

Metodologia Bayesowska

- Czym jest prawdopodobieństwo?
 1. Interpretacja klasyczna: prawdopodobieństwa są miarą częstości z jaką zjawisko występuje
 - (a) Podstawowe twierdzenie klasycznej statystyki: częstości empiryczne dążą do prawdopodobieństw
 - (b) Prawdopodobieństwo (opisane np. dystrybucją) jest granicą częstości empirycznych
 2. Interpretacja Bayesowska: prawdopodobieństwa są odbiciem subiektywnej wiedzy na temat rzeczywistości

- (a) Podstawowe twierdzenie statystyki Bayesowskiej: twierdzenie Bayesa
 - (b) Prawdopodobieństwo jest odbiciem naszej niepełnej wiedzy na temat rzeczywistości opartej na *apriorycznych* sądach i danych empirycznych
- Oba podejścia są zgodne z aksjomatami rachunku prawdopodobieństwa
 - Zasadnicza różnica w przypadku małych prób
1. Metodologia klasyczna:
- (a) problem niedostatecznej liczby stopni swobody
 - (b) problem trudnej do oszacowania różnicy między rozkładami asymptotycznymi i małopróbkowymi
2. Podejście Bayesowskie
- (a) przy małych próbach korzystamy prz estymacji głównie z wiedzy *a priori*

(b) rozkłady małopróbkowe można *w zasadzie* zawsze znaleźć

Modelowanie

- Prosty model regresji liniowej

$$y_i = x_i \beta + \varepsilon_i$$

1. W podejściu klasycznym

- (a) błędy ε_i są losowe i nieznane
- (b) parametry modelu są deterministyczne ale nieznane
- (c) szukamy takich funkcji danych, która najszybciej zbiegają do parametrów modelu (estymatory)
- (d) w małych próbkach estymatory są zawsze zmiennymi losowymi

2. W podejściu Bayesowskim:

- (a) nieznane elementy modelu są traktowane jako zmienne losowe
- (b) niepełna wiedza błędów losowych i parametrów opisana za pomocą rozkładów prawdopodobieństwa
- (c) wstępna wiedza na temat parametrów opisana rozkładem *a priori*
- (d) na podstawie twierdzenia Bayesa formułować rozkład *a posteriori* parametrów, który uwzględnia wiedzę *a priori* i posiadane dane empiryczne

Twierdzenie Bayesa

- Łączną dystrybuantę dla $\theta \in \Theta$ możemy zapisać jako

$$f(\theta, \mathbf{y}) = f(\mathbf{y}|\theta) f(\theta) = f(\theta|\mathbf{y}) f(\mathbf{y})$$

- $f(\theta)$ jest gęstością rozkładu *a priori* parametrów
- $f(\mathbf{y}|\theta) = L(\theta)$ jest funkcją wiarygodności
- Rozkład *a posteriori* parametrów spełnia równanie

$$g(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta) f(\theta)}{f(\mathbf{y})},$$

- Mianownik tego wzoru można znaleźć zauważając, że

$$\int_{\Theta} f(\theta | \mathbf{y}) d\theta = \frac{\int_{\Theta} f(\mathbf{y} | \theta) f(\theta) d\theta}{f(\mathbf{y})} = 1$$

a więc

$$f(\mathbf{y}) = \int_{\Theta} f(\mathbf{y} | \theta) f(\theta) d\theta$$

i wzór Bayesa zapisuje się niekiedy jako

$$f(\theta | \mathbf{y}) = \frac{f(\mathbf{y} | \theta) f(\theta)}{\int_{\Theta} f(\mathbf{y} | \theta) f(\theta) d\theta}$$

- Dystrybucja brzegowa $f(\mathbf{y})$ nie zależy θ
- $g(\theta | \mathbf{y})$ traktowana jako funkcja θ jest proporcjonalna do licznika $f(\mathbf{y} | \theta) g(\theta)$

- Zapisujemy to

$$f(\boldsymbol{\theta} | \mathbf{y}) \propto f(\mathbf{y} | \boldsymbol{\theta}) f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) f(\boldsymbol{\theta}),$$

gdzie \propto oznacza "jest proporcjonalne do"

- Funkcję

$$g(\boldsymbol{\theta}) = f(\mathbf{y} | \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

nazywamy jądrem rozkładu *a posteriori*. Nie jest ona zazwyczaj funkcją gęstości ponieważ nie całkuje się do jedynki.

Estymacja punktowa i przedziałowa

- Estymacja punktowa w przypadku metod bayesowskich polega najczęściej na wyznaczeniu wartości oczekiwanej θ z rozkładu *a posteriori*

$$\hat{\theta} = E(\theta | \mathbf{y}) = \int_{\Theta} \theta f(\theta | \mathbf{y}) d\theta$$

- Często za $\hat{\theta}$ przyjmujemy dominantę $f(\theta | \mathbf{y})$:

$$\hat{\theta} = \arg \max_{\theta} f(\theta | \mathbf{y})$$

- Jeśli dysponujemy dystrybuantą brzegową $f(\theta_i | \mathbf{y})$ dla θ_i to za $\hat{\theta}_i$

możemy przyjąć medianę $f(\theta_i | \mathbf{y})$:

$$\int_{-\infty}^{\hat{\theta}_i} f(\theta_i | \mathbf{y}) d\theta_i = \frac{1}{2}$$

- Znalazienie przedziału ufności dla skalarnego θ polega na znalezieniu takich θ_1 i θ_2 , że przy poziomie ufności α

$$\int_{\theta_1}^{\hat{\theta}_i} f(\theta_i | \mathbf{y}) d\theta_i = \int_{\hat{\theta}_i}^{\theta_2} f(\theta_i | \mathbf{y}) d\theta_i = \frac{1 - \alpha}{2}$$

- Przyjęcie konkretnej formy estymatora uzasadnia się niekiedy przyjęciem określonej funkcji straty $\mathcal{L}(\hat{\theta}, \theta)$, to jest funkcji, która opisuje stratę związaną z błędem estymacji.

- Wybieramy ten estymator, który minimalizuje oczekiwaną stratę przy danym rozkładzie *a posteriori*:

$$\hat{\theta} = \arg \min \mathbb{E} \left[\mathcal{L}(\hat{\theta}) \mid \mathbf{y} \right]$$

- Przykładowo kwadratowa funkcja straty $\mathcal{L}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)' Q (\hat{\theta} - \theta)$, gdzie Q jest dodatnio określona jest minimalizowana dla estymatora równego wartości oczekiwanej.
- Estymator $\hat{\theta}$ dla ustalonego \mathbf{y} jest nielosowy. W rezultacie:

$$\begin{aligned} \mathbb{E} \left[\mathcal{L}(\hat{\theta}) \mid \mathbf{y} \right] &= \left[\hat{\theta} - \mathbb{E}(\theta \mid \mathbf{y}) \right]' Q \left[\hat{\theta} - \mathbb{E}(\theta \mid \mathbf{y}) \right] \\ &\quad + \mathbb{E} \left\{ \left[\theta - \mathbb{E}(\theta \mid \mathbf{y}) \right]' Q \left[\theta - \mathbb{E}(\theta \mid \mathbf{y}) \right] \right\} \\ &= A(\hat{\theta}, \mathbf{y}) + B(\mathbf{y}) \end{aligned}$$

- Wartość $B(\mathbf{y})$ nie zależy od postaci $\hat{\boldsymbol{\theta}}$.
- Dla dodatnio określonego Q wartość $\arg \min A(\hat{\boldsymbol{\theta}}, \mathbf{y}) = E(\boldsymbol{\theta} | \mathbf{y})$

Rozkłady *a priori*

- Rozkład *a priori* powinien odbijać wiedzę jaką posiadamy na temat parametrów a pochodzącą spoza próbki.
- Określenie rozkładu *a priori* wymaga
 - wiedzy na temat wartości oczekiwanej parametrów
 - wiedzy na temat stopnia niepewności co do wartości oczekiwanej parametrów
 - opisanie tej niepewności za pomocą rozkładu prawdopodobieństwa
- Bardzo rzadko znamy rzeczywisty rozkład *a priori* parametrów. Zazwyczaj mając wiedzę na temat wartości oczekiwanej i wariancji stosujemy jakiś rozkład, który jest wygodny analitycznie

Informacyjne rozkłady *a priori*.

- Informacyjne rozkłady *a priori* to takie rozkłady, które stanowią odbicie jakiejś wiedzy na temat zjawiska
- Rozkłady *a priori* dobiera się często tak, by uzyskany rozkład *a posteriori* był stosunkowo prosty
- Takie rozkłady *a priori* nazywa się rozkładami naturalnie sprzężonymi
- Obecnie jest możliwe uzyskanie numerycznego przybliżenia rozkładu *a posteriori* dla szerokiej klasy rozkładów *a priori*
- Główne problemy związane z doбором informacyjnego rozkładu *a priori*:

- Estymatory Bayesowskie będą lepsze od estymatorów klasycznych, gdy wartość oczekiwana w parametrów w rozkładzie *a priori* będzie bliska prawdziwej wartości parametru
- W razie przyjęcia błędnego rozkładu *a priori* estymator Bayesowski może być znacznie gorszy od klasycznego
- Im mniejsza założona niepewność wiedzy *a priori*:
 - * tym większe są zyski z jeśli jest prawidłowa
 - * tym możliwe straty jeśli jest błędna
- Co robić gdy zachodzi wyraźna sprzeczność między przyjętym rozkładem *a priori* a danymi?
- Jak uzyskać porównywalność wyników uzyskanych przez badaczy, którzy przyjęli różne rozkłady *a priori*?

Nieinformacyjne rozkłady *a priori*.

- Rozwiązaniem subiektywności rozkładu *a priori* jest zastosowanie takiej jego postaci, która opisuje brak wiedzy na temat parametrów
- Znalezienie takiego rozkładu nie jest proste
- Najpopularniejszą regułą znajdowania nieinformacyjnych rozkładów *a priori* jest reguła Jeffreysa
- Mówi ona, że nieinformacyjny rozkład *a priori* parametru θ powinien spełniać warunek

$$f(\theta) \propto |\mathbf{I}(\theta)|^{\frac{1}{2}}$$

- Uzyskany rozkład jest niewłaściwym rozkładem jednostajnym proporcjonalnym do pierwiastka wyznacznika macierzy informacyjnej Fishera
- Zaletą rozkładów *a priori* znajdowanych zgodnie z regułą Jeffreysa jest to, że rozkłady *a priori* i *a posteriori* są niezmiennicze ze względu na jednoznaczne przekształcenia parametrów
- Czasami estymatory Bayesowskie uzyskane w za pomocą *priorów* nieinformacyjnych są niedopuszczalne (*nonadmissible*)
- Estymator niedopuszczalny, to estymator dla którego istnieje inny estymator o niższej wariancji dla każdego poziomu parametru estymowanego

Bayes hierarchiczny i empiryczny

- Zdarza się, że szczególnie prostą formę prawdopodobieństwa *a posteriori* jeśli przy konstruowaniu rozkładu *a priori* założymy, że część parametrów jest znana
- Rozkład *a priori* dla nieznanymi parametrów ma w takim przypadku postać rozkładu warunkowego $g(\theta_1 | \theta_2)$
- Parametry warunkowego rozkładu *a priori* nazywamy hiperparametrami
- W praktyce parametry te nie są znane
- Jednym z proponowanych rozwiązań problemu tej trudności jest wyestymowanie hiperparametrów na podstawie próby (Bayes empiryczny)

- Ponieważ parametry rozkładu *a priori* są dopasowane do danych, więc nie powinno dojść do konfliktu między rozkładem *a priori* i danymi.
- Problem z tym podejściem polega na tym, że rozkład *a posteriori* uzyskany został przy założeniu, że hiperparametry są znane a tak naprawdę są one estymowane - trudne wyznaczenie rozkładu *a posteriori*
- Zaproponowanym rozwiązaniem jest budowa modelu Bayesowskiego dla hiperparametrów (Bayes hierarchiczny)
- W tym przypadku $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2) g(\boldsymbol{\theta}_2)$
- W przypadku hierarchicznych rozkładów *a priori* rozkłady dla poszczególnych grup parametrów są od siebie zależne.

Testowanie hipotez

- Mamy m konkurencyjnych modeli
- $p(M_i)$ prawdopodobieństwo *a priori* modelu M_i
- Prawdopodobieństwo *a posteriori* modelu M_i z twierdzenia Bayesa

$$p(M_i | \mathbf{y}) = \frac{p(M_i) p(\mathbf{y} | M_i)}{\sum_{k=1}^m p(M_k) p(\mathbf{y} | M_k)}$$

gdzie

$$p(\mathbf{y} | M_i) = p_i(\mathbf{y}) = \int_{\Theta_i} p_i(\mathbf{y} | \boldsymbol{\theta}_i) p_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$$

i $\boldsymbol{\theta}_i$ jest wektorem parametrów dla modelu i .

- Zwykle wybieramy ten model, który ma największe prawdopodobieństwo *a posteriori*
- Można się do tego posłużyć iloczynem szans *a posteriori* (*posterior odds ratio*)

$$\frac{p(M_i | \mathbf{y})}{p(M_j | \mathbf{y})} = \frac{p(M_i) p(\mathbf{y} | M_i)}{p(M_j) p(\mathbf{y} | M_j)}$$

- Dwa najważniejsze przypadki szczególne:

– przyjmujemy, że $p(M_1) = p(M_2) = \dots = p(M_m)$

$$\frac{p(M_i | \mathbf{y})}{p(M_j | \mathbf{y})} = \frac{p(\mathbf{y} | M_i)}{p(\mathbf{y} | M_j)}$$

– przyjmujemy, że mniej prawdopodobne modele o większej ilości parametrów np.

$$p(M_i) \propto 2^{-l_i}$$

- Niewłaściwe rozkłady *a priori* sprawiają problemy przy formułowaniu iloczynu szans.

Analiza Bayesowska w rozkładzie dwumianowym (prior nieinformacyjny)

- Przyjmijmy, że zmienna y_i ma rozkład dwumianowy, to jest $y_i \in \{0, 1\}$ i

$$\Pr(y_i) = \theta^{y_i} (1 - \theta)^{1-y_i}$$

- Zakładamy, że poszczególne zdarzenia są niezależne, więc funkcja wiarygodności

$$L(\theta) = \Pr(\mathbf{y} | \theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} = \theta^k (1 - \theta)^{n-k}$$

gdzie $k = \sum_{i=1}^n y_i$ jest ilością sukcesów a n ilością prób.

- Załóżmy, że nieinformacyjnym rozkładem *a priori* dla tego problemu jest rozkład jednostajny na przedziale $[0, 1]$

$$g(\boldsymbol{\theta}) = \begin{cases} 1 & \boldsymbol{\theta} \in [0, 1] \\ 0 & \text{dla pozostałych} \end{cases}$$

- Łączna funkcja gęstości dla \mathbf{y} i $\boldsymbol{\theta}$ będzie miała postać funkcji

$$f(\mathbf{y}, \boldsymbol{\theta}) = \Pr(\mathbf{y} | \boldsymbol{\theta}) g(\boldsymbol{\theta}) = \boldsymbol{\theta}^k (1 - \boldsymbol{\theta})^{n-k}$$

- Prawdopodobieństwo bezwarunkowe dla \mathbf{y} jest równe

$$\Pr(\mathbf{y}) = \int_0^1 \boldsymbol{\theta}^k (1 - \boldsymbol{\theta})^{n-k} d\boldsymbol{\theta} = B(k + 1, n - k + 1)$$

przy czym skorzystaliśmy ze wzoru, że

$$B(p, q) = \frac{\Gamma(p) \Gamma(q)}{\Gamma(p+q)} = \int_0^1 x^{p-1} (1-x)^{q-1} dx$$

- Z twierdzenia Bayesa wynika, że rozkład *a posteriori* θ

$$\Pr(\theta | \mathbf{y}) = \frac{f(\mathbf{y} | \theta) g(\theta)}{f(\mathbf{y})} = \frac{\theta^k (1-\theta)^{n-k}}{B(k+1, n-k+1)} \propto \theta^k (1-\theta)^{n-k}$$

- Rozkład beta

$$g(\theta) = \begin{cases} B(p, q) \theta^{p-1} (1-\theta)^{q-1} & \theta \in [0, 1] \\ 0 & \text{dla pozostałych} \end{cases}$$

- Otrzymany przez nas rozkład *a posteriori* jest więc rozkładem beta o parametrach $p = k + 1$ i $q = n - k + 1$
- Estymator policzony jako średnia θ z rozkładu *a posteriori* będzie miał postać

$$\begin{aligned}
 \tilde{\theta} &= E(\theta) = \int \theta \frac{\theta^k (1 - \theta)^{n-k}}{B(k+1, n-k+1)} d\theta \\
 &= \frac{1}{B(k+1, n-k+1)} \int \theta^{k+1} (1 - \theta)^{n-k} d\theta \\
 &= \frac{B(k+2, n-k+1)}{B(k+1, n-k+1)} \\
 &= \frac{\Gamma(k+2)\Gamma(n+2)}{\Gamma(k+1)\Gamma(n+3)} = \frac{(k+1)!(n+1)!}{k!(n+2)!} = \frac{k+1}{n+2}
 \end{aligned}$$

przy czym skorzystaliśmy z tego, że dla naturalnego q funkcja $\Gamma(q+1) =$

$q!$

- Czy wyliczony w ten sposób estymator Bayesowski $\tilde{\theta}$ ma sens?
Rozważmy kilka kombinacji n i k

n	0	1	1	98	98
k	0	0	1	0	98
$\tilde{\theta}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{100}$	$\frac{99}{100}$

- Nie mając żadnych informacji na temat k przyjmujemy, że $\tilde{\theta} = \frac{1}{2}$
- Wraz z przybywaniem danych coraz większą wagę przywiązujemy do danych
- Zastanówmy się nad estymatorem Bayesowskim parametru θ równym

dominancie rozkładu. Wygodniej jest maksymalizować $\ln [\text{Pr}(\boldsymbol{\theta} | \mathbf{y})]$

$$\frac{\partial \ln [\text{Pr}(\boldsymbol{\theta} | \mathbf{y})]}{\partial \boldsymbol{\theta}} = \frac{\partial [k \ln \boldsymbol{\theta} + (n - k) \ln (1 - \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = \frac{k}{\boldsymbol{\theta}} - \frac{n - k}{1 - \boldsymbol{\theta}} = 0$$

W rezultacie

$$\hat{\boldsymbol{\theta}} = \frac{k}{n}$$

Taki sam jak *MNW* i *MM*!

Analiza Bayesowska w rozkładzie dwumianowym (prior informacyjny naturalnie sprzężony)

- Analizujemy ten sam problem co poprzednio ale za rozkład *a priori* dla tego przyjmujemy rozkład beta o parametrach $p = k_0 + 1$ i $q = n_0 - k_0 + 1$.
- Z twierdzenia Bayesa wynika, że rozkład *a posteriori* θ

$$\begin{aligned} \Pr(\theta | \mathbf{y}) &\propto f(\mathbf{y} | \theta) g(\theta) \propto \left[\theta^k (1 - \theta)^{n-k} \right] \left[\theta^{k_0} (1 - \theta)^{n_0 - k_0} \right] \\ &= \theta^{k+k_0} (1 - \theta)^{n+n_0-k-k_0} \end{aligned}$$

- Szukając znowu estymatora Bayesowskiego parametru θ równego

dominancie rozkładu znajdujemy

$$\frac{\partial \ln [\Pr (\boldsymbol{\theta} | \mathbf{y})]}{\partial \boldsymbol{\theta}} = \frac{k + k_0}{\boldsymbol{\theta}} - \frac{n + n_0 - k - k_0}{1 - \boldsymbol{\theta}} = 0$$

- Rozwiązując ten układ dla $\hat{\boldsymbol{\theta}}$ otrzymujemy

$$\hat{\boldsymbol{\theta}} = \frac{k + k_0}{n + n_0} = \frac{k_0}{n_0} \underbrace{\frac{n_0}{n + n_0}}_{\text{waga prioru}} + \frac{k}{n} \underbrace{\frac{n}{n + n_0}}_{\text{waga danych}}$$

- Podobnie dla średniej otrzymujemy

$$\tilde{\boldsymbol{\theta}} = \frac{k + k_0 + 1}{n + n_0 + 2} = \frac{k_0 + 1}{n_0 + 2} \underbrace{\frac{n_0 + 2}{n + n_0 + 2}}_{\text{waga prioru}} + \frac{k}{n} \underbrace{\frac{n}{n + n_0 + 2}}_{\text{waga danych}}$$

- Średnia ważona.
- Interpretacja:
 - Prior może chodzić z wcześniejszego badania z użyciem prioru nieinformacyjnego (przypadek omówiony wcześniej) w którym dla n_0 elementowej próby uzyskano k_0 sukcesów
 - Ponieważ prior jest naturalnie sprzężony więc uzyskamy ten sam rozkład *a posteriori* jeśli połączymy obie próby i wyliczymy rozkład na podstawie połączonej próby

Przykład Bayesista gra w trzy karty.

Ortodoksyjny zwolennik podejścia bayesowskiego obserwuje grę w trzy karty. Zauważa, że przy stawce 10 zł, z 8 rund 7 jest wygranych. Zakładając niezależność rund, rozkład dwumianowy p-stwa wygranej oraz nieinformacyjny rozkład *a priori* wylicza, że wartość oczekiwana p-stwa

wygranej (równa parametrowi θ) policzona na podstawie rozkładu *a posteriori* wynosi:

$$E(\theta) = \frac{k_0 + 1}{n_0 + 2} = \frac{8}{10} = \frac{4}{5}$$

Ponieważ oczekiwana wartość wygranej wynosi $\frac{4}{5} * 10 - \frac{1}{5} * 10 = 6$ zł, więc bayesista przystępuje do gry z rozkładem *a priori* równym rozkładowi *a posteriori* zbudowanemu na podstawie obserwacji. Po 20 rundach, z których tylko 2 były wygrane rozkład wartości oczekiwanej wygranej policzona na podstawie rozkładu *a posteriori* wynosi:

$$E(\theta) = \frac{k + k_0 + 1}{n_0 + n + 2} = \frac{2 + 7 + 1}{30} = \frac{1}{3}$$

a wartość oczekiwana wygranej wynosi $\frac{1}{3}10 - \frac{2}{3}10 = -3\frac{1}{3}$.

Oczywiście w tym przypadku można mieć poważne wątpliwości co

prawidłowości rozkładu *a priori*. Prawdopodobnie obserwacje, na podstawie których uformowany został rozkład *a priori* były jedynie "demonstracją".

Analiza Bayesowska w KMRL (rozkłady)

- Rozkład *chi-kwadrat* o r stopniach swobody będziemy oznaczać jako χ_r^2
- k -wymiarowy rozkład normalny o wartości oczekiwanej μ i wariancji Σ określonej dodatnio będziemy oznaczać jako $N_k(\mu, \Sigma)$
- k -wymiarowy rozkład *t-Studenta* o r stopniach swobody, wektorze niecentralności μ i macierzy precyzji $A = \Sigma^{-1}$ określonej dodatnio będziemy oznaczać jako $t_r(\mu, \Sigma)$

Analiza Bayesowska w KMRL (hierarchiczny prior nieinformacyjny)

- Możemy znowu rozpatrywać dwa najważniejsze przypadki: przypadek, kiedy σ^2 jest traktowane jako znane i przypadek, kiedy σ^2 nadejmy rozkład *a priori*.
- Patraktujemy te dwa przypadki łącznie. Rozkłady *a posteriori* β dla przypadku, kiedy σ^2 jest traktowane jako znane otrzymujemy jako rozkłady warunkowe $\beta | \sigma^2$.
- Jak wiemy z rozważań nad estymacją metodą największej wiarygodności *KMRL*

$$f(\sigma) = \left| \frac{\partial^2 \ell^2(\beta, \sigma^2)}{\partial^2 \sigma^2} \right|^{\frac{1}{2}} = \frac{\sqrt{n}}{2\sigma^2} \propto \sigma^{-2}$$

$$f(\beta | \sigma) = \left| \frac{\partial^2 \ell(\beta, \sigma^2)}{\partial \beta \partial \beta'} \right|^{\frac{1}{2}} = \left| \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} \right|^{\frac{1}{2}} \propto \text{const}$$

- Nieinformacyjny rozkład *a priori* (Jeffreys prior)

$$f(\beta, \sigma) = f(\beta | \sigma) f(\sigma) \propto \sigma^{-2}$$

- Policzmy teraz rozkłady *a posteriori* parametrów β i σ
- Funkcja wiarygodności pochodzi z rozkładu normalnego a więc

$$\mathbf{y} | \beta, \sigma^2 \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$$

- Można pokazać, że:
 - rozkład *a posteriori* σ^2 jest

$$\frac{s^2}{\sigma^2} \sim \chi_{n-k}^2$$

- rozkład *a posteriori* β warunkowy względem σ^2

$$\beta | \sigma^2 \sim N \left(\mathbf{b}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right)$$

- rozkład brzegowy *a posteriori* β

$$\beta \sim t_{n-k} \left(\mathbf{b}, s^2 (\mathbf{X}'\mathbf{X})^{-1} \right)$$

gdzie

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - k}$$

- Zauważmy, że uzyskany wynik jest całkowicie zgodny z analizą klasyczną:
 - nieobciążonym punktowym estymatorem wartości β jest \mathbf{b}
 - estymatorem σ^2 jest s^2
 - estymatorem wariancji \mathbf{b} jest $s^2 (\mathbf{X}'\mathbf{X})^{-1}$
 - rozkład \mathbf{b} przy nieznanym σ^2 ma wielowymiarowy rozkład *t-studenta* o $n - k$ stopniach swobody

Analiza Bayesowska w KMRL (hierarchiczny prior informacyjny)

- Możemy znowu rozpatrywać dwa najważniejsze przypadki: przypadek, kiedy σ^2 jest traktowane jako znane i przypadek, kiedy σ^2 mamy rozkład *a priori*.
- Prior naturalnie sprzężony są w tym kontekście rozkłady normalny dla β i odwrócony gamma dla σ^2
- Jeśli w świetle naszej wiedzy *a priori*

$$f(\beta, \sigma^2) = f(\beta | \sigma^2) f(\sigma^2)$$

gdzie rozkład warunkowy *a priori* $\beta | \sigma^2$ spełnia:

$$\beta | \sigma^2 \sim N(\beta_0, \sigma^2 V^{-1})$$

a rozkład *a priori* σ^2 spełnia:

$$\frac{s_0^2}{\sigma^2} \sim \text{gamma}\left(\frac{n_0 - k_0}{2}, \frac{n_0 - k_0}{2}\right)$$

dla naturalnego $n_0 - k_0$ mamy

$$\frac{s_0^2}{\sigma^2} \sim \chi_{n_0 - k_0}^2$$

- Funkcja wiarygodności pochodzi z rozkładu normalnego a więc

$$y | \beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$$

- Można pokazać, że:
 - rozkład *a posteriori* σ^2 spełnia:

$$\frac{\bar{s}^2}{\sigma^2} \sim \text{gamma} \left(\frac{n_0 - k_0}{2}, \frac{n_0 - k_0}{2} \right)$$

dla naturalnego $n_0 - k_0$

$$\frac{\bar{s}^2}{\sigma^2} \sim \chi_{n-k}^2$$

- rozkład *a posteriori* β warunkowy względem σ^2

$$\beta | \sigma^2 \sim N \left(\bar{\mathbf{b}}, \sigma^2 (\mathbf{V} + \mathbf{X}'\mathbf{X})^{-1} \right) \quad (1)$$

- rozkład brzegowy *a posteriori* β

$$\beta \sim t_{n-k} \left(\bar{\mathbf{b}}, \bar{s}^2 (\mathbf{V} + \mathbf{X}'\mathbf{X})^{-1} \right)$$

gdzie

$$\bar{\mathbf{b}} = (\mathbf{X}'\mathbf{X} + \mathbf{V})^{-1} (\mathbf{X}'\mathbf{X}\mathbf{b} + \mathbf{V}\beta_0) \quad (2)$$

$$\bar{s}^2 = \frac{n-2}{\bar{n}-k} s^2 + \frac{n_0-k}{\bar{n}-k} s_0^2$$

$$\bar{n} = n + n_0$$

- Zauważmy, że $\bar{\mathbf{b}}$ można przedstawić jako

$$\bar{\mathbf{b}} = \mathbf{W}\mathbf{b} + \mathbf{W}_0\beta_0$$

$$\mathbf{W} = (\mathbf{X}'\mathbf{X} + \mathbf{V})^{-1} \mathbf{X}'\mathbf{X}$$

$$\mathbf{W}_0 = (\mathbf{X}'\mathbf{X} + \mathbf{V})^{-1} \mathbf{V}$$

$$\mathbf{W} + \mathbf{W}_0 = \mathbf{I}$$

a więc estymator parametru $\bar{\mathbf{b}}$ jest znowu średnią ważoną estymatora

klasycznego MNK i danych z rozkładu *a priori*.

- Podobnie jako średnią ważoną można przedstawić estymator \bar{s}^2 .

Analiza Bayesowska w KMRL (g-prior)

- W przypadku g-prioru zakładamy, że $V = g (X' X)$, gdzie $g > 0$
- Podane w powyżej wzory dla \bar{b} ulegną znacznemu uproszczeniu

$$\bar{b} = (X' X + g (X' X))^{-1} (X' y + g (X' X) \beta_0) = \frac{1}{1 + g} \mathbf{b} + \frac{g}{1 + g} \beta_0$$

- Estymator ten jest średnią ważoną estymatora *MNK* i wartości oczekiwanej *a priori* parametru β

Analiza Bayesowka w KMRL (regresja grzbietowa)

- W przypadku występowania problemu współliniowości zaproponowano estymator

$$b_r = (\mathbf{X}'\mathbf{X} + r\mathbf{D})^{-1} \mathbf{X}'\mathbf{y}$$

- \mathbf{D} jest macierzą diagonalną z elementami diagonalnymi macierzy $\mathbf{X}'\mathbf{X}$ na przekątnej
- r jest pewną stałą większą zera
- Estymator ten uzyskamy jako estymator Bayesowski jeśli przyjmiemy, że

$$\beta_0 = \mathbf{0}$$

$$\mathbf{V} = r\mathbf{D}$$

Problemy numeryczne związane z metodologią Bayesowską

- W wyniku zastosowania twierdzenia Bayesa otrzymujemy zazwyczaj jądro funkcji gęstości dla prawdopodobieństwa *a posteriori*

$$g(\theta) = f(y|\theta) f(\theta)$$

- Zazwyczaj jednak interesuje nas wartości oczekiwane pewnych funkcji n.p. wartość oczekiwana θ albo wartość oczekiwana funkcji straty.
- Policzenie tych wartości oczekiwanych jeśli dysponujemy wyłącznie jądrem rozkładu *a posteriori* nie jest sprawą prostą.

- Jednym z najczęstszych rozwiązań jest symulacja rozkładu *a posteriori* i znajdowanie wartości oczekiwanych interesujących nas funkcji na podstawie symulacji

Metoda Monte Carlo z funkcją ważności

- Z twierdzenia Bayesa wiem, że

$$f(\boldsymbol{\theta} | y) = \frac{g(\boldsymbol{\theta})}{\int_{\Theta} g(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

- Chcemy znaleźć wartość oczekiwaną funkcji $h(\boldsymbol{\theta})$

$$I = \mathbb{E}[h(\boldsymbol{\theta}) | y] = \frac{\int_{\Theta} h(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} g(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

- Powiedzmy, że jesteśmy w stanie znaleźć funkcję gęstości (funkcję ważności) $s(\boldsymbol{\theta})$ bliską funkcji gęstości *a posteriori*. Można wtedy zapisać

ten wzór jako

$$\begin{aligned} E[h(\boldsymbol{\theta})|y] &= \frac{\int_{\Theta} h(\boldsymbol{\theta}) \frac{g(\boldsymbol{\theta})}{s(\boldsymbol{\theta})} s(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} \frac{g(\boldsymbol{\theta})}{s(\boldsymbol{\theta})} s(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\int_{\Theta} h(\boldsymbol{\theta}) w(\boldsymbol{\theta}) s(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} w(\boldsymbol{\theta}) s(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &= \frac{E[h(\boldsymbol{\theta}) w(\boldsymbol{\theta})]}{E[w(\boldsymbol{\theta})]} \end{aligned}$$

gdzie $w(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})}{s(\boldsymbol{\theta})}$ są wagami a wartość oczekiwane jest liczona dla rozkładu $s(\boldsymbol{\theta})$.

- Z prawa wielkich liczb wiemy, że średnia ze zdarzeń niezależnych pochodzących z rozkładu $s(\boldsymbol{\theta})$ dąży do wartości oczekiwanej.
- Jeśli wiemy jak wygenerować realizacje $\boldsymbol{\theta}^i$ z rozkładu $s(\boldsymbol{\theta})$, to I można

policzyć za pomocą następującego wzoru

$$\hat{I}_n = \frac{n^{-1} \sum_{i=1}^n h(\boldsymbol{\theta}^i) w(\boldsymbol{\theta}^i)}{n^{-1} \sum_{i=1}^n w(\boldsymbol{\theta}^i)} \xrightarrow{p} E[h(\boldsymbol{\theta}) | y]$$

- Można oszacować zarówno błąd standardowy tej aproksymacji jak i jej efektywność (równa 1 jeśli $s(\boldsymbol{\theta}) = f(\boldsymbol{\theta} | y)$)
- Funkcją ważności może być np. rozkład jednostajny lub t -Studenta.

Metody numeryczne (losowanie Gibbsa)

- Rozkład *a posteriori* $f(\boldsymbol{\theta} | \mathbf{y})$

1. Przyjmujemy arbitralnie wektor wartości początkowych $\boldsymbol{\theta}^0 = [\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0, \dots, \boldsymbol{\theta}_k^0]$
2. Wygenerowanie jednej realizacji $\boldsymbol{\theta}^q$ w q -tym cyklu losowania Gibbsa polega na losowaniu z warunkowych rozkładów *a posteriori*

$$\boldsymbol{\theta}_1^q = p\left(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{q-1}, \boldsymbol{\theta}_3^{q-1}, \dots, \boldsymbol{\theta}_k^{q-1}\right)$$

$$\boldsymbol{\theta}_2^q = p\left(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^q, \boldsymbol{\theta}_3^{q-1}, \dots, \boldsymbol{\theta}_k^{q-1}\right)$$

...

$$\boldsymbol{\theta}_k^q = p\left(\boldsymbol{\theta}_k | \boldsymbol{\theta}_1^q, \boldsymbol{\theta}_2^{q-1}, \dots, \boldsymbol{\theta}_{k-1}^{q-1}\right)$$

- Cykle powtarzane są aż do osiągnięcia zbieżności
- Rozkład *a posteriori* $f(\theta | y)$ jest punktem stacjonarnym tego algorytmu
- Po osiągnięciu tego punktu kolejne realizacje pochodzą z rozkładu *a posteriori*.
- Można pokazać, że dla ogólnych warunków algorytm ten jest zbieżny.
- Zazwyczaj szereg pierwszych cykli się wyrzuca - w tym czasie algorytm zbiega do θ
- Algorytm ten jest szczególnie użyteczny dla modeli, w których warunkowe rozkłady parametrów są proste
- Umożliwia on analizę problemów o wielu wymiarach

Reakcja na pożar (opis problemu)

- Na Bronxie zainstalowane są alarmy przeciwpożarowe, które są bezpośrednio podłączone do straży pożarnej
- W przypadku zarejestrowania alarmu należy natychmiast podjąć decyzję, czy wysłać jeden wóz straży pożarnej, czy też większą ilość jednostek
- Część sygnałów wysyłanych przez alarmy okazuje się fałszywa.
- W przypadku prawdziwego pożaru jedna jednostka może okazać się bezradna
- Wysłania większej ilości jednostek do fałszywego alarmu może uniemożliwić właściwą reakcję na prawdziwy pożar

- Wszystkie alarmy na Bronxie zgrupowane są w jednym z 216 okręgach.
- Oznaczmy, przez m ilość alarmów zainstalowanych w danym okręgu
- Oznaczmy jako n_i ilość pożarów sygnalizowanych przez i -ty alarm a przez k_i ilość rzeczywistych pożarów zasygnalizowanych przez ten alarm
- $n = \sum n_i$ ogólna ilość sygnałów pochodzących z danego okręgu, $k = \sum k_i$ ilość zasygnalizowanych prawdziwych pożarów w danym okręgu.
- Tradycyjna procedura reakcji na pożar wyglądała następująco:
 - p_0 jest pewną arbitralnie ustaloną liczbą
 - jeśli $\tilde{p}_i = \frac{k_i}{n_i} > p_0$ (dany alarm wysyłał często sygnały o prawdziwych pożarach), wysyłamy większą liczbę jednostek
 - jeśli $\tilde{p}_i = \frac{k_i}{n_i} < p_0$ (dany alarm wysyłał często fałszywe sygnały), wysyłamy jedną jednostkę

- jeśli $n_i = 0$ (dany alarm po raz pierwszy wysyła sygnał) zastępujemy $\frac{k_i}{n_i}$ przez $\frac{k}{n}$ i porównujemy z p_0

Model próbkowy (reakcja na pożar)

- Załóżmy, że k_i pochodzi z rozkładu dwumianowego $Bi(n_i, p_i)$, gdzie p_i oznacza prawdopodobieństwo wysłania przez i -ty alarm prawdziwego sygnału.
- Udział prawdziwych sygnałów dla alarmu i jest równa $y_i = \frac{k_i}{n_i}$.
- Wariancja y_i jest wynosi $\sigma_i^2 = n_i^{-1} (1 - p_i) p_i$.
- Graniczny rozkład y_i można uzyskać z *CTW*:

$$y_i \stackrel{a}{\sim} N(p_i, \sigma_i^2)$$

- Ponieważ ilości sygnałów n_i oraz p_i pochodzące różnią się dla poszczególnych alarmów więc różna jest też ich wariancja
- Obserwowane y_i pochodzą więc z modelu:

$$y_i | p_i = p_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_i^2)$$

- Można to zapisać jako

$$\mathbf{y} | \mathbf{p} = \mathbf{I}\mathbf{p} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$$

gdzie $\mathbf{p} = [p_1, \dots, p_m]'$, $\boldsymbol{\Sigma} = \text{diag} [n_1^{-1} (1 - p_1) p_1, \dots, n_k^{-1} (1 - p_m) p_m]$

- Nieznanym parametrem jest p_i .

Rozkład *a priori* (reakcja na pożar)

- Nasz rozkład *a priori* opisuje rozkład częstości prawdziwych alarmów w danym sąsiedztwie
- Rozkład *a priori* opisuje na ile te są do siebie podobne a na ile różnią się od generalnej średniej π
- Rozkład *a priori* jest następujący

$$p_i \sim N(\pi, \sigma_\pi^2)$$

lub

$$\mathbf{p} \sim N(\pi \mathbf{I}, \sigma_\pi^2 \mathbf{I})$$

$$p = \pi I + \eta$$
$$\eta \sim N(0, \sigma_{\pi}^2 I)$$

- Parametry rozkładu *a priori* są nieznane ale można je wyestymować (Bayes empiryczny)

Rozkład a posteriori (reakcja na pożar)

- Możemy zastosować wcześniej wyprowadzony wzór pod warunkiem, że usuniemy heteroskedastyczność

$$\mathbf{y}^* | \mathbf{p} \sim N \left(\Sigma^{-\frac{1}{2}} \mathbf{p}, \mathbf{I} \right)$$

gdzie $\Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} = \Sigma^{-1}$, $\mathbf{y}^* = \Sigma^{-\frac{1}{2}} \mathbf{y}$

- Teraz możemy zastosować standardowe twierdzenie, jeśli $\mathbf{X} = \Sigma^{-\frac{1}{2}}$, \mathbf{y}^* , $\mathbf{V} = \frac{1}{\sigma_\pi^2} \mathbf{I}$, $\beta_0 = \pi \mathbf{l}$, $\mathbf{l} = [1, \dots, 1]'$

- Rozkład a posteriori parametru p będzie dany wzorem

$$\begin{aligned}\bar{p} &= (\mathbf{X}'\mathbf{X} + \mathbf{V})^{-1} (\mathbf{X}'\mathbf{X}\mathbf{b} + \mathbf{V}\beta_0) \\ &= \left(\Sigma^{-1} + \frac{1}{\sigma_\pi^2} \mathbf{I} \right)^{-1} \left(\Sigma^{-1} \mathbf{y} + \frac{1}{\sigma_\pi^2} l\pi \right)\end{aligned}$$

- Ponieważ wszystkie macierze są diagonalne więc

$$\begin{aligned}\bar{p}_i &= \left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_\pi^2} \right)^{-1} \left(\frac{y_i}{\sigma_i^2} + \frac{\pi}{\sigma_\pi^2} \right) \\ &= \frac{\sigma_i^2}{\sigma_i^2 + \sigma_\pi^2} \pi + \frac{\sigma_\pi^2}{\sigma_i^2 + \sigma_\pi^2} y_i\end{aligned}$$

Szacowanie parametrów rozkładu a priori (reakcja na pożar)

- Po to by znaleźć wielkość tego estymatora musimy znać parametry rozkładu *a priori* π i σ_{π}^2 oraz σ_i^2
- Estymator $\sigma_i^2 = n_i^{-1} (1 - y_i) y_i$ postaci $\hat{\sigma}_i^2 = n_i^{-1} (1 - y_i) y_i$ okazuje się zbyt niedokładny (za mało obserwacji dla poszczególnych alarmów)
- Możemy więc oszacować wariancję σ_i^2 na podstawie częstości dla całego okręgu $\bar{y} = \frac{k}{n}$ tak, że $\tilde{\sigma}_i^2 = n_i^{-1} \bar{y} (1 - \bar{y})$

- Parametry rozkładu *a priori* można oszacować zauważając, że

$$\mathbf{y} = \mathbf{I} (\boldsymbol{\pi} \mathbf{I} + \boldsymbol{\eta}) + \boldsymbol{\varepsilon} = \boldsymbol{\pi} \mathbf{I} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$y_i = \pi + \eta_i + \varepsilon_i = \pi + u_i$$

$$\mathbf{u} \sim N(\boldsymbol{\Sigma} + \sigma_{\pi}^2 \mathbf{I})$$

- Model dla y_i jest modelem liniowym ale zawierającym heteroskedastyczność
- Przy znanym σ_{π}^2 i σ_i^2 można go oszacować *UMNK* (zakładamy, że $\boldsymbol{\Sigma}$ jest znane)

$$\begin{aligned} \hat{\boldsymbol{\pi}} &= \left[\mathbf{l}' (\sigma_{\pi}^2 \mathbf{I} + \boldsymbol{\Sigma})^{-1} \mathbf{l} \right]^{-1} \mathbf{l} (\sigma_{\pi}^2 \mathbf{I} + \boldsymbol{\Sigma})^{-1} \mathbf{y} \\ &= \frac{\sum_{i=1}^m (\sigma_{\pi}^2 + \sigma_i^2)^{-1} y_i}{\sum_{i=1}^m (\sigma_{\pi}^2 + \sigma_i^2)^{-1}} \end{aligned}$$

- Za σ_i^2 można postawić $\tilde{\sigma}_i^2$
- Ponieważ, z własności *UMNK* wynika, że

$$E(e'\Omega^{-1}e) = m - 1$$

więc

$$\sum_{i=1}^m E \frac{(y_i - \hat{\pi})^2}{\sigma_{\pi}^2 + \sigma_i^2} = m - 1$$

- Sensownym oszacowaniem $\hat{\sigma}_{\pi}^2$ byłoby takie σ_{π}^2 , które spełniałoby

$$\sum_{i=1}^m \frac{(y_i - \hat{\pi})^2}{\sigma_{\pi}^2 + \sigma_i^2} = m - 1$$

- Możemy iterować oszacowania $UMNK$ i oszacowania $\hat{\sigma}_{\pi}^2$ aż do uzyskania zbieżności.
- Inną możliwością jest znalezienie estymatora MNW dla parametrów w modelu liniowym z heteroskedastycznością $y_i = \pi + u_i$.

Wyniki (reakcja na pożar)

- Wynik oszacowań

$$\bar{p}_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \hat{\sigma}_\pi^2} \hat{\pi} + \frac{\hat{\sigma}_\pi^2}{\hat{\sigma}_i^2 + \hat{\sigma}_\pi^2} y_i$$

- Uzyskany estymator jest średnią ważoną ogólnej tendencji $\hat{\pi}$ i częstości zaobserwowanych dla danego alarmu y_i .
- W zależności od ilości sygnałów pochodzących z danego alarmu większą wagę ma $\hat{\pi}$ lub y_i .
- Wagi zostały dobrane optymalnie - w sensie bayesowskim.

- Przeprowadzone badanie empiryczne wykazało, że prosta metoda nieprawidłowo zaklasyfikowała 183 fałszywe alarmy jako pożarypodczas, zastosowanie estymatora \bar{p}_i dało nieprawidłową klasyfikację fałszywych alarmów w 130 przypadkach.

Korekta danych spisowych (opis problemu)

- Erickson and Kadane (1985)
- 100mld\$ funduszy federalnych rocznie rozdzielanych między stany
- Rozdzielanie dokonywane na podstawie spisu
- Około 0.25% populacji nie jest uwzględniona w spisie
- Wiadomo, że szczególnie niektóre typy osób szczególnie łatwo umykają spisowi (np. mniejszości narodowe, nielegalni emigranci)
- Stan Nowy Jork ma szczególnie zamieszkuje szczególnie dużo tego typu osób

- W związku z tym było w interesie tego stanu, by uwzględnić ten fakt w oszacowaniach wielkości populacji.
- Dysponujemy oszacowaniem procentu niepoliczonej populacji (PEP - postenumeration program)
- Oszacowanie to jest dokonywane na podstawie próby losowej (sprawdzamy, czy zostały policzone - próba P) oraz na podstawie weryfikacji danych (sprawdzamy, czy poprawne są dane dotyczące osób policzonych - próba E)
- Oszacowania procentu niepoliczonej populacji nie zostały użyte do poprawy oszacowań danych spisowych ponieważ charakteryzowały się zbyt wysoką wariancją.
- Problem: znaleźć taki estymator wielkości niepoliczonej populacji, który przy niskiej wariancji uwzględniałby fakt, że wielkość niedoszacowania

zależy od charakterystyk liczonej populacji.

- Prawdziwy procent niepoliczonej populacji oznaczmy jako β_i , gdzie i -jest numerem okręgu statystycznego.
- Pochodzące z biura spisowego oszacowania β_i oznaczmy jako y_i .
- Biuro spisowe dostarcza także oszacowań wielkości wariancji dla β_i oznaczanych jako K_{ii} .

Model próbkowy (korekta danych spisowych)

- Załóżmy, że rozkład procentu osób niepoliczonych pochodzi z modelu:

$$\begin{aligned}y_i | \beta_i &= \beta_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, K_{ii}) \\ y_i | \beta_i &\sim N(\beta_i, K_{ii})\end{aligned}$$

- Można to zapisać w postaci wektorowej jako

$$\begin{aligned}\mathbf{y} | \boldsymbol{\beta} &= \mathbf{I}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(0, \mathbf{K}) \\ \mathbf{y} | \boldsymbol{\beta} &\sim N(\mathbf{I}\boldsymbol{\beta}, \mathbf{K})\end{aligned}$$

gdzie K jest macierzą diagonalną.

Rozkład a priori (korekta danych spisowych)

- Wiemy, że procent niepoliczonej populacji zależy od charakterystyk osób zamieszkujących dany teren.
- Załóżmy więc, że rozkład *a priori* parametru β_i jest dany:

$$\beta_i = z_i \alpha + \eta_i$$

$$\eta_i \sim N(0, \sigma_\eta^2)$$

$$\beta_i \sim N(z_i \alpha, \sigma_\eta^2)$$

- z_i zawierało: procent mniejszości, wskaźnik przestępczości i procent populacji zbadanej bezpośrednio (a nie za pośrednictwem poczty)

- W zapisie macierzowym mamy:

$$\begin{aligned}\beta | \alpha &= Z\alpha + \eta \\ \eta &\sim N(\mathbf{0}, \sigma_\eta^2 \mathbf{I})\end{aligned}$$

- Rozkład *a priori* β przy znajomości α jest w związku z tym

$$\beta | \alpha = N\left(Z\alpha, \sigma_\eta^2 \mathbf{I}\right)$$

Rozkład a posteriori (korekta danych spisowych)

- W tej wersji model odbiega nieco od modelu *KMRL* z informacyjnym rozkładem *a priori*, ponieważ model charakteryzuje się heteroskedastycznością reszt $K \neq \sigma^2 I$
- Wiemy jednak, że model z heteroskedastycznością można przekształcić do modelu bez heteroskedastyczności
- Mnożąc lewostronnie obie strony równania dla y przez $K^{-\frac{1}{2}}$ otrzymujemy:

$$\begin{aligned}y^* | \beta &= K^{-\frac{1}{2}} \beta + \varepsilon^* \\ \varepsilon^* &\sim N(0, I) \\ y^* | \beta &\sim N\left(K^{-\frac{1}{2}} \beta, I\right)\end{aligned}$$

gdzie $y^* = K^{-\frac{1}{2}} y$

- Jeśli zastosujemy twierdzenia z poprzedniej części przy założeniu, że $\beta_0 = Z\alpha$ i $\sigma^2 = 1$ są znane, dla $X = K^{-\frac{1}{2}}$, $V^{-1} = \sigma_\eta^2 I$, to otrzymamy

$$\beta | \alpha \sim N(\bar{b}, P^{-1})$$

- Wartość oczekiwaną liczymy ze wzoru (2)

$$\bar{b} = P^{-1} \left(K^{-1} y + \frac{1}{\sigma_\eta^2} Z\alpha \right)$$

a wariancję ze wzoru (1)

$$P = K^{-1} + \frac{1}{\sigma_\eta^2} I$$

Szacowanie parametrów rozkładu a priori (korekta danych spisowych)

- Parametr jest α nieznany. Zauważmy, że

$$\begin{aligned}y|\alpha &= Z\alpha + \eta + \varepsilon \\y|\alpha &\sim N(Z\alpha, \sigma_\eta^2 I + K)\end{aligned}$$

- Model ten jest modelem liniowym z heteroskedastycznością. Przy znanym σ_η^2 można oszacować parametr α za pomocą *UMNK*:

$$\hat{\alpha} = (Z'\Sigma^{-1}Z)^{-1} Z'\Sigma^{-1}y$$

$$\Sigma_{\alpha} = (\mathbf{Z}'\Sigma^{-1}\mathbf{Z})^{-1}$$

dla $\Sigma = \sigma_{\eta}^2\mathbf{I} + \mathbf{K}$.

- Wielkość parametru σ_{η}^2 można uzyskać za pomocą *MNW*.
- Uznajmy, że rozkład α ma rozkład z wartością oczekiwaną $\hat{\alpha}$ i wariancją Σ_{α}

$$\alpha \sim N(\hat{\alpha}, \Sigma_{\alpha})$$

Korekta wariancji estymatora (korekta danych spisowych)

- Przy liczeniu estymatora \bar{b} podstawiamy za α jej wartość oczekiwaną $\hat{\alpha}$.
- Wielkość wariancji estymatora \bar{b} została policzona przy założeniu, że α jest znana a tak naprawdę jest szacowana.
- Wariancji rozkładu *a posteriori* β nie uwzględnia więc faktu, że α jest losowa
- Problem ten można rozwiązać używając wzoru na dokompozycję wariancji:

$$\text{Var}(\beta) = E[\text{Var}(\beta | \alpha)] + \text{Var}[E(\beta | \alpha)]$$

- Wartość $\text{Var}(\beta | \alpha)$ jest równa wartości wariancji rozkładu a posteriori przy założeniu znajomości σ_η^2 :

$$E[\text{Var}(\beta | \alpha)] = P^{-1}$$

- Wariancja $\text{Var}[E(\beta | \hat{\alpha})]$ jest równa

$$\begin{aligned} \text{Var}[E(\beta | \alpha)] &= \text{Var}[\bar{b}] = \frac{1}{\sigma_\eta^4} P^{-1} \text{Var}[Z\alpha] P^{-1} \\ &= \frac{1}{\sigma_\eta^4} P^{-1} Z \Sigma_\alpha Z' P^{-1} \end{aligned}$$

$$\text{Var}(\beta) = P^{-1} + \frac{1}{\sigma_\eta^4} P^{-1} Z \Sigma_\alpha Z' P^{-1}$$

Wyniki (korekta danych spisowych)

- Wynik wyniku dokładnego badania w sądzie uzyskane w ten sposób estymatory zostały uznane za niewystarczającą podstawę do zwiększenia funduszy federalnych przyznawanych stanowi Nowy Jork
- Powody:
 - oszacowanie $\hat{\alpha}$ nie jest zgodne - błąd losowy ε jest zależny od zmiennych z_i - zmienne, które wpływają na niedoszacowanie ilości osób na danym terenie wpływają także na błąd oszacowania tego niedoszacowania (równoczesność)
 - błędy standardowe oszacowania β policzone w wyjaśniony powyżej sposób są najprawdopodobniej zaniżone (σ_{η}^2, K_{ii} traktowane jako znane a tak naprawdę szacowane, niepewność co do formy równania regresji)

- nie bardzo wiadomo dlaczego błędy ε_i się nie znoszą przy agregacji
- akceptowalność polityczna: średnio rzecz biorąc oszacowanie za pomocą estymatora bayesowskiego może być lepsze niż y_i , jednak dla części stanów oznaczałoby to stratę a dla innych zysk - stany, które ponosiłyby na takich kerekctach straty z pewnością protestowałyby przeciwko korygowaniu oszacowań.

Przykład empiryczny - prior z Minnesoty (szeregi czasowe)

- Prior dla modeli VAR zaproponowany przez Littermana(1986)
- **Prior dla VAR stacjonarnego**
- Zakładamy, że wartości oczekiwane z rozkładu *a priori* wszystkich współczynników są równe zero
- Wariancje i kowariancje współczynników maleją wraz z wilkością opóźnienia (zakładamy, że mamy wiedzę *a priori*, że odległe opóźnienia są nieistotne)
- Wariancje współczynnika przy opóźnieniu i tej zmiennej w j tym równaniu jest proporcjonalna do ilorazu wariancji błędów losowych w tych równaniach.

- Załóżmy, że wektor α^* zawiera opóźnienia dla wszystkich równań w modelu VAR

$$\alpha^* \sim N(0, V^{-1})$$

- Zakładamy, że macierz V^{-1} jest diagonalna z elementami:

$$v_{ij} = \begin{cases} \left(\frac{\lambda}{l}\right)^2 & \text{dla } i = j \\ \left(\theta \frac{\lambda \sigma_i}{l \sigma_j}\right)^2 & \text{dla } i \neq j \end{cases}$$

gdzie

- i jest numerem równania
- j jest numerem zmiennej opóźnionej
- l jest opóźnieniem
- σ_i^2, σ_j^2 są wariancjami błędów losowych w odpowiednich równaniach

- l jest numerem opóźnienia, λ opisuje zaufanie do priora, że współczynniki przy opóźnieniach są równe zero
- θ opisuje zaufanie do priora, że poszczególne zmienne wzajemnie od siebie nie zależą

- Oszacowania *MNK* otrzymujemy dla $\lambda = \infty$ i $\theta = 1$

- Przykład: *VAR*(2)

$$y_{1t} = \alpha_{11,1}y_{1,t-1} + \alpha_{12,1}y_{2,t-1} + \alpha_{11,2}y_{1,t-2} + \alpha_{12,2}y_{2,t-2} + u_{1t}$$

$$y_{2t} = \alpha_{21,1}y_{1,t-1} + \alpha_{22,1}y_{2,t-1} + \alpha_{21,2}y_{1,t-2} + \alpha_{22,2}y_{2,t-2} + u_{2t}$$

$$\alpha^* = [\alpha_{11,1}, \alpha_{21,1}, \alpha_{12,1}, \alpha_{22,1}, \alpha_{11,2}, \alpha_{21,2}, \alpha_{12,2}, \alpha_{22,2}]'$$

- Czasami estymuje się poszczególne równia za pomocą estymatora Bayesowskiego, choć powinno się estymować cały system (łatwiejsze analitycznie)
- Zauważmy, że założyliśmy, że w modelu nie ma stałej. W praktyce albo estymujemy model na odchyleniach od średnich albo formułujemy prior dla stałych (może być nieinformacyjny)
- **Prior dla VAR niestacjonarnego**
- Identyczna macierz V ale inny jest wektor oczekiwanych wartości z rozkładu *a priori* dla α^*
- Zakładamy, że wartości oczekiwane *a priori* dla tych parametrów, które związane są z pierwszymi opóźnieniami zmiennej i w równaniu i są równe 1 (błądzenie przypadkowe) a pozostałe wartości oczekiwane parametrów są równe zero.

- Przykład: aprioryczny $VAR(2)$ niestacjonarny

$$\begin{aligned}
 y_{1t} &= \frac{1}{(\lambda)} y_{1,t-1} + \frac{0}{\theta \lambda \frac{\sigma_1}{\sigma_2}} y_{2,t-1} + \frac{0}{\left(\frac{\lambda}{2}\right)} y_{1,t-2} + \frac{0}{\theta \frac{\lambda}{2} \frac{\sigma_1}{\sigma_2}} y_{2,t-2} + u_{1t} \\
 y_{2t} &= \frac{0}{\theta \lambda \frac{\sigma_2}{\sigma_1}} y_{1,t-1} + \frac{1}{(\lambda)} y_{2,t-1} + \frac{0}{\theta \frac{\lambda}{2} \frac{\sigma_2}{\sigma_1}} y_{1,t-2} + \frac{0}{\left(\frac{\lambda}{2}\right)} y_{2,t-2} + u_{2t}
 \end{aligned}$$