Jiná pravděpodobnostní rozložení a typy modelů

- Vícerozměrné normální rozložení $y = (y_1, \ldots, y_k)$
- $\phi(y) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu)}$
  - $|.|$ je determinant, pokud počet složek y označíme k je $|2\pi\Sigma| = (2\pi)^k|\Sigma|$.
- pokud Σ není invertibilní, má závislé sloupce, tj. proměnné y_i jsou lineárně závislé.
  - vezmeme hodnost matice Σ (označ. ℓ), pak existuje matice A a vektor ν, že
  - $y = Az + \nu$ pro ℓ dimenzionální nové souřadnice z
  - my: přejdeme k novým souřadnicím, pro které má Σ plnou hodnost.

Marginalizace

- Máme: $\frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu)}$
- chceme: rozložení na veličinách $\{y_3, y_5, y_7\} \subset \{y_1, \ldots, y_n\}$
- marginální rozložení dostaneme, když 'nezajímavé' dimenze prostě vynecháme, tj.
- $\mu_{3,5,7} = (\mu_3, \mu_5, \mu_7)$ a $\Sigma_{3,5,7} = \begin{bmatrix} \Sigma_{33} & \Sigma_{35} & \Sigma_{37} \\ \Sigma_{53} & \Sigma_{55} & \Sigma_{57} \\ \Sigma_{73} & \Sigma_{75} & \Sigma_{77} \end{bmatrix}$
- tj. $\frac{1}{\sqrt{|2\pi\Sigma_{3,5,7}|}} e^{-\frac{1}{2}(y_{3,5,7}-\mu_{3,5,7})\Sigma_{3,5,7}^{-1}(y_{3,5,7}-\mu_{3,5,7})}$

Podmiňování

- Zajímá nás $\pi(A|B)$, kde
  - $A \subset \{y_1, \dots, y_n\}$ o q prvcích,
  - zbytek $B = \{y_1, \dots, y_n\} \setminus A$ má $(n-q)$ prvků.
- přeházíme řádky (a sloupce), aby A byly u sebe, pak všechna B; dostaneme: $y = \begin{bmatrix} y_A \\ y_B \end{bmatrix}$ (1 sloupec), $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$ (1 sloupec), $\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$ o dimenzích $\begin{bmatrix} q \times q & q \times (n-q) \\ (n-q) \times q & (n-q) \times (n-q) \end{bmatrix}$.
- Parametry podmíněného normálního rozložení $\pi(A|B=b) = N(\mu_{A|B=b}, \Sigma_{A|B=b})$ jsou:

$$\begin{aligned} \mu_{A|B=b} &= \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b - \mu_B) \\ \Sigma_{A|B=b} &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \end{aligned}$$

Přidání nové složky y_{n+1}

- Pokud chceme do modelu přidat proměnnou y_{n+1}, stále předpokládáme vícerozměrné gaussovské
- přidáme lineární funkci průměru (střední hodnoty):
  - $\mu_{y_{n+1}|y_{1,\ldots,n}} = c_0 + \sum_{i=1,\ldots,n} c_i y_i$
  - rozptyl $\sigma^2_{(n+1)}$
  - případně kovariance k ostatním veličinám.

Se gaussovsky rozloženými veličinami můžeme dělat 'totéž' (grafický model, rozklad, propagace) jako v diskrétním.

Změníme parametrizaci rozložení

- Gaussovské rozložení $\phi(y) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu)}$
- definujeme:
  - koncentrační matice $K = \Sigma^{-1}$
  - $h = K\mu$
  - $a = -\frac{d}{2}\log(2\pi) + \frac{1}{2} log(|K|) - \frac{1}{2}\mu^T K\mu$
- pak vícerozměrnou gaussovskou hustotu přepíšeme

$$
\begin{aligned}
\phi(y) &= (2\pi)^{-\frac{d}{2}} |K|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}(y-\mu)K(y-\mu) \right\} \\
&= (2\pi)^{-\frac{d}{2}} |K|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\mu^T K\mu + h^T y - \frac{1}{2} y^T K y \right\} \\
&= \exp\left\{ a + h^T y - \frac{1}{2} y^T K y \right\} \\
&= \exp\left\{ a + \Sigma_u h_u y_u - \frac{1}{2} \Sigma_{u,v} K_{u,v} y_u y_v \right\}
\end{aligned}
$$

Rozložitelnost

- $\phi(y) = \exp\left\{a + \Sigma_{u\in U} h_u y_u - \frac{1}{2}\Sigma_{u,v} K_{u,v} y_u y_v\right\}$
- Mějme množiny vrcholů A, B separované množinou C. Pak máme $\forall u \in A, v \in B\ k_{uv} = 0$.
- Podíváme se na vzorec a sumy rozdělíme: $\phi(y) =$
$$\exp\left\{\begin{array}{c} a + \Sigma_{u\in A\cup C} h_u y_u + \Sigma_{v\in B\cup C} h_v y_v - \Sigma_{v\in C} h_v y_v \\ -\frac{1}{2}(\Sigma_{u,v\in A\cup C} K_{u,v} y_u y_v + \Sigma_{u,v\in B\cup C} K_{u,v} y_u y_v - \Sigma_{u,v\in C} K_{u,v} y_u y_v) \end{array}\right\}$$

$$\begin{array}{cccc} & A & C & B \\ A & K_{AA} & K_{AC} & \\ C & K_{AC} & K_{CC} & K_{CB} \\ B & & K_{BC} & K_{BB} \end{array}$$

- Tedy $\phi(y) = g(A, C)h(C, B)$

Diskrétní veličiny

- Mějme tři veličiny, A, B, C, libovolnou pravděpodobnostní distribuci $p(A, B, C)$ (velkým proměnná, malým konkrétní hodnota) můžeme zapsat pomocí (neznámých) parametrů u_*^*, tzv. *interaction terms*:
  - $log(p(a, b, c)) = u + u_a^A + u_b^B + u_c^C + u_{ab}^{AB} + u_{ac}^{AC} + u_{bc}^{BC} + u_{abc}^{ABC}$ (u je víc než je třeba, protože ...)
  - aby prošlo i pro nulu, napíšeme v exponenciálním tvaru $\tilde{u} = \exp u$:
  - $p(a, b, c) = \tilde{u} \cdot \tilde{u}_a^A \cdot \tilde{u}_b^B \cdot \tilde{u}_c^C \cdot \tilde{u}_{ab}^{AB} \cdot \tilde{u}_{ac}^{AC} \cdot \tilde{u}_{bc}^{BC} \cdot \tilde{u}_{abc}^{ABC}$
- *l*og–lineární modely některé interakce z u_*^* dají napevno rovné nule
  - s úmluvou, že pokud dám rovno nule u_{AB}, tak i všechny vyšší interakce u^{ABC} atd. musí být rovny nule.
  - pro specifikaci hierarchického modelu stačí zadat maximální interakce, tzv. generátory.
  - např. $\{a, b\}$ a $\{a, c\}$ generují model:
  - $log(p(a, b, c)) = u + u_a^A + u_b^B + u_c^C + u_{ab}^{AB} + u_{ac}^{AC}$ (jedno–složkové interakce jsou tam proto, že jsou podmnožinou více–složkového generátoru).

Faktorizace

- $log(p(a,b,c)) = u + u_a^A + u_b^B + u_c^C + u_{ab}^{AB} + u_{ac}^{AC}$
- $p(a,b,c) = \tilde{u} \cdot \tilde{u}_a^A \cdot \tilde{u}_b^B \cdot \tilde{u}_c^C \cdot \tilde{u}_{ab}^{AB} \cdot \tilde{u}_{ac}^{AC}$
- model výše lze rozepsat na součin:
- $p(a,b,c) = (\tilde{u} \cdot \tilde{u}_a^A \cdot \tilde{u}_b^B \cdot \tilde{u}_{ab}^{AB}) \cdot (\tilde{u}_c^C \cdot \tilde{u}_{ac}^{AC})$
- Vidíte tam podmíněnou nezávislost $B \perp\!\!\!\perp C | A$?
- obecně: dva faktory jsou nezávislé dáno zbytek právě když interakce této dvojice je nuceně 0 (tj. nevyskytují se spolu v žádné interakci)
- graf: hrana právě když je interakce dvojice povolena (ne–nutně–nula).
- grafické modely: log–lineární modely s generátory definovanými klikami grafu.
- Najděte log–lineární model, který není grafický.
- Obecně je třeba log–lineární modely učit iterativně i pro plná data.
- Pro rozložitelné modely existuje přímý výpočet. Rozložitelný model: lze rozebrat tak, že vždy eliminujeme simpliciální uzel, existuje prefektní eliminační posloupnost, graf je triangulovaný (kružnice délky 4 a víc mají tětivu).
- Najděte grafický model, který není rozložitelný.

Mixed interaction models

- A teď diskrétní a gaussovské dohoromady.
- *podmíněná gaussovská hustota* pro $x(i, y)$, i seznam diskrétních proměnných, y seznam spojitých proměnných
- $f(i, y) = p(i)(2\pi)^{-\frac{q}{2}}|\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(y - \mu(i))\Sigma^{-1}(y - \mu(i)))$
- parametry $p(i), \mu(i), i \in \mathcal{I}, \Sigma$ se nazývají *m*oment parameters
- když rovnici převedeme do tvaru 'exponenciálních rodin'

$$
\begin{aligned}
f(i, y) &= \exp\left\{g(i) + h(i)^T y - \frac{1}{2} y^T K y\right\} \\
&= \exp\left\{g(i) + \Sigma_u h_u(i) y_u - \frac{1}{2} \Sigma_{u,v} K_{u,v} y_u y_v\right\}
\end{aligned}
$$

- parametry $g(i), h(i), i \in \mathcal{I}, K$ se nazývají *k*anonické parametry

Mixed interaction models

- Marginála už nemusí být podmíněná gaussovská hustota (je to směs gaussovských hustot)
- ještě pořád se s tím dá počítat
  - za předpokladu, že všechny diskrétní veličiny jsou 'u sebe'
  - tj. pokud existuje cesta mezi diskrétními veličinami, existuje i jejich neobsahující spojitý uzel
  - pak lze počítat likelihood a učit model.

Exponenciální rodiny

- distribuční funkce ve tvaru

$$p(x) = h(x)e^{\theta^T T(x) - A(\theta)}$$

- nebo obecněji ve tvaru

$$p(x) = h(x)e^{\eta(\theta)^T T(x) - A(\eta(\theta))}$$

- θ vektor parametrů
- η natural parameters
- $T(x)$ sufficient statistics
- $A(\theta)$ – normalizační funkce
- $h(x)$
- klíčová myšlenka: θ a x jsou společně jen v $e^{\theta^T T(x)}$

Příklady rozložení z exponenciálních rodin

(tj. dají se 'reparametrizovat')

Gaussian	$\phi(y) = \frac{1}{\sqrt{\lvert 2\pi\Sigma\rvert}} e^{-\frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu)}$	$x \in R$
Bernoulli	$p(x) = \alpha^x(1-\alpha)^{1-x}$	$x \in \{0,1\}$
Binomial	$p(x) = \binom{n}{x} \alpha^x(1-\alpha)^{n-x}$	$x \in \{0,1,\ldots,n\}$
Multinomial	$p(x) = \frac{n!}{x_1!x_2!\ldots x_n!}\Pi_{i=1}^n \alpha_i^{x_i}$	$x \in \{0,1,\ldots,n\}, \Sigma_i x_i = n$
Exponential	$p(x) = \lambda e^{-\lambda x}$	$x \in R^+$
Poisson	$p(x) = \frac{e^{-\lambda}}{x!}\lambda^x$	$x \in \{0,1,\ldots\}$
Dirichlet	$p(x) = \frac{\Gamma(\Sigma_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \Pi_i x_i^{\alpha_i - 1}$	$x \in < 0, 1 >, \Sigma_i x_i = 1$

- $\Gamma(k) = k!$, $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$

Příklad: USA senát: hledání skupin podobně hlasujících

Justin Grimmer: An Introduction to Bayesian Inference via Variational Approximations

- N senátorů $i = 1, ..., N$.
- každý v právě jednom z K klastrů $k = 1, ..., K$ podobně hlasujících
- τ_i je K rozměrný vektor, indikuje klastr i–tého senátora
- pravděpodobnost klastrů (bloků) π
- pravděpodobnost i–tý senátor v bloku k: $\tau_i | \pi \approx Multinomial(1, \pi)$
- pozorovali jsme J hlasování
  - každé hlasování má pro každý blok parametr 'pravděpodobnost pozitivního hlasování'
    $\theta_k = (\theta_{k1}, \ldots, \theta_{kJ})$
- i–tý senátor (v bloku $\tau_{ik} = 1$), j–té hlasování
  - $V_{ij} | \tau_{ik} = 1 \approx Bernoulli(\theta_{kj})$
- apriorní rozložení předpokládáme
  - $\pi \approx Dirichlet(\alpha)$
  - $\forall k, j \ \theta_{kj} \approx Beta(\gamma_1, \gamma_2)$

'Běžné metody'

- MCMC vzorkování má problém
  - pro různé permutace 'identifikátorů bloků' stejný výsledek i likelihood
  - snadno může přeskakovat a nedržet se jednoho konkrétního
  - pak jsou napočtené četnosti i hodně daleko od reality
- EM klastrování
  - hodně parametrů (s každým hlasováním další)
  - potřebuje invertovat matici $218,694 \times 218,694$
  - potřebuje hodně dat
  - senátorů ale není neomezeně
- EM v M kroku počítá maximálně věrohodný model = JEDEN
- následující algoritmus bude držet (přibližné) rozložení na 'všech' modelech

Variational approximation

- Chceme 'posterior' $p(model|data) = p(\theta, \pi, \tau)$
- aproximujeme pomocí $q(\theta, \pi, \tau)$
- a předpokládáme nezávislost parametrů, $q(\theta, \pi, \tau) = q(\pi)\Pi_{k=1}^{K}\Pi_{j=1}^{J}q(\theta_{kj}\Pi_{i=1}^{N}q(\tau_i)$
- předpokládáme rozložení (viz dříve)
  - $q(\tau_i) = Multinomial(1, r_i)$, kde $r_i = (r_{i1}, \dots, r_{iK})$
  - $q(\pi) = Dirichlet(\lambda)$, kde $\lambda = (\lambda_1, \dots, \lambda_k)$
  - $\forall k, j\ q(\theta_{kj}) = Beta(\eta_{kj1}, \eta_{kj2})$
- iterujeme podobně jako v EM, jen minimalizujeme KL divergenci $KL(q(), p(|data))$.
- Máme $\lambda_k^{old}, \theta_{kj}^{old}$, chceme r_{ik}^{new}:

$$r_{ik}^{new} \propto \exp\left\{ E(log\pi_k) \sum_{j=1}^{J} \{V_{ij}\left(E(log\theta_{kj1})\right) + (1 - V_{ij})\left(E(log\theta_{kj2})\right)\} \right\}$$

where $\mathrm{E}[\log\ \pi_k] = \Psi(\lambda_k^{\text{old}}) - \Psi(\sum_{z=1}^{K} \lambda_z^{\text{old}})$, $\mathrm{E}[\log\ \theta_{kj1}] = \Psi(\eta_{kj1}^{\text{old}}) - \Psi(\eta_{kj1}^{\text{old}} + \eta_{kj2}^{\text{old}})$, and $\mathrm{E}[\log\ \boldsymbol{\theta}_{kj2}] = \Psi(\eta_{kj2}^{\text{old}}) - \Psi(\eta_{kj1}^{\text{old}} + \eta_{kj2}^{\text{old}})$. $\Psi(\cdot)$ is the *Digamma* function, the derivative of the *Gamma* function. Next, we set λ_k^{new} to $\lambda_k^{\text{new}} = \alpha_k + \sum_{i=1}^{N} r_{ik}^{\text{new}}$. And finally, we set $\boldsymbol{\theta}_{kj}^{\text{new}}$ to, $\theta_{kj1}^{\text{new}} = \gamma_1 + \sum_{i=1}^{N} r_{ik}^{\text{new}} V_{ij}$ and $\theta_{kj2}^{\text{new}} = \gamma_2 + \sum_{i=1}^{N} r_{ik}^{\text{new}} (1 - V_{ij})$.

$$\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

Příklad: USA senát: hledání skupin podobně hlasujících

Table 1 Voting blocs in U.S. Senate

Label	*Example senators*	*Distinctive Vote*	*%*
Cons. Rep	Coburn, DeMint, Inhofe, Sessions	Amendment 521: Reduce Federal Debt	37.7
Mod. Rep	Coleman, Hagel, Lugar, Murkowski	Amendment 2662: Prohibit Canyon Funds	12.2
Mod. Dem	Bayh, McCaskill, Lieberman, Ben Nelson	Cloture on S. 2633: Iraq Redeployment	17.0
Lib. Dem	Clinton, Kennedy, Obama, Sanders	Table Amendment 4388: Mortgages	33.0

Součin členů exp. rodin

- součin dvou členů exp. rodin je také z exp. rodin,
- ale nemusí mít hezkou parametrickou formu

$$h(x)e^{\boldsymbol{\theta_1}^T T(x)-A(\theta)} \cdot h(x)e^{\boldsymbol{\theta_2}^T T(x)-A(\theta)} = \tilde{h}(x)e^{(\boldsymbol{\theta_1}+\boldsymbol{\theta_2})^T T(x)-\tilde{A}(\boldsymbol{\theta_1},\boldsymbol{\theta_2})}$$

- součet dvou členů exp. rodin už do exp. rodin patřit nemusí.

Bayesovské rozhodování, konjugované hustoty

- $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$
- máme–li 'hezké' $p(\theta)$ a $p(x|\theta)$, součin $p(x|\theta)p(\theta)$ už být hezký nemusí
- ale může: když zvolíme vhodné hustoty
- Dirichlet a Multinomial
- gausovské a gausovské
- beta a bernoulli
- a další.