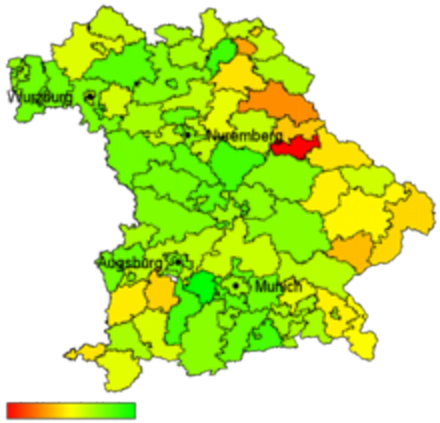
Statistische Methoden in der Epidemiologie

Disease Mapping
Analyse geographischer Variabilität

Prof. Dr. Ulrich Mansmann
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie
mansmann@ibe.med.uni-muenchen.de

Analyse geographischer Variabilität

- Analyse von Mortalitätsraten und Inzidenzraten
- Geographische Verteilung und Identifikation von Regionen mit hohen oder niederen Raten
- Zwei klassische Vorgehensweisen:
  Verwendung der Poissonverteilung
  Karten mit standardisierten Raten
  Karten mit statistischer Signifikanz (Abweichung der stand. Rate von 1)



niedrig hoch

Standardisierte Nutzungsrate der Screening-Koloskopie in Bayern, adjustiert nach Alter und Geschlecht.

Karten standardisierter Raten

- Vorteil:
  - Karten stellen Schätzer für interessierende Größen dar.
- Nachteil:
  - Variabilität der Beobachtungen ist größer als die aufgrund des Poisson-Modells erwartete Variabilität:
    Variabilität durch Poisson-Sampling und Mischung verschiedener Poisson-Verteilungen aufgrund verschiedener individueller Risikoniveaus.
  - Konventionelle Poisson-Modellierung berücksichtigt kaum die räumliche Struktur der Daten:
    benachbarte Gebiete haben ähnliche Raten
    lokale geographische Abhängigkeiten.

Bayesianische Verfahren für Disease Mapping

Verwendung von Apriori-Information zu

- der Variabilität von Raten über die Karte hinweg
  Ziel: Liegen Daten zu vielen Ereignissen vor, so wird der Schätzer nahe an der standardisierten Rate liegen, wurden nur wenige Ereignisse beobachtet, so wird der Schätzer maßgeblich von der apriori-Verteilung bestimmt
  Shrinkage
- Verteilungen mit Berücksichtigung der Beeinflussung zwischen benachbarten Regionen
  Wurden nur wenige Ereignisse in einer Region beobachtet, so ist deren Ratenschätzer nahe am Mittelwert der Raten aus den benachbarten Regionen.
  Shrinkage

$$P[\text{Parameter}|\text{Daten}] \sim P[\text{Daten}|\text{Parameter}] \cdot P[\text{Parameter}]$$

$$\text{Posterior} \sim \text{Likelihood} \cdot \text{Prior}$$

Modellierungsansätze: Poisson

Karte ist in n benachbarte Regionen aufgeteilt, die mit i = 1, …, n indiziert sind.

$y = (y_1, \ldots, y_n)$ ist der Vektor mit den beobachteten Todes- (Erkrankungs-)zahlen in den n Regionen während des Studienzeitraums.

$e = (e_1, \ldots, e_n)$ ist der Vektor mit den erwarteten Todes- (Erkrankungs-)zahlen in den n Regionen während des Studienzeitraums.

Die erwartete Ereignisanzahl läßt sich für eine Region aus der Mischung der alters- und geschlechtsspezifischen erwarteten Ereignisanzahlen berechnen. Wobei die Population währen der Studie hinsichtlich der adjustierten Parameter als stabil angesehen werden kann.

$r = (r_1, \ldots, r_n)$ ist der Vektor der unbekannten gebietsspezifischen Mortailtäts- oder Morbiditätsrisiken.

Wenn die Erkrankung nicht ansteckend und selten ist, so wird die Verteilung der Ereignishäufigkeit mittels Poisson-Verteilung modelliert und als unabhängig zwischen den Gebieten angenommen. Hieraus ergibt sich die folgende Likelihood für das unbekannt relative Risiko

$$[y_i|r_i] = \exp\{-e_i \cdot r_i\} (e_i \cdot r_i)^{y_i} / (y_i!)$$

Notation

[X|Y]	Bedingte Verteilung von X gegeben Y
[X]	Marginale Verteilung von X
[X,Y]	Gemeinsame Verteilung von X und Y

ML - Schätzungen für relative Risiken

ML-Schätzer: $r_{ML} = r^* = y/e$

Standardfehler: $s = y^{0.5}/e$

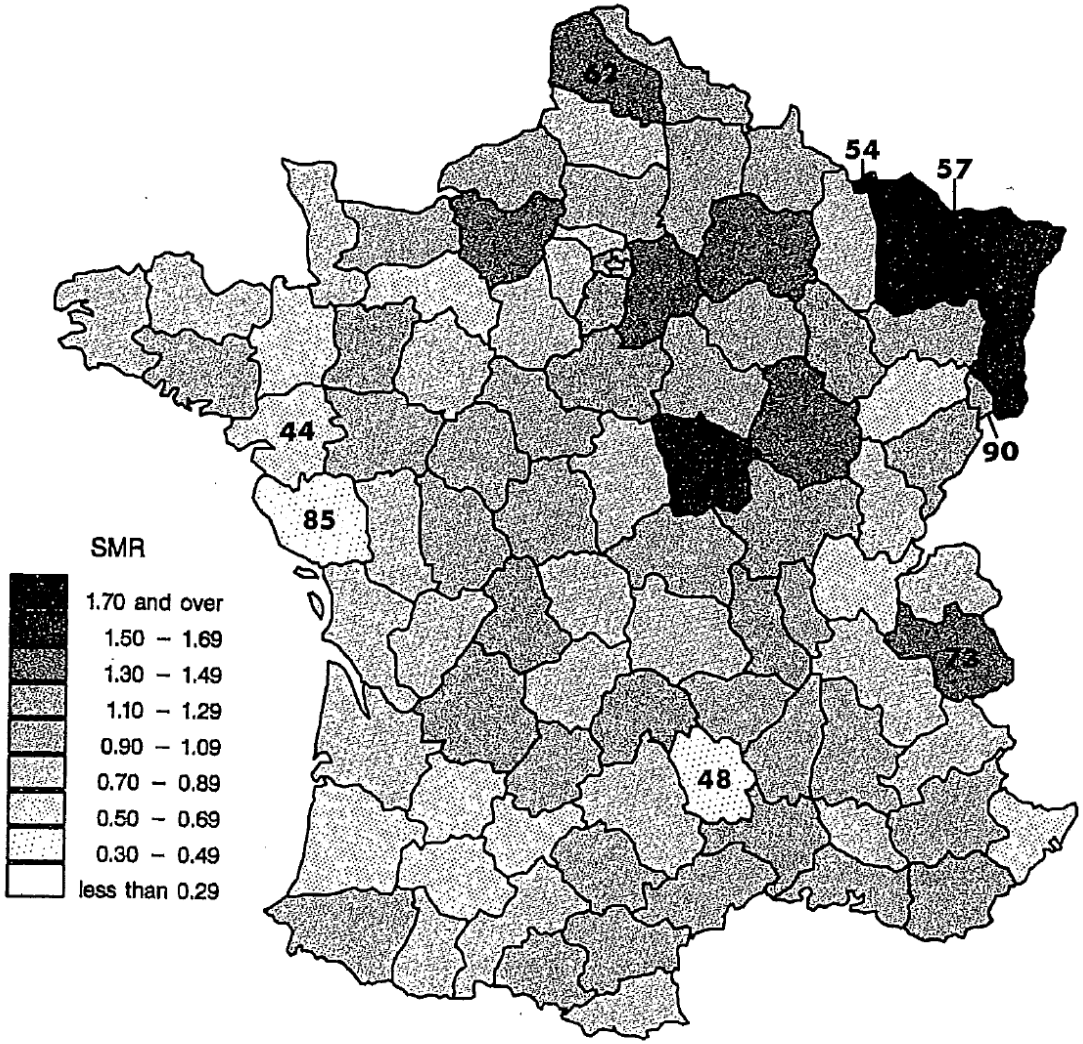
p-Wert: $p = 2 \cdot [1-\text{pnorm}(\text{abs}(p^*-r_{ML})/s,0,1)]$

Gallenblasen- und Gallengangstumore (ML)

number	département	expected number of deaths	SMR	standard error of SMR	p-value
44	Loire-Atlantique	48.34	0.58	0.11	0.002
48	Lozère	6.46	0.31	0.22	0.088
54	Meurthe-et-Moselle	37.66	1.75	0.22	$< 10^{-4}$
57	Moselle	46.40	1.72	0.19	$< 10^{-4}$
62	Pas-de-Calais	73.59	1.33	0.13	0.006
73	Savoie	18.05	1.38	0.28	0.092
75	Paris	158.35	0.92	0.08	0.348
85	Vendée	28.94	0.35	0.11	$< 10^{-4}$
90	Territoire-de-Belfort	6.74	1.19	0.42	0.476

Mollié A., Richardson S. (1991) Empirical Bayes estimates of cancer mortality rates using spatial models, Statist. Med.,10:95-112

Gallenblasen- und Gallengangstumore (ML)



Mollié A., Richardson S. (1991) Empirical Bayes estimates of cancer mortality rates using spatial models, Statist. Med.,10:95-112

Gallenblasen- und Gallengangstumore (ML)

- Die SMRs variieren
  Minimum 0.31 im Departement 48 und Maximum 1.75 im Departement 54
- Standardfehler der SMRs variieren
  Minimum 0.076 im Departement 75 und Maximum 0.419 im Departement 90
- Die Bevölkerungszahlen variieren
- Gibt es einen Süd-West → Nord-Ost Trend?
- SMR berücksichtigt nicht die Größe der zugrundeliegenden Population:
  Extreme SMRs bei kleinen Populationen mit wenigen oder vielen Fällen?
- p-Werte hängen stark von der Populationsgröße ab:
  Kleinste p-Werte finden sich in Regionen mit den größten Populationen?
- Wenn seltene Erkrankungen in Regionen mit geringer Population untersucht werden, machen die beiden letztgenannten Punkte die epidemiologische Interpretation von SMR- oder p-Wertkarten schwer oder unmöglich.

Gallenblasen- und Gallengangstumore (ML)

Genauer Blick auf die Daten:

- Gallenblasen- und Gallengangstumore haben eine geringe Mortalitätsrate bei französischen Männern: $1.57 \cdot 10^{-5}$.
- SMR_{48}: ist klein (0.31) aber nicht signifikant von 1 verschieden (p=0.088)
- SMR_{44}: ist näher an der 1 (0.58) aber signifikant von 1 verschieden (p=0.002) Die Population dieses Departements ist 12 mal so groß wie die von D 44.
- SMR_{73}: ist groß (1.38) aber nicht signifikant von 1 verschieden (p=0.092) aufgrund der geringen Bevölkerungsdichte.
- SMR_{62}: ist näher an der 1 (1.33) aber signifikant von 1 verschieden (p=0.006) aufgrund der hohen Bevölkerungsdichte.

Diese Situation erschwert die epidemiologische Interpretation der SMRs und p-Werte.

Weiteres Problem: multiples Testen!

Overdispersion

Seltene Erkrankungen in Gebieten mit geringer Einwohnerzahl zeigen oft *Overdispersion*: Es leigt eine Variabilität vor, die größer ist als aufgrund des verwendeten Poissonmodells erwartet.

Die zusätzliche Variabilität kann ins Modell eingebaut werden, wenn man die relativen Risiken r_i innerhalb des Gebietes variable macht oder Verteilungen für Zähldaten verwendet, die eine größere Variabilität als die Poisson-Verteilung haben.

1.) Hierarchische Modelle

2.) Negativbinomial-Verteilung

Bayesianische hierarchische Modelle für RRs

Verwendung Bayesianischer Methoden zur Berechnung regularisierter (geglätteter) Schätzer für SMRs.

Selbst unter Vernachlässigung geographischer Strukturen führt dieses Verfahren zu Schätzungen mit einem kleineren quadratischen Fehler verglichen zur dargestellten ML-Schätzung.

Dies gilt schon für Probleme ab 3 Regionen.

Bayesianische Verfahren kombinieren 2 Formen von Information:

- Information der in einer Region beobachteten Fälle mittel Poisson-Likelihood [y|r]
- Information über die allgemeine Variabilität relativer Risien, gegeben in einer Apriori-Verteilung [r]

Bayesianische hierarchische Modelle für RRs

Es wird angenommen, dass die y_i voneinander unabhängig sind, falls die r_i bekannt sind:

$$[y \mid r] = \prod_{i=1}^{n} [y_i \mid r_i]$$

Die Aprioi-Verteilung zur Variabilität der SMRs im betrachteten Gebiet wird mit einem Parameter ρ (Hyperparameter) parametrisiert.

Die gemeinsame Aposteriori-Verteilung der Parameter (r,ρ) wird dann wie folgt beschrieben:

$$[r,\rho|y] \sim [y|r]\ [r|\rho]\ [\rho].$$

Die marginale Verteilung von r gegeben die Beobachtungen y ist damit:

$$[r \mid y] = \int [r,\rho \mid y]\, d\rho = \int [y \mid r] \cdot [r \mid \rho] \cdot [\rho]\, d\rho = \int \prod_{i=1}^{n} [y_i \mid r_i] \cdot [r \mid \rho] \cdot [\rho]\, d\rho$$

Regionale Abhängig- oder Unabhängigkeit?

Bayesianische hierarchische Modelle für RRs

Schätzer für die Menge der relativen Risiken:

- Mittelwert der Aposteriori-Verteilung E[r|y]
- Modus der Aposteriori-Verteilung (MAP): Maximiere [r|y] bezüglich r

Umgang mit dem Hyperparameter ρ:

- Der Wert von ρ ist in den seltensten Fällen bekannt
- Empirical Bayes: Der Hyperparameter ist nicht bekannt und stammt aus einer nicht spezifizierten Verteilung [ρ]
- Bayes: Spezifizieren die die Verteilung [ρ] (Hyperprior) , eventuell durch Einführen weiterer Hyperparameter.
  Führt zu mehrstufigen hierarchischen Modellen, mindestens 3 Stufen:
  (1) Daten - (2) SMRs - (3) ρ

Spezifikation der Apriori-Verteilungen

$$[r \mid y] = \int [r, \rho \mid y]\, d\rho = \int [y \mid r] \cdot [r \mid \rho] \cdot [\rho]\, d\rho = \int \prod_{i=1}^{n} [y_i \mid r_i] \cdot [r \mid \rho] \cdot [\rho]\, d\rho$$

Wie sieht [r|ρ] genau aus?

Unstrukturierte Heterogenität der RRs (1)

Falls ρ bekannt ist, sind die relativen Risiken austauschbar: $[r \mid \rho] = \prod_{i=1}^{n} [r_i \mid \rho]$

Dabei wir angenommen, dass die Apriori-Verteilungen $[r_i|\rho]$ für alle Gebiete gleich sind.

In der Praxis werden zwei Ansätze verwendet:

- Gamma-Verteilung G(ν,α) mit Mittelwert ν/α und Varianz ν/α²
  $[r_i|\rho] = [r_i|\nu,\alpha] \sim \alpha^{\nu} \cdot (r_i)^{\nu-1} \cdot \exp\{-\alpha \cdot r_i\}$

$$[r \mid \rho] = [r \mid \nu, \alpha] = \prod_{i=1}^{n} [r_i \mid \nu, \alpha]$$

- Log-normal-Verteilung: Normalverteilung mit Mittelwert μ und Varianz σ² für die log-transformierten RRs ($x_i = \log[r_i]$).

$$[x \mid \rho] = [x \mid \mu, \sigma^2] = \prod_{i=1}^{n} [x_i \mid \mu, \sigma^2]$$

Unstrukturierte Heterogenität der RRs (2)

Es ist im log-normalen Modell leicht möglich, spezifische Kovariablen für die einzelnen Gebiete zu berücksichtigen:

$$\mu = Z\phi$$

Dabei ist Z eine nxp Matrix, die für die n Gebiete die Ausprägungen von p Kovariaten enthält. Der Vektor $\varphi = (\varphi_1, \ldots, \varphi_p)$ enthält die Effekte der Kovariablen. Mit $\mu_i = (Z\varphi)_i$ gilt

$$[x \mid \rho] = [x \mid \mu, \sigma^2] = \prod_{i=1}^{n} [x_i \mid \mu_i, \sigma^2]$$

Im Gamma-Apriori-Modell können auch Kovariablen betrachtet werden. Schätzer aus dem Modell mit dem Gamma-Prior sind besonders robust.

Morris CN (1983) Parametric empirical Bayes inference, JASA, 78:47-65

Das Gamma-Modell kann nicht leicht auf eine räumliche Abhängigkeitsstruktur erweitert werden.

Räumlich strukturierte Variabilität von RRs

Geographisch benachbarte Gebiete haben oft ähnliche relative Risiken, die Variabilität der RRs ist somit lokal strukturiert.

Die formale Modellierung dieses Sachverhaltes erfolgt mittels Markov Random Fields (MRF).

MRF:
Die bedingte W‘keit für das RR im Gebiet i gegeben die Werte aller anderen Gebiete ist gleich der bedingten W‘keit für das RR im Gebiet i gegeben die Werte aller seiner direkten Nachbarn (δ_i).

Die gemeinsame Verteilung $r = (r_1, \ldots, r_n)$ ist bis auf die Normalisierungs-konstante durch die auf die Nachbarn bedingte Verteilungen charakterisiert.

Gauss‘sche MRF für log-transformierte RRs:
Mittelwert von x_i in Region i hängt von den Mittelwerten der benachbarten Regionen ab. Bedingte Varianzen sind konstant (homogene Karte, Gebiete ähneln sich)

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems JRSS B, 36:192-236.
Clayton D., Kaldor J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, Biometrics, 43: 671-681

Intrinsische Gauss‘sche Autoregression

Bernardinelli L, Montomodi C (1992) Empirical Bayes versus fully Bayeian analysis of geographical variation in disease risk. Statist. Med. 11: 983-1007

Irreguläre Karten: Die bedingte Varianz von x_i gegeben x_{-i} ist inverse proportional zur Anzahl der angrenzenden Gebiete w_{i+}.

Definition der gemeinsamen Apriori-Verteilung:

$$[x \mid \rho] = [x \mid \sigma^2] \propto \frac{1}{\sigma^n} \cdot \exp\left[-\frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j<i} w_{ij} \cdot (x_i - x_j)^2 \right]$$

$E[x|\rho] = 0$

Inverse Kovarianzmatrix hat die Diagonalelemente w_{i+}/σ^2 und außerhalb der Diagonalen an Position (i,j) den Eintrag $-w_{ij}/\sigma^2$.

$w_{ij} = 0$ falls i und j nicht benachbart sind, sonst $w_{ij}=1$. (Andere nicht-negative Gewichtung möglich)

$$w_{i+} = \sum_{j=1}^{n} w_{ij}$$

Intrinsische Gauss‘sche Autoregression

Betrachte den Fall: w_{ij} = 1 falls i und j benachbart sind.

$$E[x_i \mid x_{-i}, \sigma^2] = E[x_i \mid x_i, i \in \delta_i, \sigma^2] = \frac{1}{|\delta_i|} \sum_{j \in \delta_i} x_j$$

$$Var[x_i \mid x_{-i}, \sigma^2] = Var[x_i \mid x_i, i \in \delta_i, \sigma^2] = \frac{\sigma^2}{w_{i+}}$$

Gauss‘sches Konvolution Prior

Im praktischen Fall fällt die Entscheidung zwischen einem räumlich strukturierten und einem unstrukturierten Prior schwer.

Besag J et al. (1991) Bayesian image restauration, with two applications in spatial statstics, Ann. Inst. Statist. Math, 43:1-21

Im Konvolutionsprior ist x eine Summe aus zwei unabhängigen Komponenten

$$x = u + v$$

Die Variable v wird nach der Idee der log-normalen Verteilung bei unstrukturierter Heterogenität entwickelt (Folie 17)

$$[v \mid \lambda] = [v \mid 0, \lambda^2] = \prod_{i=1}^{n} [x_i \mid 0, \lambda^2]$$

Die Variable u wird nach der intrinsischen Gauss‘schen Autoregression (Folie 20) modelliert.

$$[u \mid \kappa^2] \propto \frac{1}{\kappa^n} \cdot \exp\left[-\frac{1}{\kappa^2} \sum_{i=1}^{n} \sum_{j<i} w_{ij} \cdot (x_i - x_j)^2 \right]$$

Gauss‘sches Konvolution Prior

Die bedingte Varianz von x_i ist:

$$\mathrm{Var}[x_i \mid x_{-i}, \kappa, \lambda] = \mathrm{Var}[x_i \mid x_j, j \in \delta_i, \kappa, \lambda] = \frac{\kappa^2}{w_{i+}} + \lambda^2$$

Die Parameter λ und κ kontrollieren die Stärke und den Einfluß der einzelnen Komponenten.

Falls κ^2/λ^2 nahe beim Mittelwert der w_{i+} liegt, so haben beide Komponenten den gleiche Einfluß.

Falls κ^2/λ^2 klein ist, so hat die unstrukturierte Komponenten den Haupteinfluß.

Falls κ^2/λ^2 groß ist, so dominiert die räumlich strukturierte Komponente.

In der Komponente v (unstrukturierte Komponente) können wie in Folie 18 dargestellt auch Kovariateneffekte mitmodelliert werden.

Graphische Modelle

Graphische Modelle helfen bei der Formalisierung komplexer Verteilungen durch die systematische Darstellung der beteiligten Komponenten und deren Interaktionen.

Für die dargestellten hierarchische Modelle geben Graphen Details der Verteilung wider.

In jedem Graphen werden Annahmen zur bedingten Unabhängigkeit dargestellt.

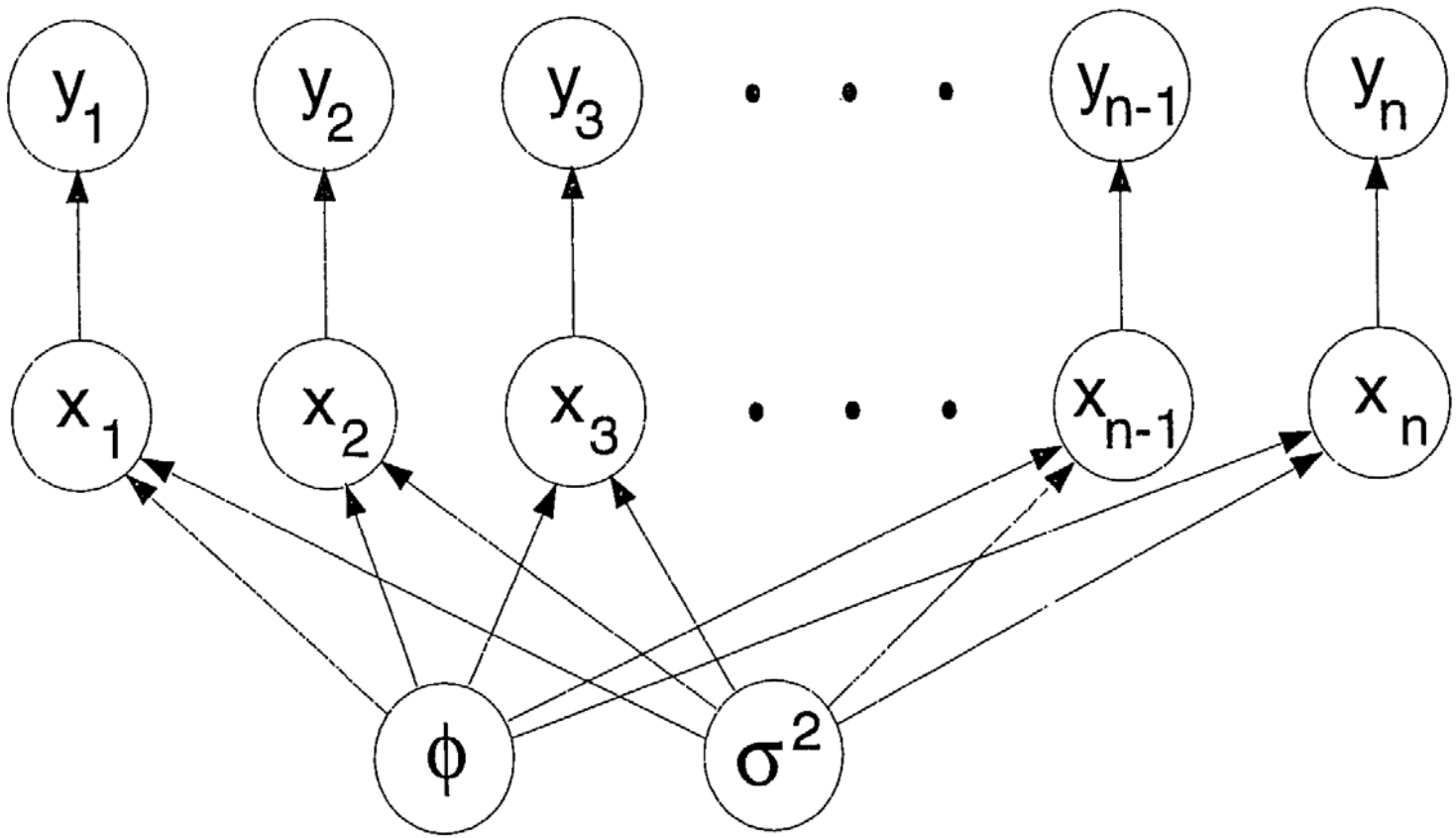
Unterschiede bestehen in der Spezifikation der Priors und in der Verwendung von Modellen für direkten oder log-transformierten RRs.

Graphische Modelle

Bisher wurde angenommen:

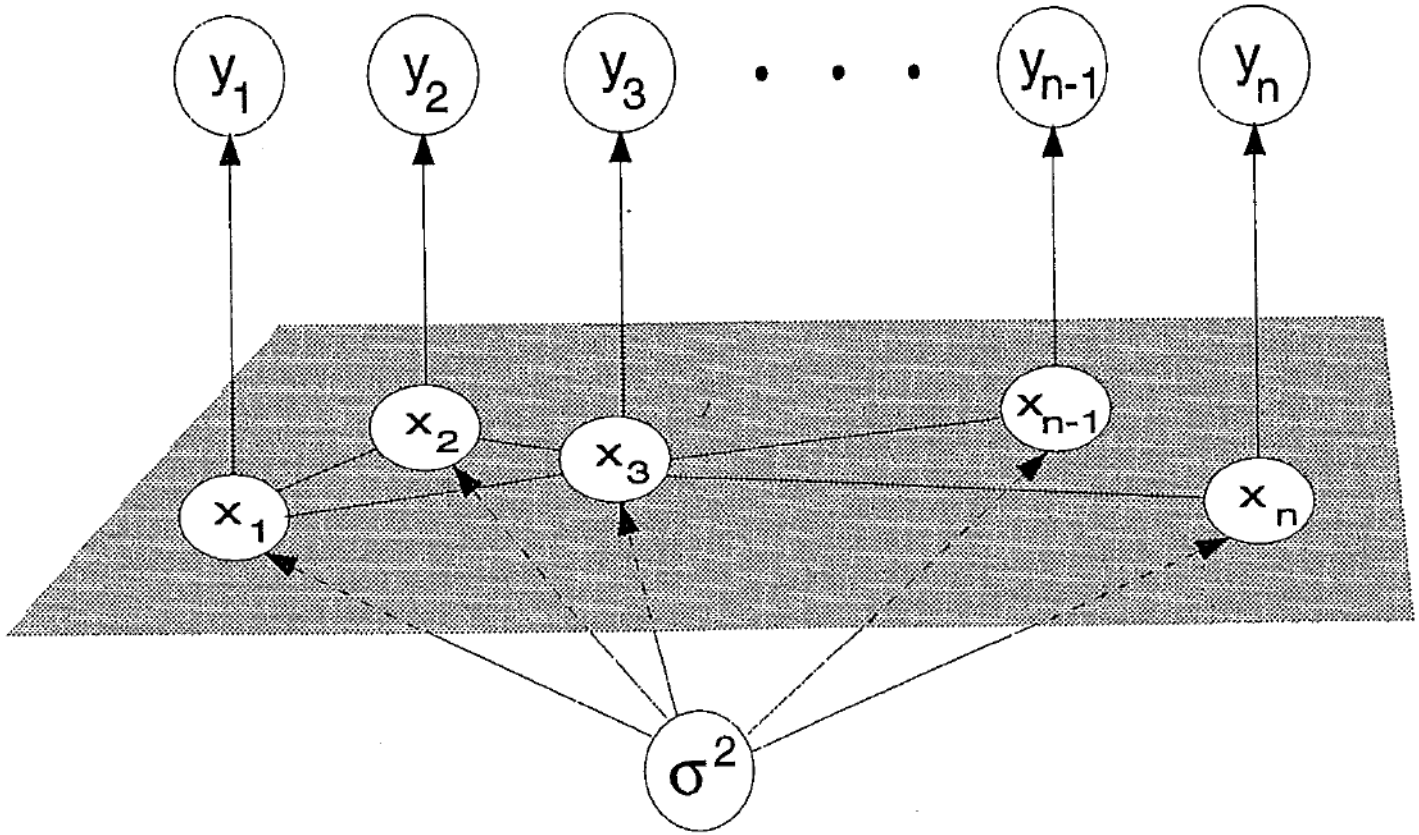
- y ist unabhängig von ρ falls x oder r gegeben ist.
  Somit kann es keine direkten Pfeile zwischen den y und den Hyperparametern geben
- Die Komponenten von y sind bedingt unabhängig gegeben r oder x
  Es kann keine Pfeile zwischen den y-Komponenten geben
- y_i ist unabhängig von x_{-i} gegeben x_i
  Es kann keine Pfeile zwischen y_i und Komponenten in x_{-i} geben.
- Bedingte Unabhängigkeit für die Priors,
  MRF-Bedingung: Direkte Abhängigkeit nur von den direkten Nachbarn
  Unstrukturierte Prior: Komponenten von x sind unabgängig bedingt auf die Hyperparameter. Es gibt aber direkte Pfeile von den relevanten Hyperparametern auf jede x-Komponente.

Graphische Modelle



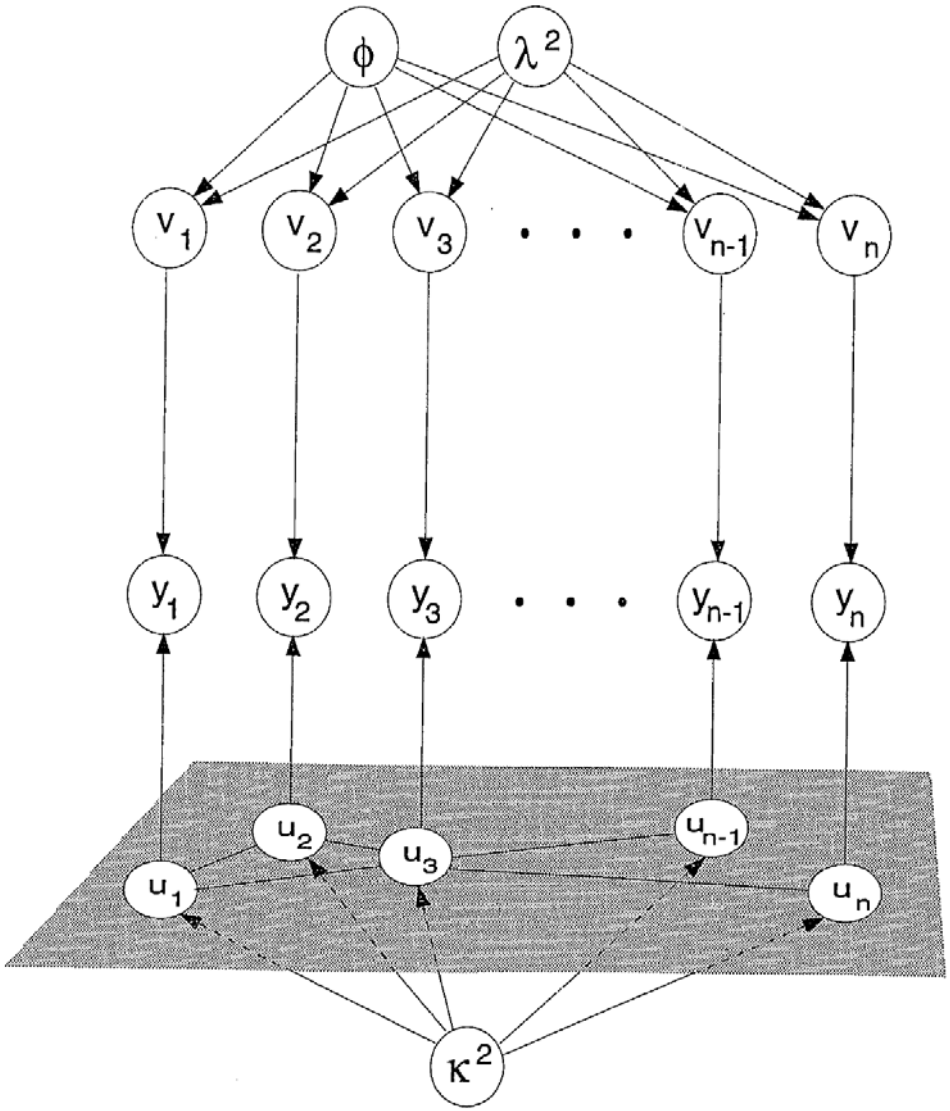
Unstrukturiertes hierarchisches Modell mit log-normalen RRs

Graphische Modelle



Intrinsisches Gauss‘sches Autoregressions Modell mit log-normalen RRs

Graphische Modelle



Hierachisches Modell mit
Gauss‘scher Konvolutions
Apriori-Verteilung

Empirische Bayes Schätzer für RRs

Bayes-Ansatz:

$$[r|y,\rho] \sim [y|r]\,[r|\rho]$$

Ersetze den unbekannten Hyperparameter ρ durch einen geeigneten Schätze ρ*.

In der Regel sind die Schätzer dieser Hyperparameter ML-Schätzer, die aus Information abgeleitet werden, die es über die globale Karte gibt.

Herausintegrieren der Wirkung einzelner Regionen und verwenden der marginalen Likelihood von ρ:

$$[y \mid \rho] = \int [y \mid r] \cdot [r \mid \rho]\, dr$$

Der *empirical Bayes* Schätzer (EB) ist dann der Mittelwert der Aposteriori-Verteilung [r|y,ρ*]: E [r|y,ρ*].

Konjugierte Gamma-Verteilung

- Die Poissonverteilung beschreibt das Auftreten seltener Ereignisse in grossen Populationen.
- $P_\lambda(y) = \exp\{-\lambda\}\ \lambda^y/y!$ wobei der Parameter λ die mittlere Anzahl der Ereignisse (pro Einheit) angibt: $\lambda = e \cdot r$.
- Die Gamma-Verteilung hat die Dichte

$$\text{Gamma}(r \mid \alpha, \nu) = r^{\nu-1} \cdot e^{-\alpha r} \cdot \alpha^{\nu} / \Gamma(\nu)$$

- Gibt man dem Parameter r als a priori Verteilung eine Gamma-Verteilung, so ist die a posteriori Verteilung gegeben die Beobachtung von y Ereignissen wieder eine Gamma-Verteilung:

$$P(r \mid y) = \text{Gamma}(r \mid y + \nu, \alpha + e)$$

- Somit ist die Gamma-Verteilung eine konjugierte Familie für das Poisson-Modell
- Gamma(α,ν)-Verteilung: Mittelwert ν/α und Varianz = ν/α^2

Konjugierte Gamma-Verteilung

Auf Folie 17 haben wir die unstrukturierte Version der Apriori-Verteilung wie folgt formuliert:

$$[r \mid \rho] = [r \mid \nu, \alpha] = \prod_{i=1}^{n} [r_i \mid \nu, \alpha]$$

Für diese Apriori-Verteilung ist

$$[r \mid y, \rho] = [r \mid y, \nu, \alpha] = \prod_{i=1}^{n} [r_i \mid y_i, \nu, \alpha]$$

Die Posterior $[r_i|y_i,\nu, \alpha]$ ist eine Gamma-Verteilung $Ga[y_i+\nu,e_i+\alpha]$ mit Mittelwert

$$E[r_i \mid y_i, \alpha, \nu] = \frac{y_i + \nu}{e_i + \alpha} = \omega_i \cdot SMR_i + (1 - \omega_i) \cdot \frac{\nu}{\alpha}$$

Dabei ist $SMR_i = y_i/e_i$ und $\omega_i = e_i / (e_i+\alpha)$

Somit ist der aposteriori Mittelwert des RR in Region i das gewichtete Mittel der SMR aus Region i und des globalen Mittelwertes ν/α über die ganze Karte.

Das Gewicht ist invers zur Varianz des SMR_i.

Konjugierte Gamma-Verteilung

Berechnung der MLEs für α und ν:

Maximiere die marginale Likelihood $[y|\alpha,\nu] \rightarrow \alpha^*$ und ν^*
Diese ist das Produkt von n negativ-binomial Verteilungen $[y_i|\alpha,\nu]$.

Der EB-Schätzer des RR in Region i ist dann

$$E[r_i|y_i, \alpha^*, \nu^*]$$

Gallenblasen- und Gallengangskrebs (EB)

Berechnung der MLEs für α und ν: $\alpha^* = 25.508$ und $\nu^* = 25.223$

Die EB-Schätzer der RRs zeigen weniger Variabilität als bei der ML-Analyse.

Vergleich mit ML-Tabelle!

Variabilität von 0.65 (D 85) bis 1.46 (D 57)

EB-Schätzer haben MW 0.99 und STD 0.15 (0.27 bei den ML geschätzten SMRs)

Gallenblasen- und Gallengangskrebs (EB)

number	département	EB estimate	Bayes estimate	95% Bayes credible interval
44	Loire-Atlantique	0.72	0.72	0.57 – 0.90
48	Lozère	0.85	0.88	0.68 – 1.13
54	Meurthe-et-Moselle	1.44	1.55	1.28 – 1.88
57	Moselle	1.46	1.63	1.34 – 1.96
62	Pas-de-Calais	1.24	1.25	1.03 – 1.50
73	Savoie	1.15	1.03	0.81 – 1.35
75	Paris	0.93	0.95	0.81 – 1.08
85	Vendée	0.65	0.69	0.53 – 0.88
90	Territoire-de-Belfort	1.03	1.21	0.92 – 1.59

Diskussion (EB)

Ist die Apriori-Verteilung keine konjugierte Verteilung so muss [r|y,ρ] auf kompliziertere Weise berechet werden (Approximation). Clayton und Kaldor benutzen eine multivariate Normalverteilung als Approximation an [x|y,ρ].

Berechnung der marginalen Likelihood bei nicht-konjugierten Priors wird ebenso schwer und verwendet oft den EM Algorithmus.

Die Darstellung des EB-Schätzers als gewichtetes Mittel bleibt bestehen auch wenn nicht-konjugierten Priors vorliegen.

Weil die EB-Verfahren keine Unsicherheitsannahmen bezüglich ρ machen, wird die Variabilität des Schätzers in der Regel unterschätzt. Die Punktschätzer sind gut.

Eine bessere Schätzung der Variabilität ist mit Bootstrapverfahren und einer Modifikation der Delta-Regel möglich.

Laird NM, Louis TA, 1987, JASA, 82:739-750
Louis TA, 1991, Statist. Med., 10: 811-829

Der EB-Ansatz erlaubt keine angemessene Beschreibung der wahren Overdispersion in den Raten.

Devine OJ, Louis TA, 1994, Statist. Med. 13:1119-1133

Bayesianische Analyse für RRs

Bayesianische Analyse relativer Risiken basiert auf der Aposteriori-Verteilung

$$[r|y].$$

Diese ist oft schwer algorithnisch zu behandeln und verlangt komplizierte Approximationsverfahren.

Weiterhin handelt es sich um eine Marginale Verteilung nach Ausintegrieren anderer Nuisance-Parameter.

MCMC-Verfahren erlauben eine schnelle Berechnung eines Samples aus

$$[r,\rho \mid y]$$

und liefern die marginale Verteilungen

$$[r \mid y] \text{ und } [\rho \mid y].$$

MCMC - Grundidee

- Konstruktion einer a posteriori Verteilung aus der Gleichgewichtsverteilung einer Markovkette.
- Mit Hilfe eines geeigneten Algorithmus ziehen wir eine Stichprobe $\theta_0, \theta_1, \theta_2, \theta_3, \ldots$ aus der unbekannten a posteriori Verteilung.
- Die unbekannte a posteriori Verteilung lässt sich dann durch das Historgramm der Stichprobenwerte $\theta_0, \theta_1, \theta_2, \theta_3, \ldots$ approximieren.
- Zentrales Object ist $\pi(\theta) = p(\theta|\text{Data})$, die gemeinsame a posteriori Verteilung aller Parameter gegeben die Daten.
- Aussagen über einen bestimmten Parameter erhält man durch die Betrachtung der entsprechenden eindimensionalen Randverteilung.
- Anstelle der a posteriori Verteilung werden oft Kenngrößen dieser Verteilung angegeben: Momente oder Quantilen.
- Die Folge θ kann durch irgendeinen Prozess generiert werden, solange es sich um eine Stichprobe aus π handelt. Eine Möglichkeit: Markovkette mit π als stationärer Verteilung.

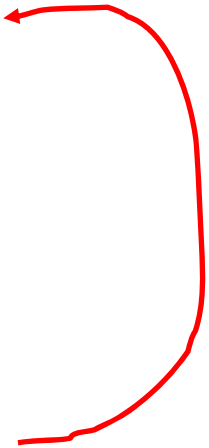
MCMC - Grundidee

- Gibbs-Sampling
  Erzeugen einer Kette für den r-dimensionalen Parametervektor

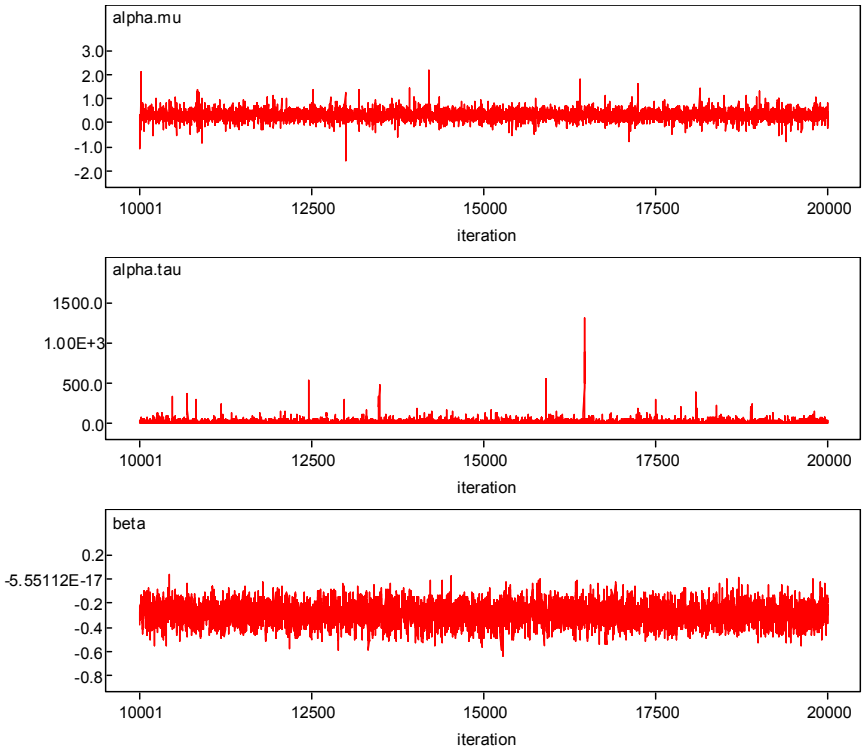
$$\theta_k = (\theta_k^1, \theta_k^2, ..., \theta_k^r)$$

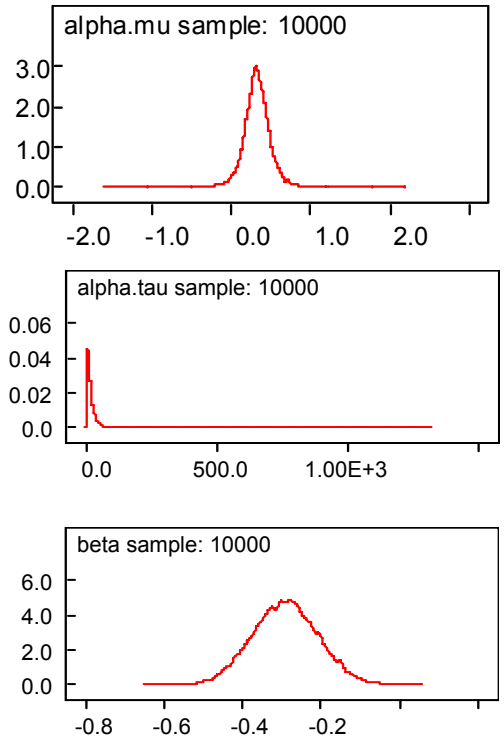
- Durchlaufe folgende Zyklen um aus θ_k den nächsten Kandidaten θ_{k+1} zu erzeugen:

$$\theta_{k+1}^1 \sim P[\cdot | \theta_k^2, ..., \theta_k^r, Data]$$
$$\theta_{k+1}^2 \sim P[\cdot | \theta_{k+1}^1, \theta_k^3, ..., \theta_k^r, Data]$$
$$\downarrow$$
$$\theta_{k+1}^{r-1} \sim P[\cdot | \theta_{k+1}^1, ..., \theta_k^{r-2}, \theta_k^r, Data]$$
$$\theta_{k+1}^r \sim P[\cdot | \theta_{k+1}^1, ..., \theta_k^{r-1}, Data]$$



MCMC - Grundidee





MCMC - Grundidee

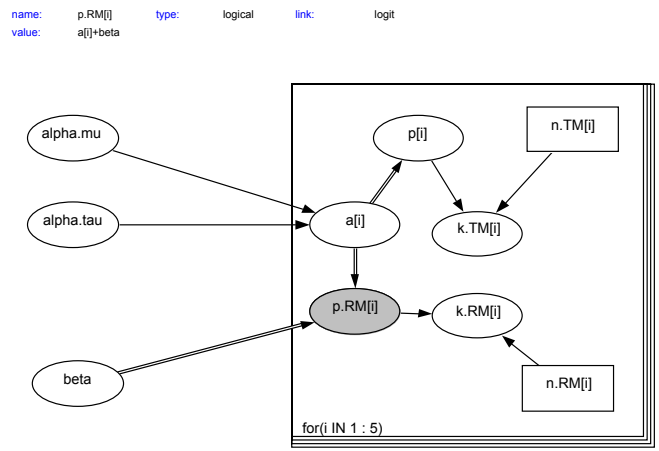
- Stationär: selbstähnlich, es gibt keine Entwicklung mehr.
- Markovkette: Es gibt eine Regel, wie der Zustand zur Zeit k+1 aussieht, wenn der Zustand zur Zeit k bekannt ist:

  $$\theta_{k+1} \sim P(\bullet|\theta_k)$$
- Die Markovkette vergisst ihre Vergangenheit.
- Unter milder Voraussetzungen besitzt eine Markovkette eine stationäre Verteilung – Sie vergisst Ihren Ursprung

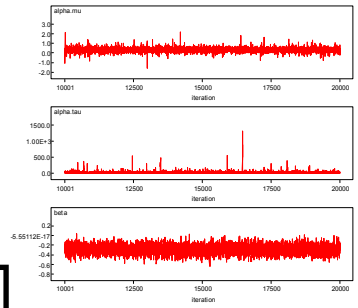
  $$\pi(\theta) = \lim_{k->\infty} P[\theta_k=\theta|\theta_0]$$
- Wie kann man einen Übergangsmechanismus für eine Markovkette konstruieren, damit deren stationäre Verteilung genau die von mir gesuchte a posteriori Verteilung ist?
  Metropolis Hastings
  Gibbs-Sampling

MCMC - Grundidee

Modellüberlegung
Verteilungen
Struktur
a priori Annahmen

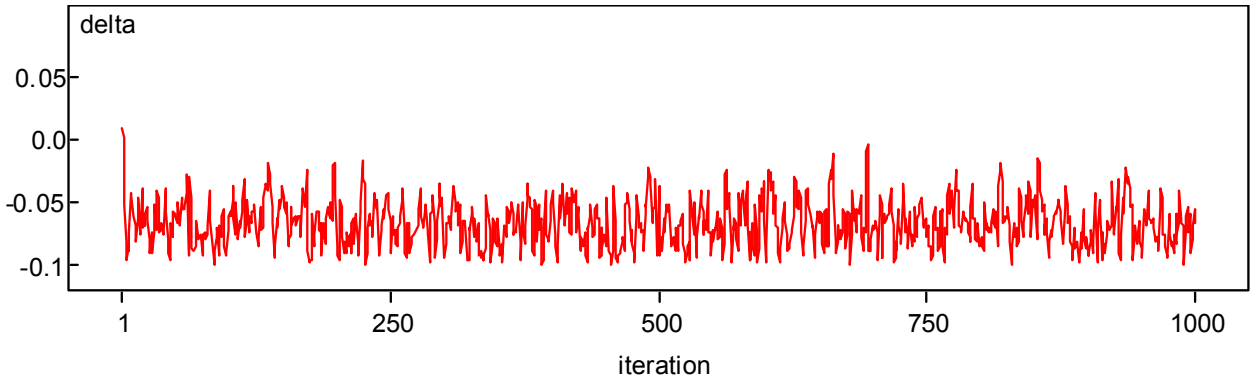


Bestimmung der Likelihood, Verbinden mit a priori Verteilung, Bau einer Übergangsvorschrift für eine Markovkette deren stationäre Verteilung der gesuchten a posteriori Verteilung entspricht.



Ableitung wichtiger Kenngrößen für die Inferenz

MCMC - Grundidee



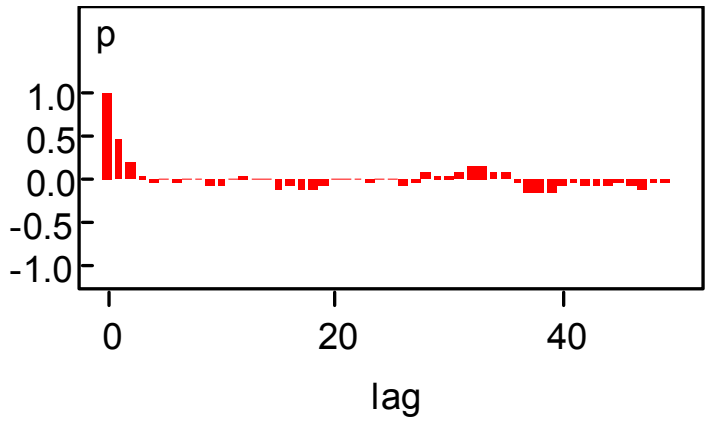
Burn In:
Wandern der Kette in das Gleichgewicht:
stationäre Maß.

Sampling:
Verwenden der stationären Verteilung der Markovkette zum Sampeln der gesuchten a posteriori Verteilung.

MCMC – Praktische Anwendung

Eine lange oder eine kurze Kette?

Eine lange Kette ist oft besser um Feinheiten der a posteriori Verteilung zu sehen. Oft sind auch die aufeinanderfolgenden Kettenglieder stark voneinander abhängig.
So muß die Kette manchmal ausgedünnt werden um „unabhängige“ Infomationsbeiträge zu erhalten.



Viele kurze Ketten sind meist besser, um die Konvergenz zu überprüfen: Vergleich der Variabilität zwischen den Ketten mit der innerhalb der Ketten.

MCMC – Praktische Anwendung

- Wahl der Startwerte
  Eigentlich egal, Ziel: kurze Burn-in Periode
- Burn-In
  Wie groß ist der Burn-In zu wählen?
  Beachte Konvergenzverhalten der Markov-Kette
  Konvergenz-Diagnostik
- Wann stoppen?
  Wann ist ein adäquater Erwartungswert und adäquate Varianz für einen interessierenden Parameter erreicht?
  MC Error: Kette solange laufen lassen, bis MC Error kleiner als 5% der a posteriori Stichprobenstandardabweichung ist.

$$\sqrt{n} \cdot [\bar{f}_k - E_\pi(f(\theta))] \rightarrow N(0, \sigma^2)$$

- σ^2 wird geschätzt, $\sigma / \sqrt{n}$ heißt Monte Carlo Standard Error MC error.

Konstruktion des Gibbs-Samplers (1)

Apriori-Verteilungen für Varianzen von Normalverteilungen:

Konjugierte Gamma-Verteilungen G(a,b) für die inverse Varianz $\theta = \sigma^{-2}$.
MW = a/b, Varianz=a/b^2
Die Parameter a und b erhalten feste positive Werte

Der Wert von a/b soll eine gute Schätzung für die erwartete inverse Varianz θ^* darstellen.

Der Wert von $a/b^2 = \theta^*/b$ soll etwa eine Schätzung für die Unsicherheit im Parameter θ^* quantifizieren. Der Wert von b wird in der Regel klein gewählt.

Im Gauss'schen autroregressiven Modell wird empfohlen $\theta^* = 1 / (w' \cdot s_x^2)$ zu setzen. Dabei ist w' der Mittelwert der w_{i+}.

Im Gauss'schen Konvolutionsmodell wird empfohlen für den Parameter $1/\kappa^2$ den Wert $2 / (w' \cdot s_x^2)$ zu zu wählen (Dabei ist w' der Mittelwert der wi_+) und für $1/\lambda^2$ den Wert $2 / (s_x^2)$.

Konstruktion des Gibbs-Samplers (2)

Gemeinsame Verteilung der log-transformierten RRs als Produkt von einfachen bedingten Verteilungen:

$$[x, \gamma \mid y] \propto \prod_{i=1}^{n} [y_i \mid x_i][x \mid \gamma][\gamma]$$

γ ist der Vektor der Hyperparameter

Konditionale Verteilung für die x_i

Konditionale Verteilung für die Hyperparameter

Konstruktion des Gibbs-Samplers (3)

Bedingte Verteilungen für die x_i:

$$[x_i \mid x_{-i}, y, \gamma] \propto \exp\left\{ y_i x_i - e_i \exp\{x_i\} - \frac{(x_i - \mu_i)^2}{2\sigma^2} \right\}$$

Intrinsisches Gauss'sches autoregressives Modell: $\bar{x}_i = \frac{1}{w_{i+}} \sum_{j=1}^{n} w_{ij} x_j$

$$[x_i \mid x_{-i}, y, \gamma] \propto \exp\left\{ y_i x_i - e_i \exp\{x_i\} - \frac{w_{i+}(x_i - \bar{x}_i)^2}{2\sigma^2} \right\}$$

Gauss'sches Konvolutions Modell: $\bar{u}_i = \frac{1}{w_{i+}} \sum_{j=1}^{n} w_{ij} u_j$

$$[u_i \mid u_{-i}, v, y, \gamma] \propto \exp\left\{ y_i u_i - e_i \exp\{u_i + v_i\} - \frac{w_{i+}(u_i - \bar{u}_i)^2}{2\kappa^2} \right\}$$

$$[v_i \mid v_{-i}, u, y, \gamma] \propto \exp\left\{ y_i v_i - e_i \exp\{u_i + v_i\} - \frac{(v_i - \mu_i)^2}{2\lambda^2} \right\}$$

Konstruktion des Gibbs-Samplers (4)

Bedingte Verteilungen für die Hyperparameter:

φ Regressionskoeffizienten der Kovariaten für $\mu=(\mu_i)$

$$[\varphi|x,y,\sigma^2] \sim [x|\varphi, \sigma^2]\,[\varphi] \sim N[(Z^tZ)^{-1}Z^tx, \sigma^2(Z^tZ)^{-1}]$$

σ^2: Die inverse Varianz $\theta=1/\sigma^{-2}$ hat die Verteilung

$$Ga[a+n/2, b+0.5\cdot(x-Z\varphi)^t(x-Z\varphi)] \qquad \text{MRF mit Kov}$$

$$Ga[a+n/2, b+0.5\cdot\sum_{i=1}^{n}\sum_{j<i} w_{ij}(x_i-x_j)^2] \qquad \text{intrinsisch}$$

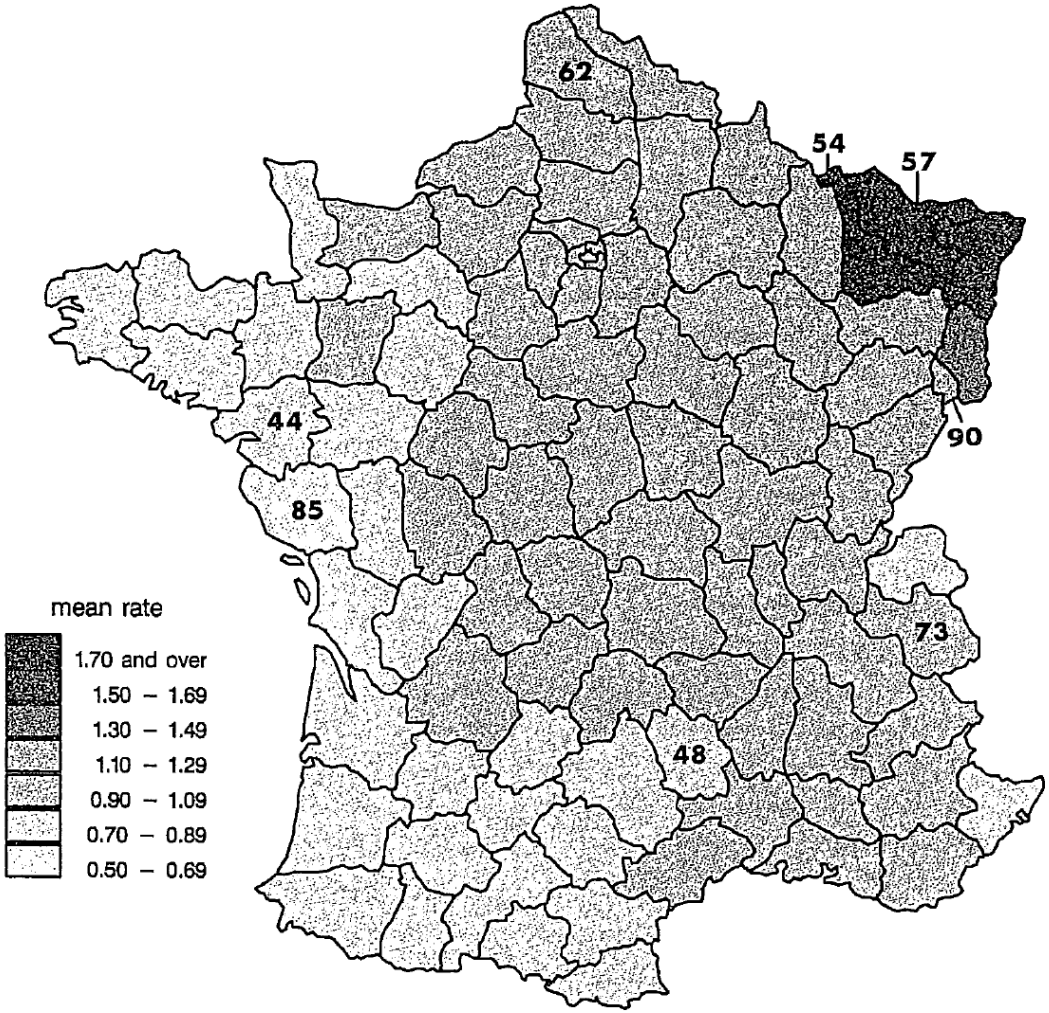
Hyperparameter im Gauss'schen Konvolutionsmodell:

$$[\varphi|u,v,y,\lambda^2,\kappa^2] \sim N[(Z^tZ)^{-1}Z^tv, \lambda^2(Z^tZ)^{-1}]$$

$$[1/\lambda^2|u,v,y,\varphi,\kappa^2] \sim Ga[a_1+n/2, b_1+0.5\cdot(v-Z\varphi)^t(v-Z\varphi)]$$

$$[1/\kappa^2|u,v,y,\varphi,\lambda^2] \sim Ga[a+n/2, b+0.5\cdot\sum_{i=1}^{n}\sum_{j<i} w_{ij}(u_i-u_j)^2]$$

Bayesianische Analyse für RRs



Bayesianische Schätzungen der relativen Mortalitätsrisiken an Gallenblasen und Gallengangkarzinomen bei französischen Männern zwischen 1971 und 1978.

Zusammenfassung

- Im Falle der Analyse seltener Erkrankungen in kleinen Regionen erlaubt eine Bayesianische Strategie eine Modellierung der Überdispersion in den klassischen SMRs
- Shrinkage der SMRs erlaubt eine bessere Interpretation der Ergebnisse.
- Die EB und Bayes Methode ergeben vergleichbare Punktschätzer. Der Bayes-Ansatz liefert jedoch bessere Intervallschätzer.
- MCMC-Verfahren erlauben eine Handhabung der komplexen Berechnungen im Bayes-Ansatz.
- Die Verwendung des Bayes-Ansatzes verlangt jedoch die Wahl geeigneter Apriori-Verteilungen. Die Apriori-Verteilungen bestimmen die Gewichtung zwischen regionalen und globalen Eigenschaften der Karte.
- Das Gauss‘sche Konvolutions Modell bietet hier gute Modellierungs-möglichkeiten
- Entwicklung neuer Prioris: Beachtung natürlicher Diskontinuitäten in den Daten: Gebirge, Flüsse, natürliche Grenzen, …