Züchtungslehre - Genomische Selektion

Peter von Rohr

2016-12-09

Einleitung

Zuchtprogramme werden betrieben um die fundamentalen Probleme in der Tierzucht zu lösen. Wir hatten in früheren Kapiteln gesehen, dass die **Auswahl der Elterntiere** einer zukünftigen Generation in der Tierzucht als ein fundamentales Problem bezeichnet werden.

Zuchtziel

Die Kriterien nach welchen die Tiere bewertet und anschliessend ausgewählt werden, sind relativ bezüglich zu einer gegebenen Umwelt. Sobald sich eine Organisation auf eine Auswahl von Bewertungskriterien geeinigt hat, werden diese zu einem **Zuchtziel** zusammengefasst. Das Zuchtziel hat meistens die Form einer sprachlichen Beschreibung des idealen oder besten Zuchttieres. Für eine gezielte und systematische Auswahl von Elterntieren braucht es eine quantitative Bewertung von potenziellen Selektionskandidaten. Für eine solche quantitative Bewertung muss das Zuchtziel in eine mathematisch erfassbare Form verwandelt werden.

Gesamtzuchtwert

Die mathematische Formulierung des Zuchtziels wird als **Gesamtzuchtwert** bezeichnet und meist mit H abgekürzt. Die im Zuchtziel enthaltenen Bewertungskriterien müssen im Gesamtzuchtwert auf konkrete Merkmale von Tieren übertragen werden. Der Gesamtzuchtwert H ist definiert als gewichtetes Mittel von wahren Zuchtwerten der berücksichtigten Merkmale. Die Gewichtungsfaktoren entsprechen den wirtschaftlichen Gewichten der einzelnen Merkmalen im Gesamtzuchtwert. Die wirtschaftlichen Gewichte für die Merkmale im Gesamtzuchtwert entsprechen der Veränderung des Gewinns bei einer kleinen Veränderung des Populationsmittels. Durch die Gewichtung der wahren Zuchtwerte mit den wirtschaftlichen Gewichten im Gesamtzuchtwert, ist die Einheit von H eine monetäre Grösse.

Für ein Tier i kann der Gesamtzuchtwert H_i berechnet werden als

$$H_i = \sum_{j=1}^{n} v_j * a_{ij} = v^T * a_i \tag{1}$$

mit
- n Anzahl Merkmale im Gesamtzuchtwert
- v_j Wirtschaftliches Gewicht vom Merkmal j im Gesamtzuchtwert
- a_{ij} wahrer Zuchtwert von Tier i für Merkmal j
- v Vektor der Länge n mit wirtschaftlichen Gewichten aller Merkmale
- a_i Vektor der Länge n mit wahren Zuchtwerten aller Merkmale für Tier i

Werden die Tiere nach ihrem Gesamtzuchtwert H_i beurteilt und rangiert und werden die Eltern der nachfolgenden Generation aufgrund dieser Rangierung ausgewählt, so wird der zu erwartende Selektionserfolg optimiert.

Selektionsindex

In der Praxis sind die wahren Zuchtwerte a_i unbekannt und müssen deshalb aus Daten geschätzt werden. Da die wahren Zuchtwerte unbekannt sind, ist auch der wahre Gesamtzuchtwert unbekannt. Der Gesamtzuchtwert

kann aber geschätzt werden. Die Schätzung des Gesamtzuchtwertes H wird als Selektionsindex I bezeichnet. Die Schätzung $I = \hat{H}$ des Gesamtzuchtwertes H soll so sein, dass die Abweichung zwischen I und H minimal ist, d.h. die Fehlervarianz $var(H - I)$ soll minimiert werden.

Für einen Ansatz von I ist es naheliegend, den Selektionsindex auch als gewichtete Summe anzunehmen. Da wir bei der BLUP-Zuchtwertschätzung gesehen habe, dass diese Voraussagen optimale Eigenschaften haben, werden wir diese anstelle der wahren unbekannten Zuchtwerte einsetzen. Somit ist unser Selektionsindex definiert als

$$\hat{H}_i = I_i = \sum_{j=1}^{m} b_j * \hat{a}_{ij} = b^T * \hat{a}_i \tag{2}$$

mit
- m Anzahl Merkmale im Selektionsindex
- b_j Gewichtung vom Merkmal j im Selektionsindex
- $\hat{a}_{ij}$ geschätzter Zuchtwert von Tier i für Merkmal j
- b Vektor der Länge m mit Gewichtungsfaktoren der Merkmale im Index
- $\hat{a}_i$ Vektor der Länge m mit geschätzten Zuchtwerten aller Merkmale für Tier i

In der Definition des Selektionsindexes (2) sind die Gewichtungen im Vektor b unbekannt. Diese können aufgrund der Anforderung der minimalen Fehlervarianz abgeleitet werden. Das Resultat dieser Ableitung sind die sogenannten Indexgleichungen aus welchen die Indexgewichte b berechnet werden können.

$$Pb = Gv \tag{3}$$

mit
- P die Covarianzmatrix zwischen den Merkmalen im Selektionsindex ist
- G die Covarianzmatrix zwischen den Merkmalen im Gesamtzuchtwert und den Merkmalen im Selektionsindex ist

Aus (3) folgt b als

$$b = P^{-1}Gv \tag{4}$$

Aufgrund der Indexgleichungen (3) wird klar, dass die Merkmale im Gesamtzuchtwert und im Selektionsindex nicht identisch sein müssen. Sind die gleichen Merkmale im Selektionsindex, wie im Gesamtzuchtwert, dann vereinfacht sich die Berechnung der Gewichte in (4). In diesem Fall sind die Matrizen P und G gleich und somit ist $b = v$. Recht häufig finden wir im Gesamtzuchtwert Merkmale, welche nicht einfach und nicht kostengünstig zu erheben sind. Solche Merkmale werden häufig durch Hilfsmerkmale im Selektionsindex ersetzt. Ein Hilfsmerkal im Selektionsindex muss einfach zu erheben sein und sollte eine möglichst hohe Korrelation zum entsprechenden Originalmerkmal im Gesamtzuchtwert aufweisen.

Genomische Selektion

Durch die rasante Entwicklung von Technologien in der Molekularbiologie wurde es möglich eine grosse Anzahl an Selektionskandidaten typisieren zu lassen. Der Begriff **Typisierung** bedeutet die Ermittlung von Genotypen an einer grossen Anzahl an Genorten. Diese Genorte werden auch als **Marker** bezeichnet, da diese wie Vermessungspunkte auf der Landkarte des gesamten Genoms betrachtet werden können. An den ermittelten Genorte werden sogenannte SNP (Single Nucleotide Polymorphisms) beobachtet. Standardmässig werden die Genotypen an rund 50000 (50K) SNP-Markern ermittelt. Wird diese genomische Information an einer genügend grossen Anzahl von Tieren mit ausreichender Diversität erhoben, können gewisse Varianten an bestimmten Genorten mit erwünschten Ausprägungen von phänotypischen Merkmalen assoziiert werden.

Die genomische Selektion kann einen wichtigen Beitrag zum Erfolg von Zuchtprogrammen leisten, vor allem wenn gewisse Merkmale im Zuchtziel schwer zu erheben sind. Genomische Selektion liefert genauere

Zuchtwerte zu einem sehr frühen Zeitpunkt im Leben eines Tieres. Vor allem in der Milchviehzucht, wo wichtige Merkmale nur bei den weiblichen Tieren nach deren Reproduktionsleistung beobachtet werden können. Da hat die genomische Selektion ein grosses Potential den Selektionsfortschritt erheblich zu steigern.

Die Vorhersage von Zuchtwerten aufgrund von genomischer Information setzt eine sogenannte **Referenzpopulation** voraus. Diese muss möglichst gross sein und muss repräsentativ sein für die gesamte Zuchtpopulation. Die Genauigkeit der vorausgesagten genomischen Zuchtwerten ist abhängig von der Grösse der Referenzpopulation und von der Anzahl Verwandtschaftsbeziehungen eines Tieres zu anderen Tieren in der Referenzpopulation. Hat ein Tier i viele Verwandtschaftsbeziehungen zu anderen Tieren in der Referenzpopulation, so werden die vorausgesagten genomischen Zuchtwerte von i sehr genau sein.

Bei der aktuell übliche Dichte an SNP-Markern (50K pro Genom) ist eine zuverlässige Voraussage von genomischen Zuchtwerten nur innerhalb einer Rasse möglich. Genomische Selektion über die Grenzen einer Rasse hinaus braucht noch dichtere Marker-Information und ist vielleicht erst bei der Verwendung der kompletten genomischen Sequenz als Informationsquelle möglich.

Schätzung des Selektionsfortschrittes in einem Zuchtprogramm

In einem Zuchtprogramm werden praktisch immer mehrere Eigenschaften oder Merkmale von Zuchttieren gleichzeitig bearbeitet. Es kann gezeigt werden, dass eine Kombination der Merkmale mit einer individuellen Gewichtung der einzelnen Merkmale, wie dies im Gesamtzuchtwert gemacht wird, zu einem besseren Selektionsfortschritt führt, als das mit anderen Selektionstrategien der Fall wäre. Wie wir einen bestimmten Gesamtzuchtwert durch den Selektionsindex schätzen wird in der Selektionsindextheorie im nächsten Abschnitt besprochen.

Selektionsindextheorie

Potentielle Gewinne durch die Berücksichtigung von genomischer Information bei den Selektionsentscheidungen können durch die Selektionsindextheorie abgeschätzt werden. Allgemein funktioniert die Selektionsindextheorie so, dass wir alle verfügbaren Informationsquellen wie zum Beispiel geschätzte Zuchtwerte oder genomische Informationen in einem Vektor x zusammenfassen. Diese Informationen werden dann mit bestimmten Gewichtungsfaktoren versehen und zu einem gewichteten Mittel zusammengefasst. Dieses gewichtete Mittel dient dann zur Schätzung einer unbekannten Grösse, wie zum Beispiel dem Gesamtzuchtwert.

Als erstes definieren wir die Matrix P als Covarianzmatrix zwischen den Informationsquellen im Vektor x. Es gilt also

$$P = var(x).$$

Soll unser Selektionsindex den Gesamtzuchtwert schätzen, so definieren wir die Matrix G als Covarianz zwischen den Merkmalen im Index und den Merkmalen im Gesamtzuchtwert. Somit haben wir

$$G = cov(x, a^T)$$

wobei a der Vektor der wahren Zuchtwerte der Merkmale im Gesamtzuchtwert darstellt.

Zur Herleitung der Indexgleichungen verwenden wir die Bedingung, dass die Fehlervarianz $var(H - I)$ minimal sein soll. Daraus folgt dann, dass

$$Pb = Gv \tag{5}$$

Lösen wir (5) nach b auf, erhalten wir die unbekannten Indexgewichte. Der erwartete Selektion pro Standardabweichung σ_I des Selektionsindexes beträgt

$$\Delta G = b^T G/\sigma_I$$

wobei die Standardabweichung des Indexes $\sigma_I = \sqrt{b^T P b}$. Die Genauigkeit des Indexes ist definiert als

$$\frac{\sigma_I}{\sigma_H} = \sqrt{\frac{b^T P b}{v^T C v}}$$

wobei $\sigma_H = \sqrt{v^T C v}$ die Standardabweichung des Gesamtzuchtwertes darstellt und die Matrix C die Covarianz der wahren Zuchtwerte beinhaltet. Der Selektionsfortschritt ΔG_i für ein bestimmtes Merkmal i entspricht

$$\Delta G_i = b^T G_i/\sigma_I$$

Berücksichtigung von genomischen Zuchtwerten

Für ein bestimmtes Merkmal m kann genomische Information in Form eines geschätzten genomischen Zuchtwertes q_m als weitere Informationsquelle zum Vektor x hinzugefügt werden. Wir nehmen an, dass der Effekt q_m ein Anteil ρ der genetisch additiven Varianz σ^2_{am} beim Merkmal m erklärt. Der Varianzanteil V_{am} erklärt durch q_m ist $V_{am} = \rho\sigma^2_{am}$. Die Diagonalelemente von P entsprechen somit V_{am} und Offdiagonalelemente von P enthalten die Covarianzen zwischen dem genomischen Zuchtwert q_m und den übrigen Informationsquellen im Vektor x, somit gilt

$$cov(q_m, x) = a_{ij}\rho\sigma^2_{am}$$

wobei a_{ij} dem Verwandschaftsgrad zwischen Individuum j, zu welchem die entsprechende Information im Vektor x gehört und dem Selektionskandidaten i, für welchem der genomische Zuchtwert q_m geschätzt wurde.

Ein Beispiel

Nehmen wir an, dass wir nur ein Merkmal im Selektionsindex und im Gesamtzuchtwert haben. Die Covarianzmatrizen reduzieren sich in diesem Fall zu einfachen Zahlen und betragen $P = \sigma^2_p = 1$ und $G = \sigma^2_a = 0.25$. Der Anteil $\rho = 0.3$, was bedeutet, dass 30% der genetisch additiven Varianz durch den Effekt des genomischen Zuchtwerts erklärt wird.

Ein Selektionsindex, welcher als einzige Informationsquelle eine phänotypische Beobachtung beinhaltet, führt zu einem Indexgewicht von

$$b = \frac{\sigma^2_a}{\sigma^2_p} = 0.25,$$

was der Heritabilität h^2 entspricht. Die Genauigkeit des Indexes wird berechnet als,

$$r_{IH} = \sqrt{\frac{b^T P b}{\sigma^2_a}} = 0.5$$

Berücksichtigen wir nun zusätzlich zur phänotypischen Beobachtung auch noch den genomischen Zuchtwert, so können wir den Selektionsindex erweitern. Die Bestandteile des erweiterten Indexes lauten:

$$P = \begin{bmatrix} 1.000 & 0.075 \\ 0.075 & 0.075 \end{bmatrix}$$

$$Gv = \begin{bmatrix} 0.250 \\ 0.075 \end{bmatrix}$$

Der Vektor der Indexgewichte bestimmen wir mit

$$b = P^{-1}Gv.$$

Setzen wir die Zahlen ein, dann erhalten wir als Resultat

$$b = \begin{bmatrix} 0.189 \\ 0.811 \end{bmatrix}$$

Die Genauigkeit des erweiterten Selektionsindexes ist

$$r_{IH} = 0.658$$

Somit hat die zusätzliche Berücksichtigung des genomischen Zuchtwertes im Selektionsindex eine Erhöhung der Genauigkeit um 31.5% gebracht.

Schätzung genomischer Zuchtwerte

Das Problem bei der Schätzung genomischer Zuchtwerte liegt, darin, dass die Anzahl der SNP-Effekte k viel grösser ist als die Anzahl Beobachtungen n. Obwohl von der Datenstruktur ein einfaches Regressionsmodell ausreichend wäre, können die Effekte nicht mit Least Squares geschätzt werden, da $n << k$.

In der Tierzucht werden die folgenden zwei Möglichkeiten berücksichtigt, um das Problem von $n << k$ zu lösen.

1. BLUP von zufälligen Effekten in einem gemischten linearen Modell (gBLUP)
2. Bayes'sche Ansätze für die Effektschätzung.

Der Ausgangspunkt für beide Ansätze bildet das gemischte lineare Modell

$$y = X\beta + Z\alpha + e \tag{6}$$

mit	y	Vektor der Länge n mit Beobachtungen
	β	Vektor der Länge p mit fixen Effekten
	X	$n \times p$ Inzidenzmatrix der fixen Effekte β
	α	Vektor der Länge k mit zufälligen Alleleeffekten
	Z	$n \times k$ Inzidenzmatrix für α
	e	Vektor der Resteffekte

Genomisches BLUP (gBLUP)

Basierend auf dem linearen gemischten Modell (6) werden in gBLUP die genomischen Zuchtwerte als zufällige Effekte modelliert. Analog zum BLUP-Tiermodell können wir die genomischen Zuchtwerte als Lösungen der zufälligen Effekte vorhersagen. Schon im BLUP-Tiermodell war die Anzahl vorherzusagender Zuchtwerte grösser als die Anzahl Beobachtungen. Tiere, welche keine Beobachtung haben, erhalten ihren Zuchtwert über die Verknüpfungen, welche durch die bekannte Covarianzstruktur über die Verwandtschaft gegeben ist. Somit erzielen wir eine Regularisierung des Problems $n << k$ durch den Einbezug der Covarianz für die Voraussage der zufälligen Effekte.

Im Gegensatz zum Pedigree-basierten Tiermodell, welche die Covarianzen zwischen den Individuen über die Verwandtschaft definiert, ist in gBLUP die Covarianzstruktur der zufälligen Effekte α über die sogenannte **Genomische Verwandtschaftsmatrix** festgelegt. Die Covarianzmatrix für α entspricht also $var(\alpha) = G\sigma_g^2$, wobei G die genomische Verwandtschaftsmatrix ist und σ_g^2 der genetischen Varianz entspricht.

Im Gegensatz zur Pedigree-basierten Verwandtschaftsmatrix A basiert die genomische Verwandtschaftsmatrix G auf den beobachteten SNP-Markerinformationen.

Bayes'sche Ansätze

Bayes'sche Ansätze zur Schätzung von unbekannten Grössen, wie in unserem Falle die genomischen Zuchtwerte, verwenden die A-Posteriori Verteilung der unbekannten Grössen gegeben die beobachteten Daten. Als Schätzungen oder Voraussagen werden häufig die Erwartungswerte der A-Posteriori Verteilungen angegeben.

Für eine bestimmte unbekannte Grösse - für unsere Anwendungen sind das meistens die unbekannten Parameter - wird die A-Posteriori-Verteilung aufgrund des Satzes von Bayes aus der A-Priori-Verteilung des unbekannten Parameters und der Likelihood der Daten gegeben der Parameter errechnet. Für einen unbekannten Parameter θ, welcher wir aus einem Datensatz y schätzen wollen, entspricht die A-Posteriori-Verteilung nach dem Satz von Bayes dem folgenden Ausdruck:

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)}$$

mit	$P(\theta)$	A-Priori-Verteilung des unbekannten Parameters
	$P(y\|\theta)$	Likelihood
	$P(y)$	konstante Normalisierungskonstante

Die Regularisierung des Problems $n << k$ bei den Bayes'schen Ansätzen erfolgt über die Verwendung der A-Priori-Information. Spezifisch in der Vorhersage der genomischen Zuchtwerte können wir über die Wahl der geeigneten A-Priori-Verteilung die vorhergesagten genomischen Zuchtwerte so modellieren, dass nur eine kleine Anzahl Genorte einen Einfluss auf den Zuchtwert haben, oder dass wir eine bestimmte Varianzstruktur zwischen den Genorten vorgeben.

Analog wird auch hier die Regularisierung über die Vorgabe von bestimmten Strukturen erreicht.