def and theorem is relevant

Definition 8. Sei $n \in \mathbb{N}$. Der m -Schritt Kingman n -Koaleszent ist die zeitstetige $[n]$ -wertige Markov-Kette $\{|\Pi_t^n|, t \geq 0\}$ mit Q -Matrix gegeben durch

$$q_{i,j} = \begin{cases} \beta_1^2 \binom{i}{2} & \text{falls } j = i - 1 \\ -\beta_1 \binom{i}{2} & \text{falls } i = j \\ 0 & \text{sonst} \end{cases} \quad (1.2)$$

wobei $\beta_1 := 1/\mathbb{E}[B] = 1/(\sum_{i=1}^m ib_i)$ *define beta*

Satz 1.2 (Haupttheorem). Sei $|\Pi_t^n|$ der m -Schritt Kingman Koaleszent und $|\Pi_k^{n,N}|$ wie in 7. Dann gilt für $N \rightarrow \infty$

$$|\Pi_{[Nt]}^n| \rightarrow |\Pi_t^n| \quad (1.3)$$

in Verteilung auf dem Raum der càdlàg-Pfade auf $[n]$

add preamble about showing c2n and c3n for classic kingman coalescent has same necessity as here Wir berechnen nun die Wahrscheinlichkeiten, dass von einer Generation auf die nächste, keine, eine oder mehrere Verschmelzungen vorkommen. Die Wahrscheinlichkeit, dass in einem Schritt eine Verschmelzung Zelle define vorkommt, gegeben r Bälle landen im nächsten Schritt in dieser Zelle und es gibt l alte Bälle, ist gegeben durch:

$$\begin{aligned} \mathbb{P}(\text{keine Verschmelzung}) &= (1 - \frac{l}{N})(1 - \frac{l+1}{N}) \dots (1 - \frac{l+r-1}{N}) \\ &= (1 - \frac{\sum_{i=l}^{r+l-1} i}{N}) + \mathcal{O}(\frac{1}{N^2}) \\ &= 1 - \frac{1}{N}(lr + \binom{r}{2}) + \mathcal{O}(\frac{1}{N^2}) \end{aligned}$$

Analog folgt

$$\begin{aligned} \mathbb{P}(\text{Genau eine Verschmelzung}) &= (1 - \frac{l}{N})(1 - \frac{l+1}{N}) \dots (1 - \frac{l+r-2}{N})(\frac{l}{N}) \\ &\quad + (1 - \frac{l}{N})(1 - \frac{l+1}{N}) \dots (1 - \frac{l+r-2}{N})(\frac{l+1}{N}) \\ &\quad + \dots + (1 - \frac{l}{N})(1 - \frac{l+1}{N}) \dots (1 - \frac{l+r-2}{N})(\frac{l+r-1}{N}) \\ &= (1 - \frac{l}{N})(1 - \frac{l+1}{N}) \dots (1 - \frac{l+r-2}{N})(\frac{\sum_{i=l}^{r+l-1} i}{N}) \\ &= \frac{1}{N}(lr + \binom{r}{2}) + \mathcal{O}(\frac{1}{N^2}) \end{aligned} \quad (1.4)$$

Und somit

$$\mathbb{P}(\geq 2 \text{ Verschmelzungen}) = \mathcal{O}(\frac{1}{N^2})$$

Diese Bedingung ähnelt der Bedingung im Wright Fischer Modell, wobei $< \dots$ add later durch ... ersetzt haben.

Nun betrachten wir diese Wahrscheinlichkeiten zusammen mit der Bedingung auf die multinomiale Verteilung der Bälle. Dafür definieren wir $R(k+1) := (R_1(k+1), R_2(k+1), \dots, R_m(k+1))$, als die Anzahl verlegten Individuen die in der ersten, zweiten ... m -ten Generation nach der $-(k+1)$ -ten Generation landen. Nach Konstruktion vom Prozess, gilt:

$$R(k+1) \sim Mult(X_1(k); b_1, \dots, b_m)$$

Gegeben $X(k)$ und $R(k+1)$, gilt:

$$X_i(k+1) = \begin{cases} X_{i+1}(k) + R_i(k+1), & \text{falls keine Verschmelzung} \\ X_{i+1}(k) + R_i(k+1) - 1, & \text{falls eine Verschmelzung,} \end{cases} \quad (1.5)$$

wobei der zweite Fall wegen 1.4 mit Wahrscheinlichkeit

$$\frac{1}{N} \left(X_{i+1}(k) R_i(k+1) + \binom{R_i(k+1)}{2} \right) + \mathcal{O} \left(\frac{1}{N^2} \right)$$

passiert (Die Fälle mit Wahrscheinlichkeit $\mathcal{O} \left(\frac{1}{N^2} \right)$ wurden ignoriert). Falls wir nun σ als ein Verschiebungsoperator

$$\sigma(X_1(k), \dots, X_m(k)) := (X_2(k), \dots, X_m(k), 0)$$

und

$$a_i(k) := X_{i+1}(k) R_i(k+1) + \binom{R_i(k+1)}{2}$$

definieren so bekommen wir insgesamt für die Übergänge von $X^n(k)$

$$X^n(k+1) = \begin{cases} \sigma(X^n(k)) + R(k+1), & \text{mit W-keit } 1 - \frac{1}{N} \sum_{i=1}^m a_i(k) + \mathcal{O} \left(\frac{1}{N^2} \right) \\ \sigma(X^n(k)) + R(k+1) - e_i, & \text{mit W-keit } \frac{1}{N} a_i(k) + \mathcal{O} \left(\frac{1}{N^2} \right) \\ \text{sonstige Konfiguration,} & \text{mit W-keit } \mathcal{O} \left(\frac{1}{N^2} \right) \end{cases} \quad (1.6)$$

Falls wir also die Koaleszenz-Wahrscheinlichkeit gegeben einer Konfiguration berechnen wollen, so bekommen wir

$$\mathbb{P}(\text{genau eine Verschmelzung} | X^n(k)) \quad (1.7)$$

$$= \mathbb{E}[\mathbb{P}(\text{genau eine Verschmelzung} | X^n(k), R(k+1)) | X^n(k)] \quad (1.8)$$

$$= \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^{m-1} X_{i+1}(k) R_i(k+1) + \frac{1}{N} \sum_{i=1}^m \binom{R_i(k+1)}{2} + \mathcal{O} \left(\frac{1}{N^2} \right) | X^n(k) \right] \quad (1.9)$$

$$= \frac{1}{N} X_1(k) \sum_{i=1}^{m-1} X_{i+1}(k) b_i + \frac{1}{N} \binom{X_1(k)}{2} \sum_{i=1}^m b_i^2 + \mathcal{O} \left(\frac{1}{N^2} \right) \quad (1.10)$$

$$(1.11)$$

wobei $R_i(k+1) \sim \text{Bin}(X_1(k), b_i)$ ausgenutzt wurde. Also für eine konkrete Konfiguration bekommen wir:

$$\mathbb{P}(\text{genau eine Verschmelzung im nächsten Schritt} | X^n(k) = (x_1, \dots, x_m)) \quad (1.12)$$

$$= \frac{1}{N} \left(x_1 \sum_{i=1}^{m-1} x_{i+1} b_i + \binom{x_1}{2} \sum_{i=1}^m b_i^2 \right) + \mathcal{O} \left(\frac{1}{N^2} \right) \quad (1.13)$$

$$(1.14)$$

Satz 1.3. Seien E, E_1, E_2, \dots metrische Räume mit E lokal kompakt und separabel. Für $n = 1, 2, \dots$ sei $\gamma_n : E_n \rightarrow E$ messbar, sei $\mu_n(x, \Gamma)$ eine Übergangs Funktion auf $E_n \times \mathcal{B}(E_n)$ und $\{Y_n(k), k = 0, 1, 2, \dots\}$ eine Markov Kette in E_n , welche Übergänge durch μ_n gegeben hat. Sei $\epsilon_n > 0$ mit $\lim_{n \rightarrow \infty} \epsilon_n = 0$. Definiere $X_n(t) = \gamma_n(Y_n([t/\epsilon_n]))$,

$$T_n f(x) = \int f(y) \mu_n(x, dy), \quad f \in B(E_n),$$

und $\pi_n : B(E) \rightarrow B(E_n)$ durch

Literaturverzeichnis

- [1] Jochen Blath. Stochastische modelle der populationsgenetik. 2024.
- [2] Ingemar Kaj, Stephen M Krone, and Martin Lascoux. Coalescent theory for seed bank models. *Journal of Applied Probability*, 38(2):285–300, 2001.