

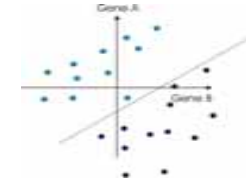
Microarrays



Genomische Datenanalyse
15. Kapitel

Von 1 über 2 nach 30.000

Wir hatten Diagnose mit einem Markergen



... und mit Zweien

Das ist noch keine Genomforschung!

Es sind Details des Stoffwechsels der Patienten

Vogelperspektive

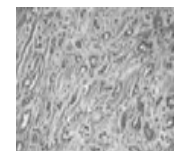
Wünschenswert ist es sich zunächst einen **Überblick über den gesamten Stoffwechsel des Patienten** zu verschaffen und dann zu schauen was nicht stimmt. Statt Diagnose mit der Lupe, Diagnose aus der Vogelperspektive



Wir messen nicht die Genexpression von einen oder zwei Genen, sondern von allen (... die wir kennen)

Wie? Parallele Hybridisierung auf einem **Microarray**

Expressionsprofil



Gewebe

Microarray



genome:~/ISBCoriginal		
ER+Nevsins4	d31628_s_at	253,3
ER+Nevsins4	d31628_s_at	1386,0
ER+Nevsins4	d31628_s_at	209,5
ER+Nevsins4	d31716_at	655,3
ER+Nevsins4	d31716_at	118,5
ER+Nevsins4	d31716_at	596,3
ER+Nevsins4	d31716_at	113,5
ER+Nevsins4	d31762_at	573,3
ER+Nevsins4	d31762_at	104,7
ER+Nevsins4	d31762_at	507,8
ER+Nevsins4	d31762_at	88,1
ER+Nevsins4	d31763_at	698,0
ER+Nevsins4	d31765_at	149,9
ER+Nevsins4	d31765_at	593,3
ER+Nevsins4	d31763_at	115,8
ER+Nevsins4	d31764_at	2993,5
ER+Nevsins4	d31764_at	426,6
ER+Nevsins4	d31764_at	2882,8
ER+Nevsins4	d31764_at	508,0
ER+Nevsins4	d31765_at	846,5
ER+Nevsins4	d31765_at	140,1
ER+Nevsins4	d31765_at	1039,5
ER+Nevsins4	d31765_at	207,3

Expressionsprofil:

Der diagnostische Befund

Eine Liste von 30.000 Genen und ihren Expressionswerten

Klinische Studie mit Expressionsprofilen



Ist das etwas qualitativ anderes als die Diagnose mit zwei Genen? **Oder müssen wir nur die Methoden von 2 auf 30.000 Dimensionen verallgemeinern?**

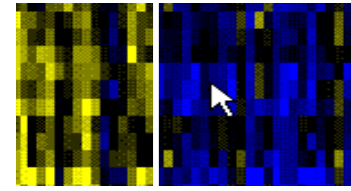


$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

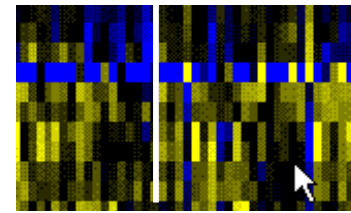
Normalenvektor im R^2
Normalenvektor im R^{30000}

$$\beta_0 + \sum_{i=1}^{30000} \beta_i x_i$$

Diagnose mit mehr als zwei Genen

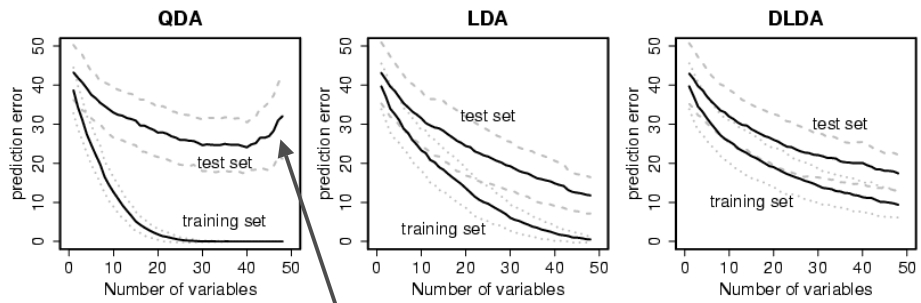


Differenziell exprimierte Gene ...



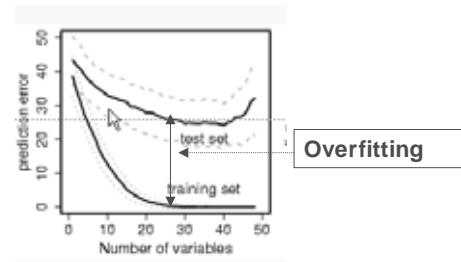
... oder multivariate Signatur

Diagnose mit 2-50 Genen



The test error can increase !

Overfitting nimmt zu

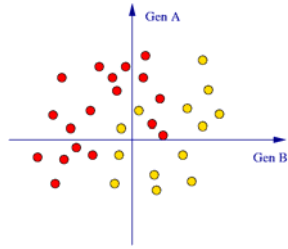


Mit mehr Genen hat man auch mehr Information über den Patienten ...

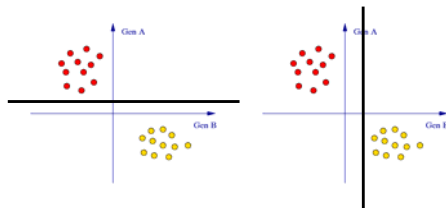
... Man sollte meinen, daß man jetzt weniger Fehler macht ...

... Das Gegenteil ist der Fall.

Probleme in **zwei** Dimensionen

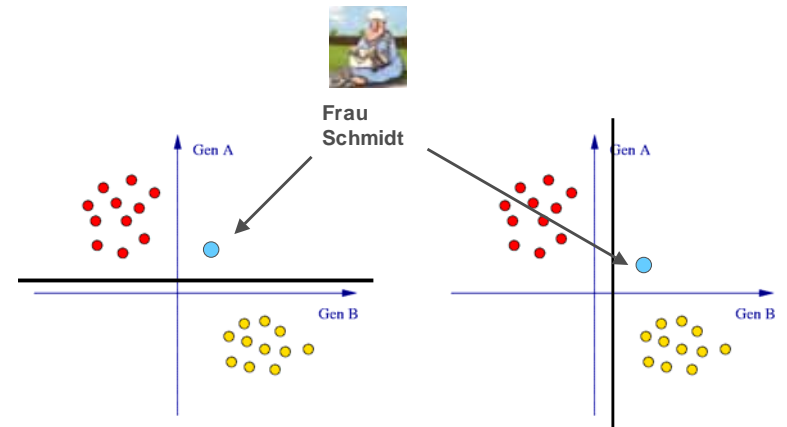


Problem 1:
Keine gute
Trenngerade

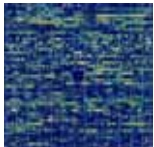


Problem 2:
Viele gute Trenngeraden
Wieso ist das ein
Problem?

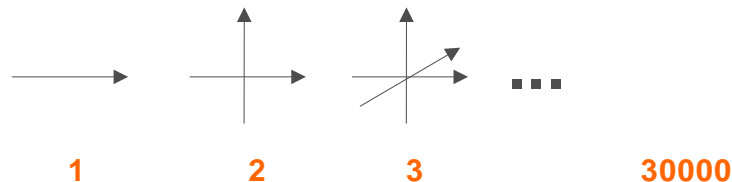
Welche Diagnose hat Frau Schmidt?



Der 30.000 dimensionale Raum



Bei den Microarrays arbeiten wir nicht in zwei, sondern in 30000 Dimensionen



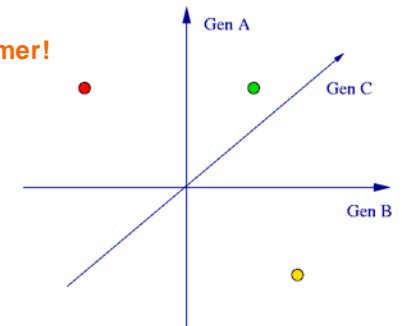
Und im 30000-dimensionalen Raum herrschen andere Gesetze

Mehr Gene als Patienten

- **Problem 1 entsteht nie!**
- **Problem 2 entsteht praktisch immer!**

Überlegen Sie sich das einmal kurz in drei Dimensionen

Also für drei Gene, zwei Patienten mit bekannter Diagnose, einem neuen Patienten und einer Trennebene statt einer Trenngerade

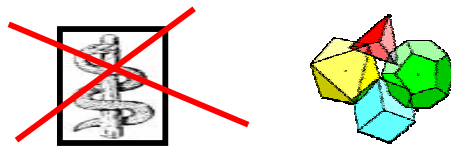


OK! Wenn alle Punkte auf einer Geraden liegen, geht es nicht immer. Das ist bei Messungen aber sehr unwahrscheinlich und kommt praktisch nie vor.

Das Overfitting-Desaster

Aus den Daten alleine kann man weder feststellen, welche Gene wichtig für die Diagnose sind, noch kann man zweifelsfrei eine Diagnose für den nächsten Patienten stellen.

Dieses Problem hat mit Medizin wenig zu tun. Es ist ein geometrisches Problem.



Bedeutungslose vs. Relevante Signaturen

Hat man eine perfekt trennende Signatur für die Trainingsdaten gefunden, heißt das im allgemeinen noch gar nichts. Es gibt ja immer eine. Und die meisten werden wohl völlig bedeutungslos sein. Und überhaupt nicht generalisieren.

Andererseits gibt es krankheitsbedingte Veränderungen in der Genexpression und die muß man auf dem Microarray auch sehen können.

Es gibt:

1. **Bedeutungslose Signaturen**
2. **und biologisch (klinisch) relevante Signaturen**

Wie kann man die auseinander halten?

Eine **biologische** und eine **statistische** Frage

Wie sehen biologisch relevante Signaturen aus?

... da müssen Sie einen Biologen fragen

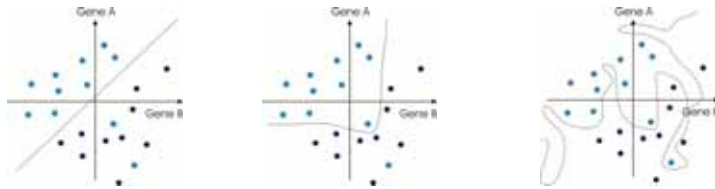
In der Statistikvorlesung stellen wir uns deshalb eher die Frage:

Wie sehen bedeutungslose Signaturen aus?

Eigenschaften bedeutungsloser Signaturen

- Hoher Testfehler
- Schlechte Generalisierung
- **Hohe Varianz** der Schätzer von Modellparametern
- **Flache Likelihoodfunktion** (viele gleichgute Lösungen, die Daten enthalten nicht genügend Information um ein Model zu identifizieren)
- *Unter Umständen aber niedriger Trainingsfehler*

Trainingsfehler und zu erwartender Testfehler



Training: 2 Fehler

1 Fehler

Kein Fehler

Test: ?

?

Hoch

Signaturraum und Lernregel

Wir definieren eine Klasse möglicher Signaturen, den **Signaturraum**. Zum Beispiel eine parametrisierte Familie von Entscheidungsfunktionen.

Beispiel: Alle Geraden im R^2

Parameter: β_0 (Offset), β_1 und β_2 (Normalenvektor)

Dann geben wir ein Verfahren an um aus Trainingsdaten eine Signatur aus dem Signaturraum auszuwählen. Zum Beispiel durch Angabe von Schätzern für die Modellparameter. Dies ist die **Lernregel**.

Signaturräume

- Alle quadratischen Flächen

- Alle linearen Ebenen

- Alle linearen Ebenen, die von höchstens 20 Genen abhängen

- Alle linearen Ebenen, die von 20 festen Genen abhängen

Hohe Wahrscheinlichkeit eine gut fittende Signatur zu finden

Niedrige Wahrscheinlichkeit, daß die Signatur etwas bedeutet



Niedrige Wahrscheinlichkeit eine gut fittende Signatur zu finden

Hohe Wahrscheinlichkeit, daß die Signatur etwas bedeutet

Zwei Strategien

Wie findet man klinisch relevante (generalisierende) Signaturen?

Wir werden nur zwei Strategien hier ansprechen

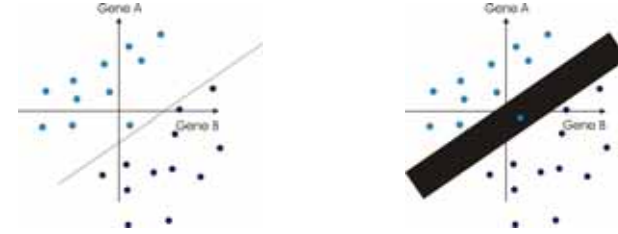
1. Geneselektion
2. Large Margins

Konzept **Genselektion**

Zieht man alle möglichen linearen Ebenen als Trennflächen in Betracht findet man immer eine die perfekt trennt, ohne daß uns die Biologie dabei helfen müßte.

Schränkt man sich auf Ebenen ein, die nur von 20 Genen abhängen ist dies nicht unbedingt der Fall. Finden wir trotzdem eine gut diskriminierende Ebene, haben wir eine gute Chance, daß sich hier eine krankheitsbedingte Abnormalität in der Genexpression widerspiegelt.

Konzept: **Large Margins**



Fette Ebenen: Man kann hochdimensionale Daten zwar immer linear trennen, braucht dafür aber eine unendlich dünne Ebene, mit einer „fetten Ebene“ ist dies nicht unbedingt möglich

Existiert so eine Large-Margin-Separation doch, stehen die Chancen gut, daß wir etwas relevantes gefunden haben.

Large margin classifiers, Support Vektor machines

Gemeinsame Idee

Die grundlegende Idee ist bei der Genselektion und bei Support Vector Machines (SVM) die gleiche.

Wir schränken die Menge aller möglichen Signaturen a priori ein. Dann ist es nicht mehr selbstverständlich eine fittende Signatur zu finden. Findet man sie doch, muß es einen Grund dafür geben. Ist es der für das Krankheitsbild entscheidende molekularbiologische Sachverhalt, haben wir gewonnen.

Regularisierung

Der Trick ist neben einer guten Trennung der Krankheitsklassen schon im Training **mehr zu verlangen.**

Diese zusätzlichen Vorschriften, die man den Modellen macht, werden nicht von den Daten unterstützt. Sie dürfen auch biologisch falsch sein (z. B. es sind mehr als 20 Gene die eine Rolle spielen).

Sie sind systematische Fehler, machen aber die Modelle einfacher.

Höherer Bias aber niedrigere Varianz der Schätzer.

Diese zusätzlichen Vorschriften nennt man **Regularisierung**

Regularisierte Modelle klingt besser als systematisch verfälschte Modelle, meint aber das gleiche.

Statistische **Lerntheorie**

Neben Geneselektion und SVM gibt es noch viele andere regularisierte Modelle:

Ridge Regression, LASSO, Kernel based methods, additive Models, classification trees, bagging, boosting, neural nets, relevance vector machines, nearest-neighbors, transduction etc. etc.

Das Fach, dass sich mit regularisierten Modellen beschäftigt heißt statistische Lerntheorie oder Machine Learning.

Moral von der **Geschicht'**

Nur der Voreingenommene kann etwas lernen!

Nearest Centroid Classification mit **zwei** Genen

$a_{1,1}, \dots, a_{1,100}, a_{2,1}, \dots, a_{2,100}$ group a
 $b_{1,1}, \dots, b_{1,100}, b_{2,1}, \dots, b_{2,100}$ group b

$$\bar{a} = (\bar{a}_1, \bar{a}_2)$$

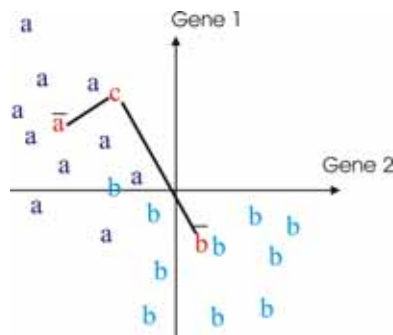
$$\bar{b} = (\bar{b}_1, \bar{b}_2)$$

$c = (c_1, c_2)$ Patient without diagnosis

Compare : $d_a = (\bar{a}_1 - c_1)^2 + (\bar{a}_2 - c_2)^2$ and

$$d_b = (\bar{b}_1 - c_1)^2 + (\bar{b}_2 - c_2)^2$$

Diagnosis : a if $d_a < d_b$
 b else



Nearest Centroid mit **n** Genen

$a_{i,j}$ Gene i in Patient j from group a
 $b_{i,j}$ Gene i in Patient j from group b

$$\bar{a} = (\bar{a}_1, \dots, \bar{a}_N)$$

$$\bar{b} = (\bar{b}_1, \dots, \bar{b}_N)$$

c_1, \dots, c_N Patient without diagnosis

Compare distances to the centroids :

$$d_a = \sum_{i=1}^N (\bar{a}_i - c_i)^2$$

$$d_b = \sum_{i=1}^N (\bar{b}_i - c_i)^2$$

Diagnosis : a if $d_a < d_b$
 b else

Der **Einfluß** einzelner Gene auf die Diagnose

$a_{i,j}$ gene i in patient j from group a
 $b_{i,j}$ gene i in patient j from group b

$$d_a = \sum_{i=1}^N (\bar{a}_i - c_i)^2$$

$$d_b = \sum_{i=1}^N (\bar{b}_i - c_i)^2$$

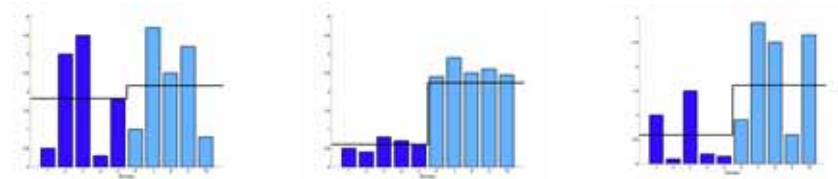
Diagnosis : a if $d_a < d_b$
 b else

Alle N Gene steuern im gleichen Maße zur Diagnose bei ...

Gewichte

Gene mit niedriger Varianz sollten **mehr Gewicht** bekommen als Gene mit hoher Varianz

$$d_a = \sum_{i=1}^N w_i (\bar{a}_i - c_i)^2 \quad d_b = \sum_{i=1}^N w_i (\bar{b}_i - c_i)^2$$



→ DLDA

Gene mit **wenig Varianz**

Die Varianz muß geschätzt werden

$$\sigma_i^2 = \frac{1}{n-2} \sum_{j=1}^{n/2} (a_{i,j} - \bar{a}_i)^2 + (b_{i,j} - \bar{b}_i)^2$$

pooled in class variance

In our case :

$$n = 200$$

Die geschätzte Varianz ist nicht die wahre Varianz. Sie kann größer oder auch kleiner sein. Wird eine ohnehin kleine Varianz noch unterschätzt, kann σ_i^2 sehr sehr klein sein und w_i unnatürlich hoch

Für ein einzelnes Gen passiert so etwas selten, aber bei 30000 Genen passiert es immer einige Male

Fudge Factor

$$w_i = 1/(\sigma_i + \sigma_0)^2$$

$$\sigma_0^2 = \text{median}(\sigma_1^2, \dots, \sigma_N^2)$$

Um diesem Large-Scale-Phänomen entgegen zu wirken, nehmen wir einen verfälschten Schätzer für die Varianz ... wir addieren eine Konstante σ_0 dazu ... so können keine übergroßen Gewichte w_i mehr entstehen.

σ_0 wird auch **Fudge Factor** genannt

Problem



Ist c ein a oder ein b?

c liegt näher am a-Zentroid als am b-Zentroid

c ist umkreist von b's

Es gibt viel mehr b- als a-Patienten

Wenn das nicht nur ein Artefakt der Stichprobe ist, sondern die Verhältnisse in der Population widerspiegelt, dann sollte c wohl als b klassifiziert werden.

Baseline Correction

π_a = relative size of group a
i.e. relative frequency of type a samples in the study, or expert knowledge
 $\pi_b = 1 - \pi_a$

$$d_a(c) = \sum_{i=1}^N \frac{(\bar{a}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_a$$

$$d_b(c) = \sum_{i=1}^N \frac{(\bar{b}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_b$$

Wir können dies erreichen, indem wir eine **baseline correction** durchführen

Modifiziere die Likelihood L durch multiplizieren mit der Prior Wahrscheinlichkeit π_k

Schätze π_k mit der relativen Häufigkeit von Typ-k Patienten in den Trainingsdaten

Discriminant Score

Abstand zum Zentroid

$$d_a(c) = \sum_{i=1}^N \frac{(\bar{a}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_a$$

$$d_b(c) = \sum_{i=1}^N \frac{(\bar{b}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_b$$

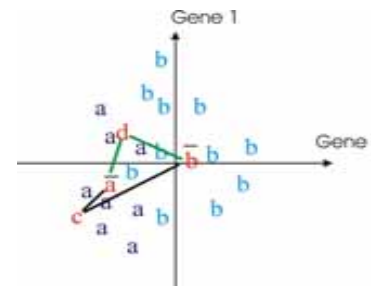
baseline correction

gepoolte Varianz

Varianz Regularisierungs Parameter

Klare und Unklare Diagnosen

Wieviel Evidenz gibt es für die Diagnose?



Sowohl c als auch d werden als a diagnostiziert, aber im Fall von d war das eine knappe Entscheidung

Klassifikations- Wahrscheinlichkeiten

Zwei Wege die Evidenz zu quantifizieren:

1. Likelihood Ratios (Anzahl weiße Bälle)
2. Klassifikations-Wahrscheinlichkeiten

$$\text{Prob} [Group(c) = a] = \frac{e^{-\frac{1}{2}d_a(c)}}{e^{-\frac{1}{2}d_a(c)} + e^{-\frac{1}{2}d_b(c)}},$$
$$\text{Prob} [Group(c) = b] = 1 - \text{Prob} [Group(c) = a]$$

Overfitting

$d_a(c) = d_b(c)$ definiert eine lineare Ebene

Wir arbeiten immer noch mit 30000 Genen

→ Overfitting

→ Die so konstruierte Ebene ist nicht unbedingt die optimale Trennebene

Das kann ein Vorteil oder auch ein Nachteil sein

Wir haben schon etwas regularisiert ... aber noch nicht genug

Variablenselektion

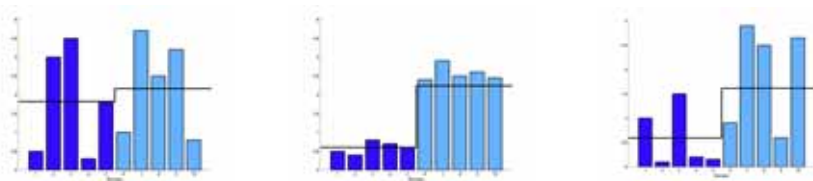
30000 Gene sind zuviel. Sie verursachen nur Overfitting

Gene, die mit der Krankheit nichts zu tun haben, verursachen nur Rauschen. Ihre Gewichte sind zwar niedrig ... aber wenn es Tausende davon gibt ...

Nicht alle Gene können eine Rolle spielen

→ Suchen wir welche aus

Welche ?



Die mit den höchsten Gewichten

Kleine Gewichte werden ganz auf Null gesetzt

Thresholding

Sagen wir, wir nehmen die 100 Gene mit den höchsten Gewichten

Gen Nr. 100 ist dabei und Gen Nr. 101 ist draußen,

Aber beide Gene waren in etwa gleich informativ

es gibt zwei Wege sich der Gene zu entledigen:

1. Abhacken

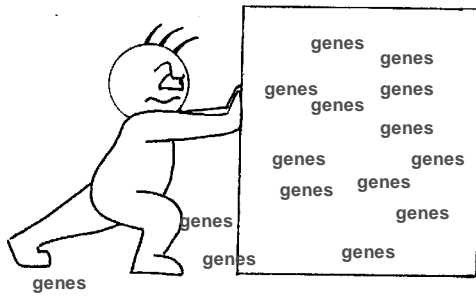


2. Hinausschieben

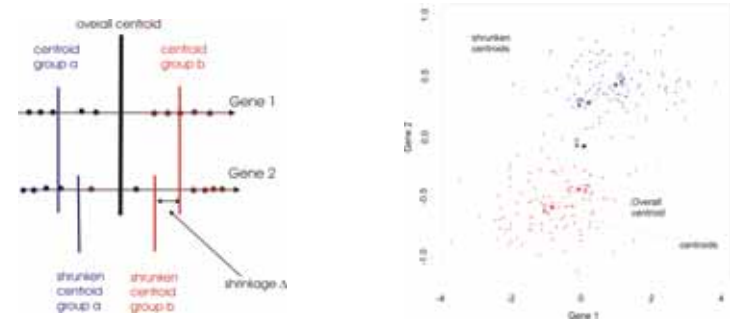


Die Shrunken Centroid Method und das Programm PAM

Tibshirani et al 2002



Centroid Shrinkage



Bewege die Klassenmittelwerte pro Gen auf den gesamt Mittelwert zu ... wenn man am gesamt Mittelpunkt angekommen ist, läßt man das Gen ganz verschwinden. Durch dieses genweise Shrinkage bewegen sich die Klassenzentroide auf den gesamt Zentroid zu

Notation

\bar{a}_i mean of gene i in group a
 \bar{b}_i mean of gene i in group b
 \bar{x}_i mean of gene i using all data
 Let

$$D_{i,a} = \frac{\bar{a}_i - \bar{x}_i}{m_a(\sigma_i + \sigma_0)}, \quad m_a = \sqrt{1/n_a + 1/n}$$

$$D_{i,b} = \dots$$
 or

$$\bar{a}_i = \bar{x}_i + m_a(\sigma_i + \sigma_0) D_{i,a}$$

$$\bar{b}_i = \dots$$

Shrinkage

group centroid overall centroid
 scaling factor

$$\bar{a}_i = \bar{x}_i + m_a(\sigma_i + \sigma_0) D_{i,a}$$
 offset

$$\bar{a}_i = \bar{x}_i + m_a(\sigma_i + \sigma_0) D'_{i,a}$$
 shrunken offset

$$D'_{i,a} = \text{sign}(D_{i,a}) (|D_{i,a}| - \Delta)_+$$
 shrinkage parameter

$$(\dots)_+ = \text{truncation at zero}$$

Tuning

Die Stärke des Shrinkage wird durch den Parameter Δ kontrolliert

Viel Shrinkage:

- Wenig Gene im Modell
- Höherer Trainingsfehler
- Bessere Generalisierung

Wenig Shrinkage:

- Viele Gene im Modell
- Niedriger Trainingsfehler
- Schlechte Generalisierung

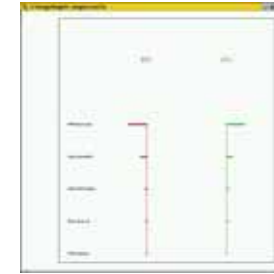
PAM = Predictive Analysis of Microarrays

Dieses ganze Verfahren ist in dem R-Paket pamr implementiert

Brustkrebs

Brustkrebsdiagnose mit Microarrays

Die Aufgabe war es den Estrogen-Rezeptor-Status des Tumors am Expressionsprofil zu bestimmen



In der Studie waren:

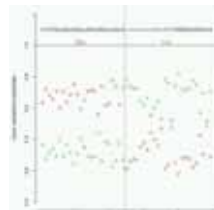
- 7000 Gene
- 49 Brusttumore
- 25 ER+
- 24 ER-

Nach dem Shrinkage bleiben im Modell nur noch 5 Gene übrig

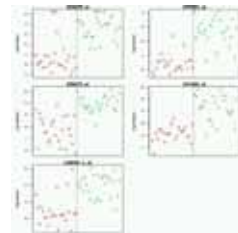
Die Länge der Balken gibt ihr Gewicht an, die Richtung ob sie in der Tumorklasse erhöht oder reduziert gemessen werden

Mehr Information

Rechnet man die Klassenwahrscheinlichkeiten aus, erhält man folgendes Bild

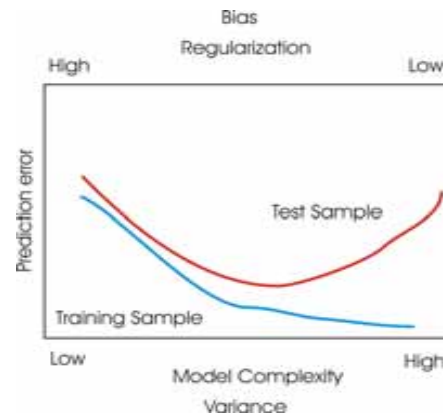


Darin sind die Expressionswerte der 5 Gene, die das Shrinkage überlebt haben kombiniert



Der bias variance trade off

Wie stark soll man regularisieren?

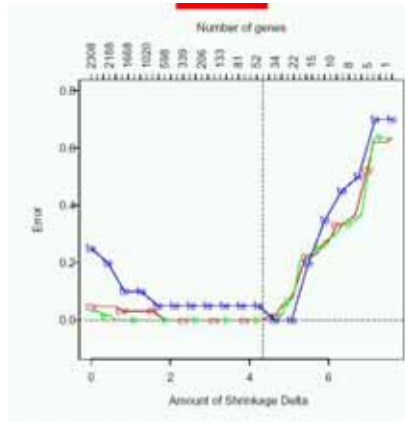


Modelkomplexität:

- Anzahl Gene
- Shrinkageparameter
- Minimal Margin
- etc

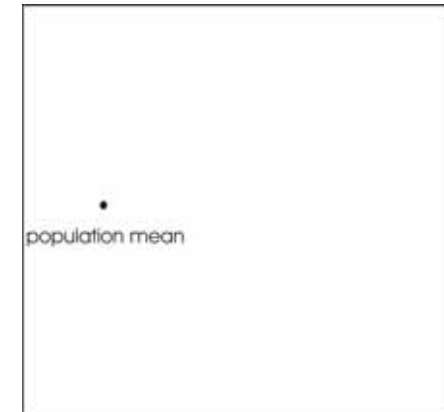
Beispiel

Small round blue cell tumors
4 classes
(Data: Khan et al. 2001)
(Analysis (PAM): Hastie et al 2002)



Populations-Parameter

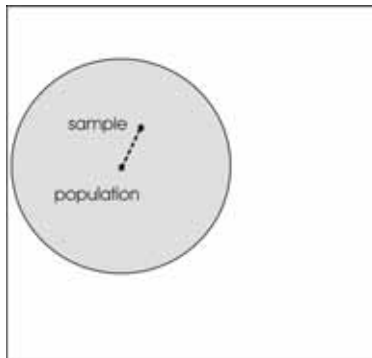
Gene haben eine bestimmte mittlere Expression und Korrelation in der Population



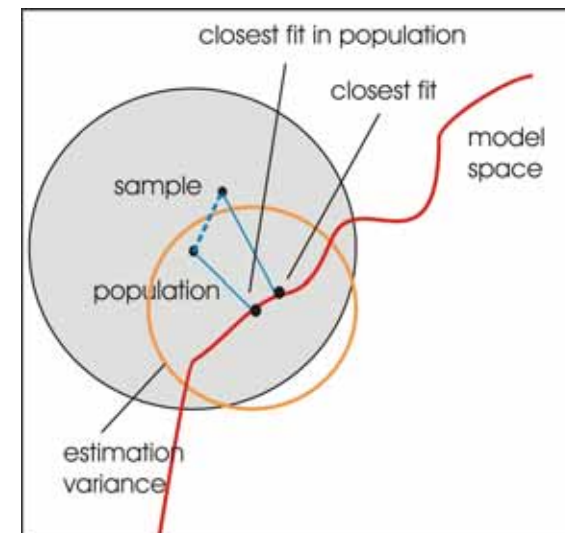
Geschätzte Parameter

Die entsprechenden empirischen Größen, die wir aus einer Stichprobe ermitteln weichen von den Populations-Parametern ab

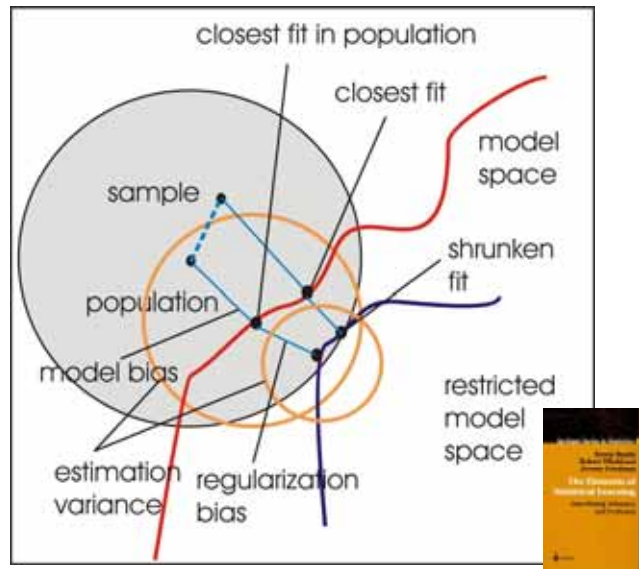
Varianz des Schätzers



Das trainierte Modell



Das **regularisierte** Modell



Zusammenfassung:

- Regularisierung
- Variablenselektion
- Baseline Correction
- Centroid Shrinkage