

Statisztikai tanulás az idegrendszerben, 2017.

Inferencia valószínűségi modellekben

Bányai Mihály

banyai.mihaly@wigner.mta.hu

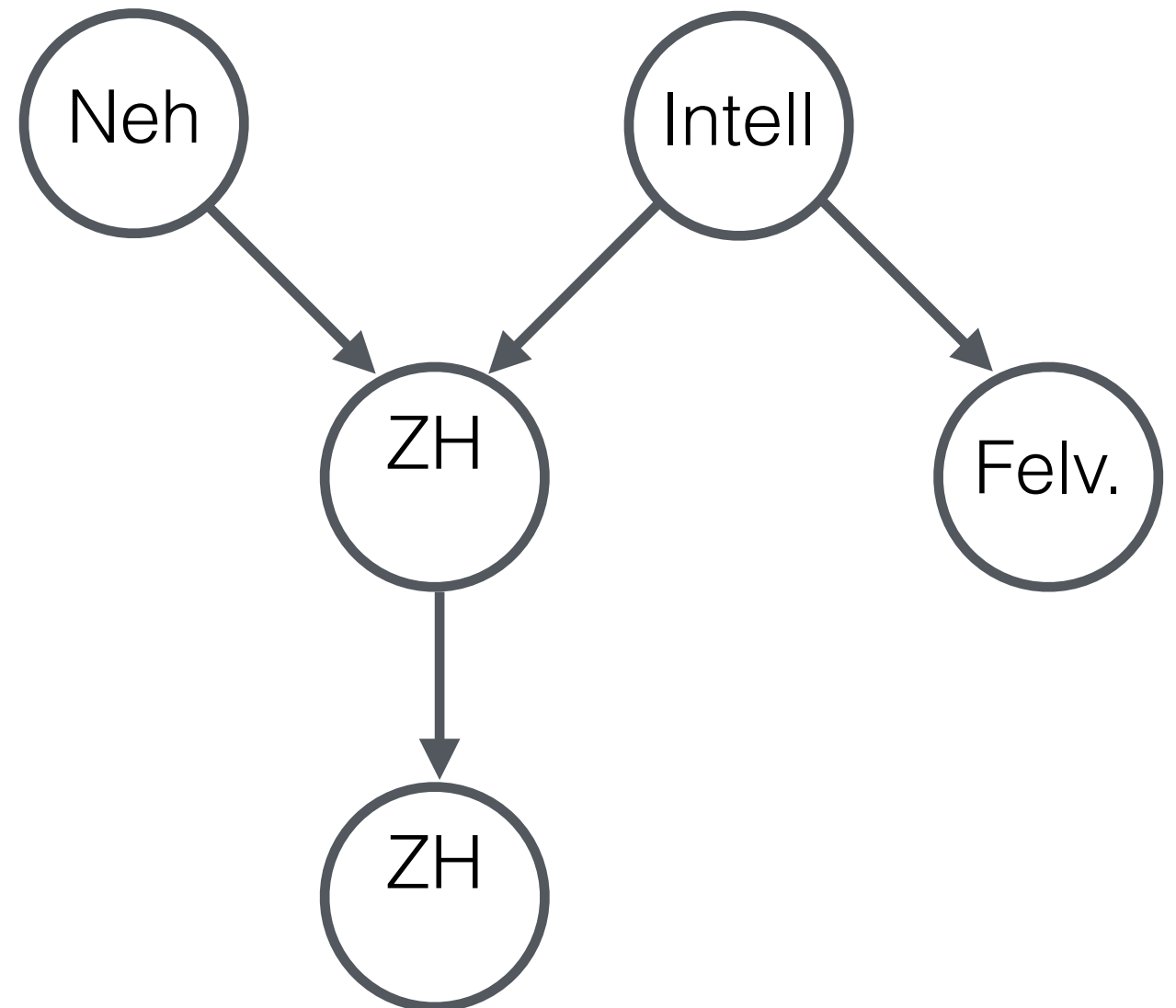
<http://golab.wigner.mta.hu/people/mihaly-banyai/>

- Inferencia valószínűségi modellekben
- Közelítő inferencia
- Algoritmusok pontbecslésekhez
- Inferencia az agyban

- Inferencia valószínűségi modellekben
- Közelítő inferencia
- Algoritmusok pontbecslésekhez
- Inferencia az agyban

Generatív modellek

- Mi a valószínűségi modell?
- Miért érdekes számunkra a valószínűségszámítás?

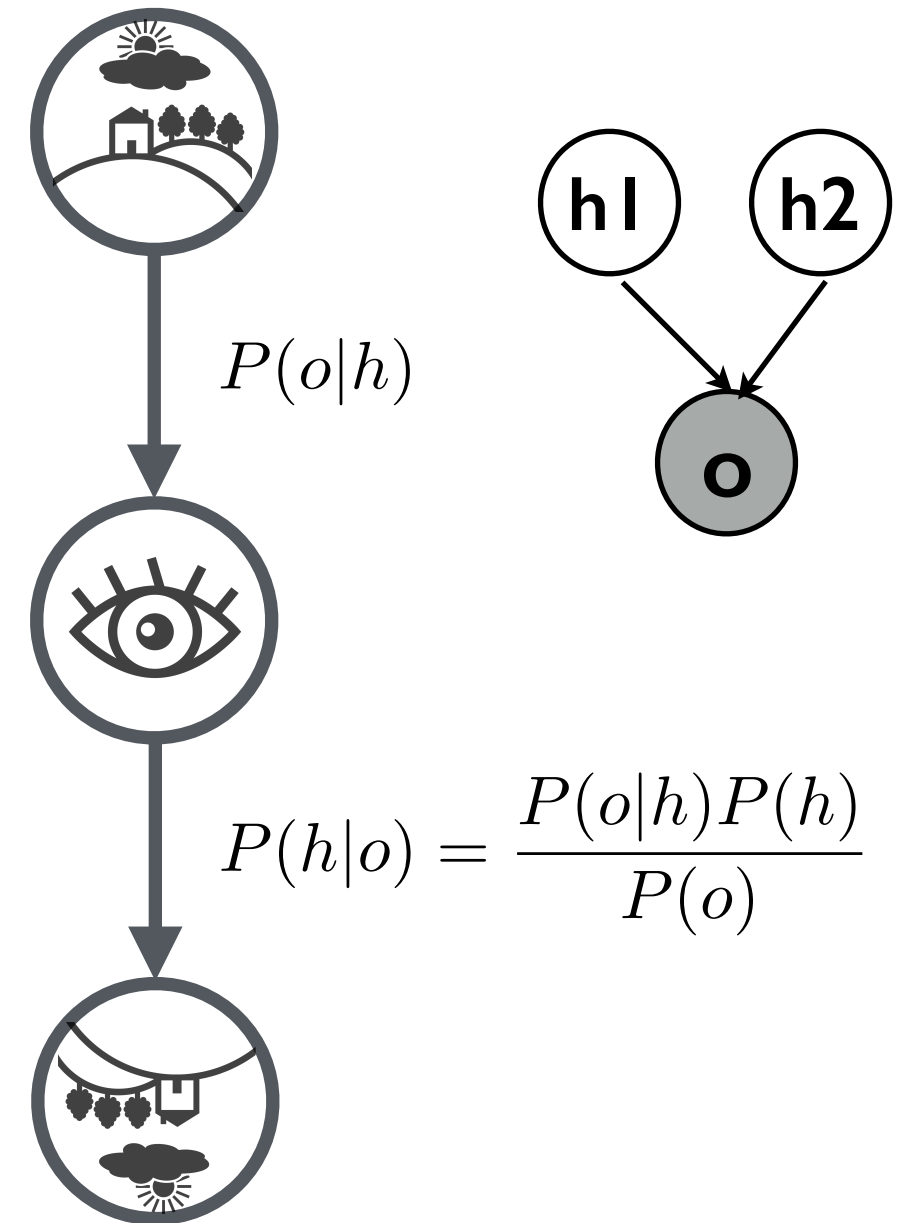


Mit jelentenek a nyilak?

- Generatív irányban a függetlenség
- a kauzalitás intuíciója

Megfigyelt és rejtett változók

- bizonyos változókról rendelkezünk adatokkal
 - jelölés: sötét kör
- a többi csak a jelenség struktúrájáról alkotott feltételezéseinket reprezentálja
 - látens, rejtett változók: üres kör
- Mire lehetünk kíváncsiak
 - poszterior eloszlás a látensek fölött
 - marginális poszteriorok
 - a prediktív eloszlás: a megfigyelt változók marginális eloszlása



Bayes-i inferencia

- A generatív valószínűségi modell egy hipotézis (vagy hipotéziscsalád) formalizációja
- Ha hiszünk a Cox- és/vagy dutch book jellegű érvelésben, akkor a nem közvetlenül megfigyelt mennyiségekről rendelkezésünkre álló információ konzisztens karakterizációja a poszterior eloszlás
- a hipotézistesztesztelés más formái heurisztikák, vagy speciális feltételezéshalmazok beépítései a (sokszor implicit) generatív modellekbe

Az inferencia nehézségei

- Zárt alakban nem megadható eloszlások
 - integrálok, marginalizáció
 - szorzatalakok Bayes tétel sokmegfigyelésen történő alkalmazásából

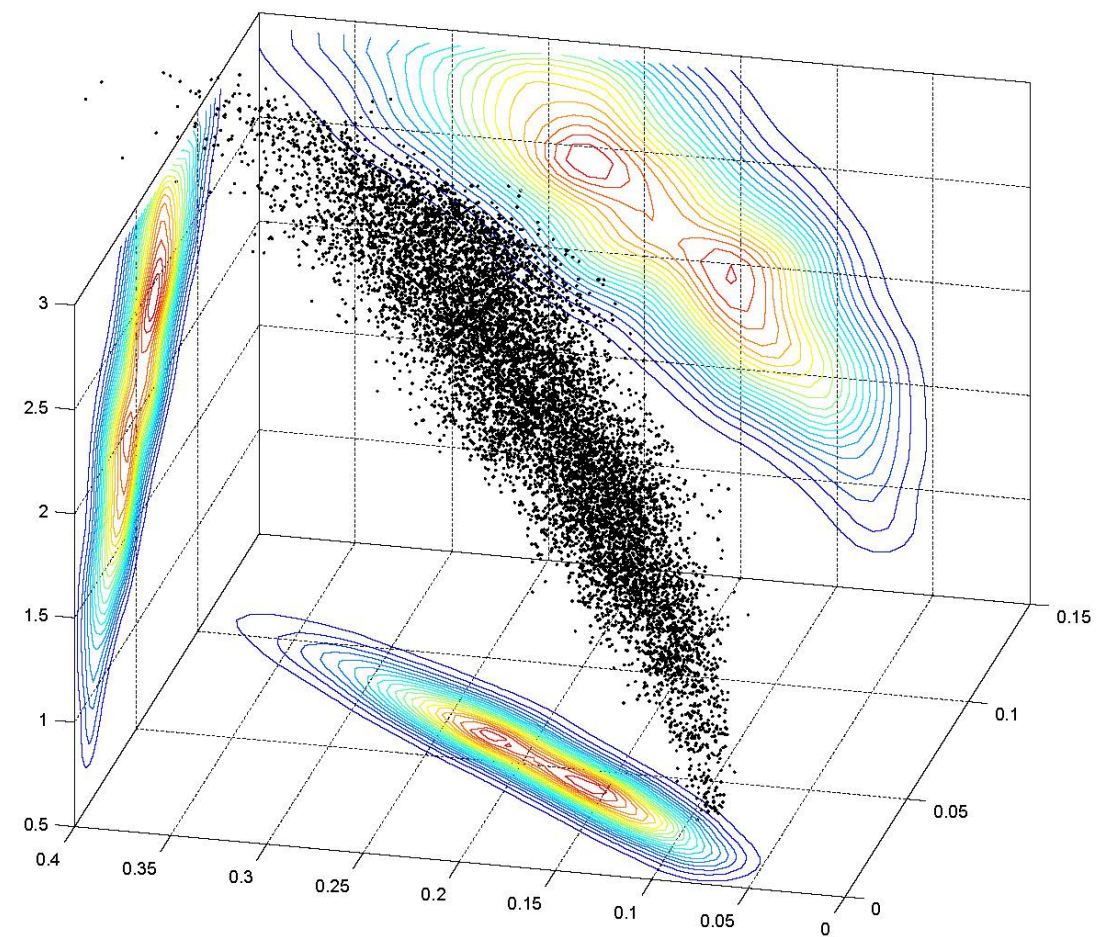
- Inferencia valószínűségi modellekben
- **Közelítő inferencia**
- Algoritmusok pontbecslésekhez
- Inferencia az agyban

Poszterior közelítő becslése

- Sztochasztikus
 - mintákat veszünk belőle
 - véletlenszámgenerátorra van szükség
 - a minták számától függ a pontosság
- Determinisztikus
 - egyszerűbben kezelhető eloszlásokkal közelítjük
 - nem használunk véletlenszámokat
 - a közelítő eloszlás formájától függ a pontosság

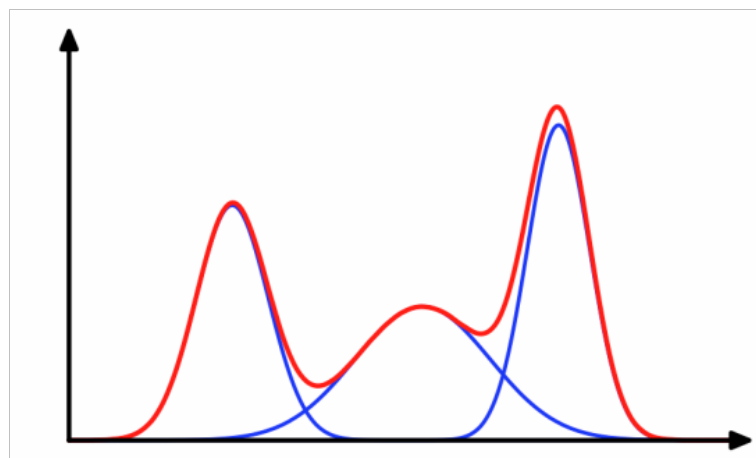
Sztochasztikus becslés

- A poszterior eloszlást az abból vett mintákkal ábrázoljuk
- Aszimptotikusan egzakt
- Könnyen implementálható
- Számításigényes lehet
- A következő órán bővebben

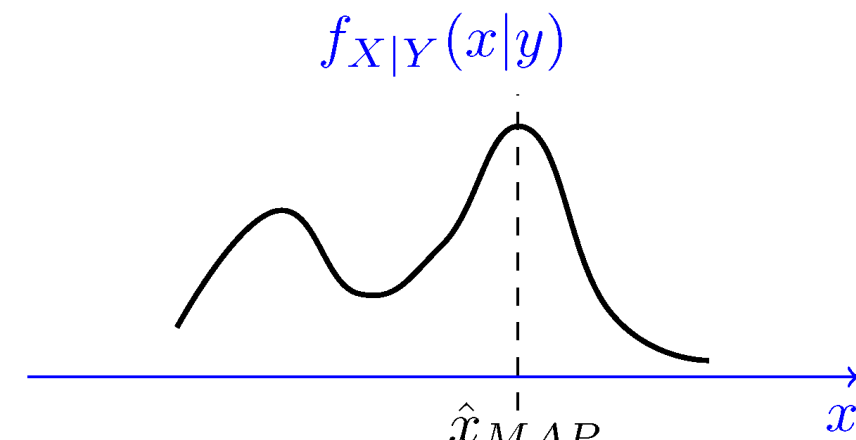
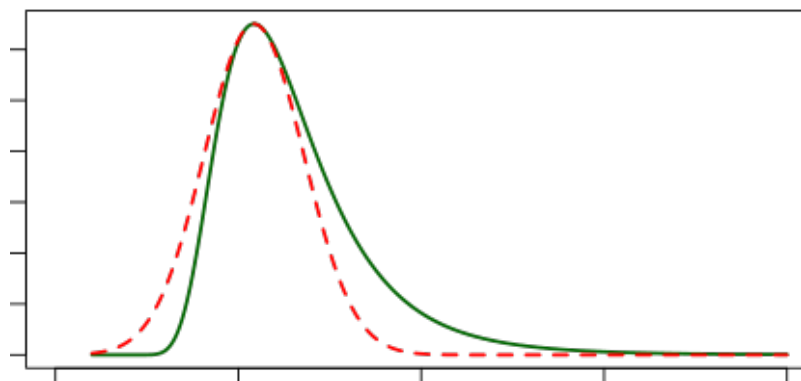


Determinisztikus közelítés - variációs módszerek

- A kérdés, hogy mennyi információt dobunk el a poszteriorból?
 - Gyakran a függőségek teszik nehezzé az inferenciát -> faktorizált közelítés:
 $p(h_1, h_2 | o) \approx p(h_1 | o) p(h_2 | o)$
 - Jól kezelhető parametrikus eloszlásokkal is közelíthetünk
 - bonyolultabbakkal, pl többkomponensű keverékeloszlások
 - egyszerűbbekkel, pl Gauss (Laplace-közelítés)
- Egyetlen jellemző értéken kívül mindent eldobunk -> delta eloszlással közelítünk -> pontbecslés



Laplace approximation of posterior of Normal(10,σ)

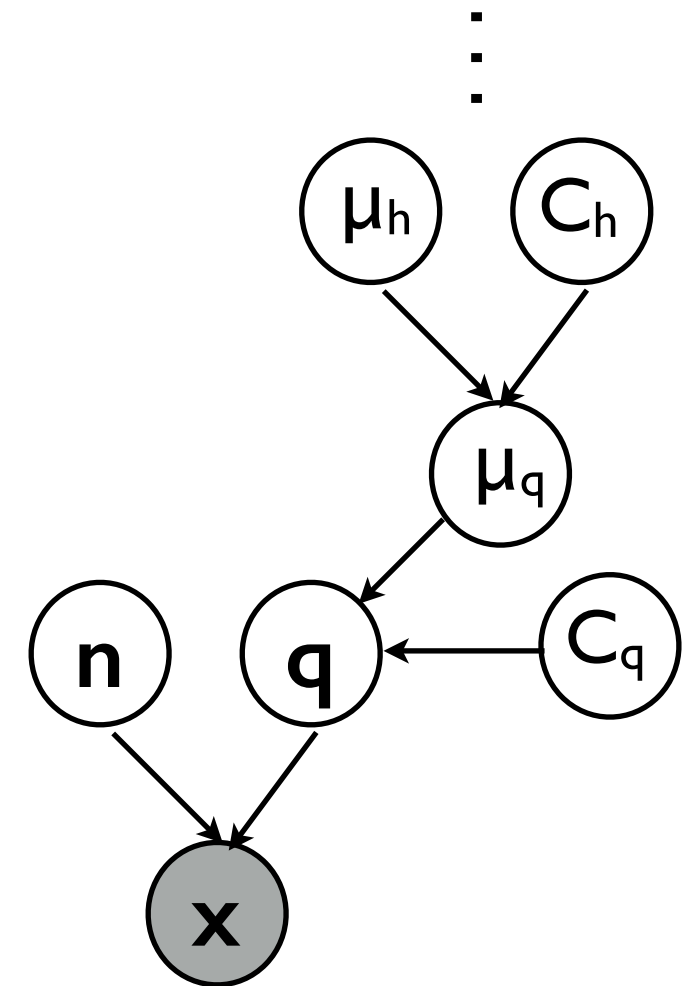


Probabilisztikus programozási nyelvek

- Deklaratív nyelvek (mint pl. a PROLOG)
- A generatív modellt és a megfigyeléseket kell specifikálni
- Látenis változók poszterior eloszlására vonatkozó inferencia nyelvi elemként használható
- Ismertebb példák: BUGS, Church, Stan

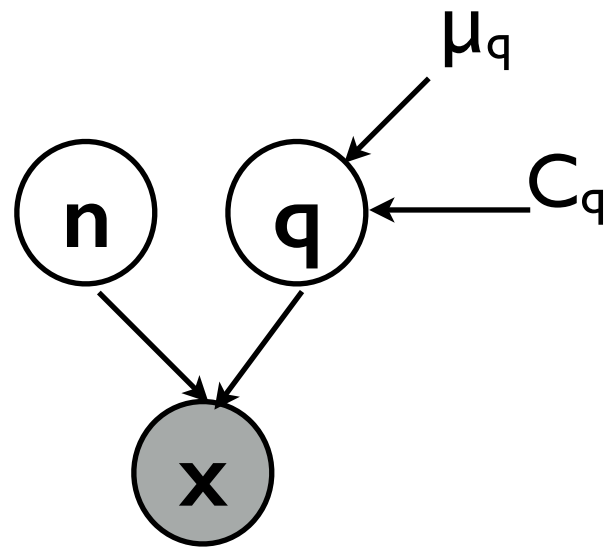
Paraméterek mint valószínűségi változók

- a valószínűségi modelben specifikálnom kell a változók prior és kondicionális eloszlásait
 - ezek gyakran parametrikus eloszláscsaládok példányai lesznek, pl. Gauss, Poisson, stb
 - ezek paramétereit szintén látens változók, amennyiben nem fixálom az értékeket teljesen
- a változó prior eloszlásának ismét lesznek paramétereit -> hierarchikus modell
 - valahol meg kell állnom
- vagy olyan priort választok, amiben nincs paraméter, illetve ami mégis, azt nem illesztem az adatra, hanem a világ konstans tulajdonságának tételezem fel az általam választott értékét



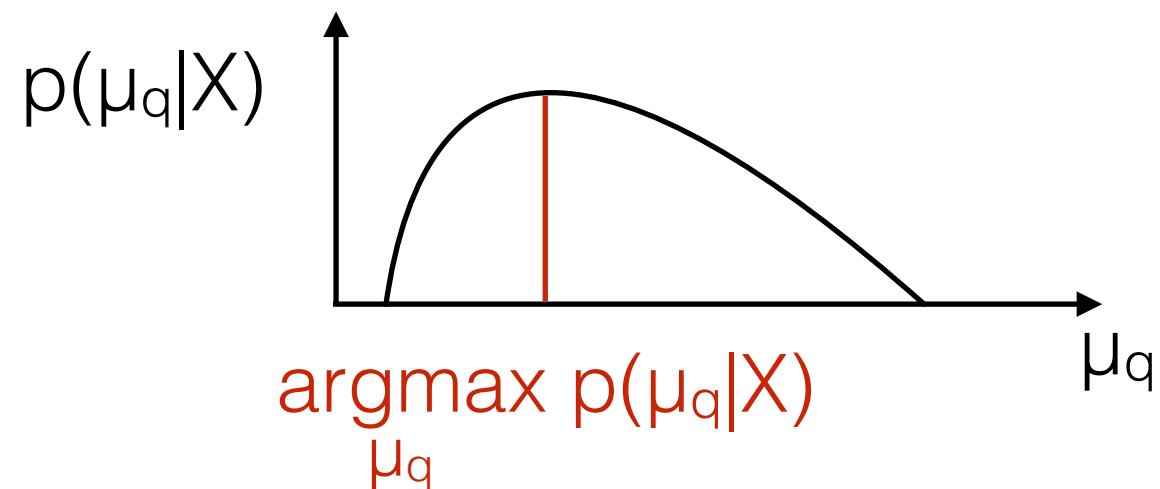
Pontbecslések

- Az inferencia általában a poszterior keresése
 - ha nem kell a teljes poszterior, csak egy pontbecslés, akkor szélsőértékkeresséssé redukálhatom
 - ha nincs prior: maximum likelihood
- Hogyan változik pontbecslésnél a prediktív eloszlás bizonytalansága?



$$p(\mu_q, C_q \mid X) \sim p(X \mid \mu_q, C_q) \cancel{p(\mu_q, C_q)}$$

konstans



Maximum likelihood tanulás

- Általában több mérésünk van a megfigyelt változókról, amiket függetlennek tekintünk

- A likelihood faktorizálódik
$$p(X | \theta) = \prod_{n=1}^N p(x_n | \theta)$$

- A log-likelihood maximumát keressük

- numerikusan stabilabb
$$\ln p(X | \theta) = \sum_{n=1}^N \ln p(x_n | \theta)$$

- exponenciális formájú eloszlásoknál egyszerűbb

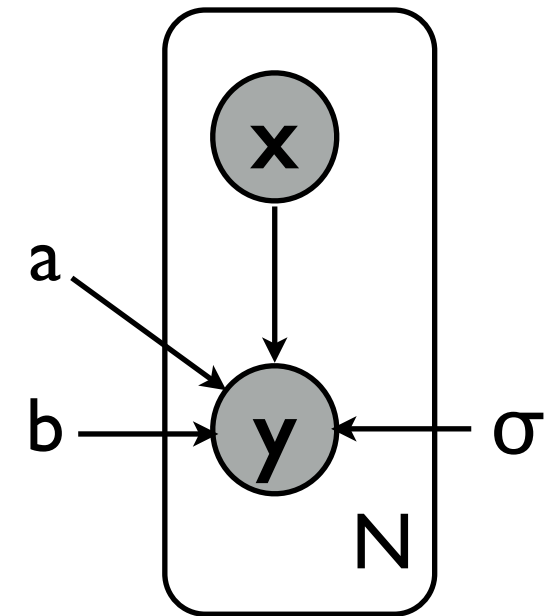
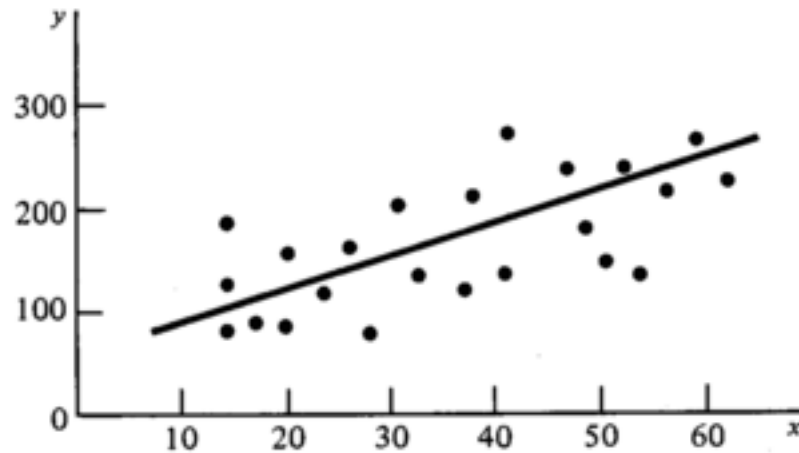
- Ki kell integrálni a látens változókat (várható érték)

$$p(X | \theta) = \int_{-\infty}^{\infty} p(X | Z, \theta) p(Z | \theta) dZ$$

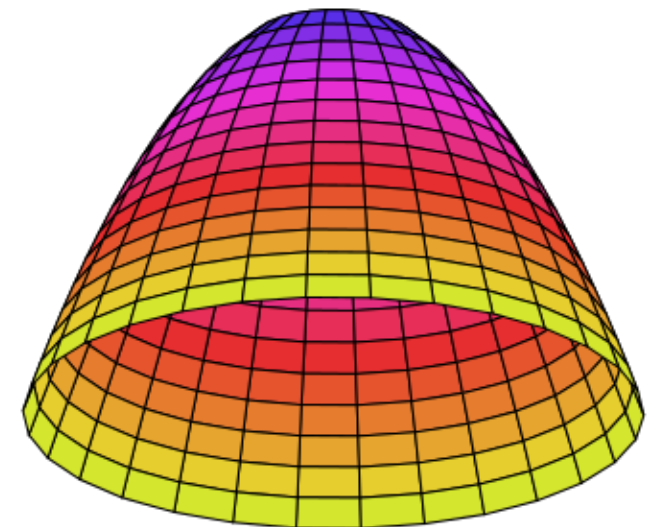
Mikor hasznosak a pontbecslések?

- Ha jellemző a becsült pont értéke a poszterior alakjára
 - unimodális
 - szimmetriája a pontbecslés természetével egyező

Példa - lineáris regresszió



- Nincs rejtett változó a paramétereken kívül
- Gauss maximum likelihood = négyzetes hiba
- polinomokra változtatás nélkül kiterjeszthető



$$p(y \mid x, a, b, \sigma) = \mathcal{N}(y; ax + b, \sigma)$$

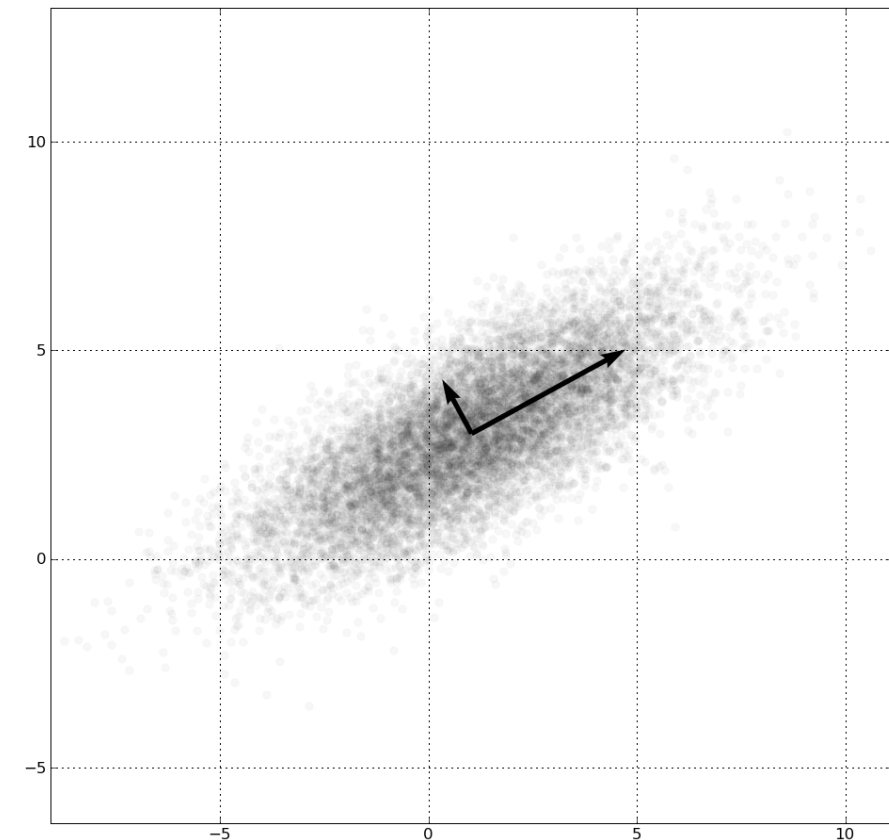
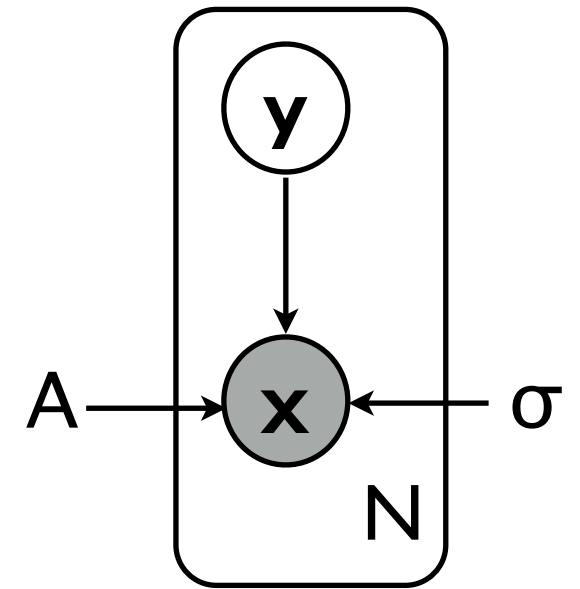
$$\ln p(Y \mid X, a, b, \sigma) \sim - \sum_{n=1}^N [(ax_n + b) - y_n]^2$$

Példa - PCA

$$p(x \mid y) = \mathcal{N}(x; Ay, \sigma I)$$

$$p(y) = \mathcal{N}(y; 0, I)$$

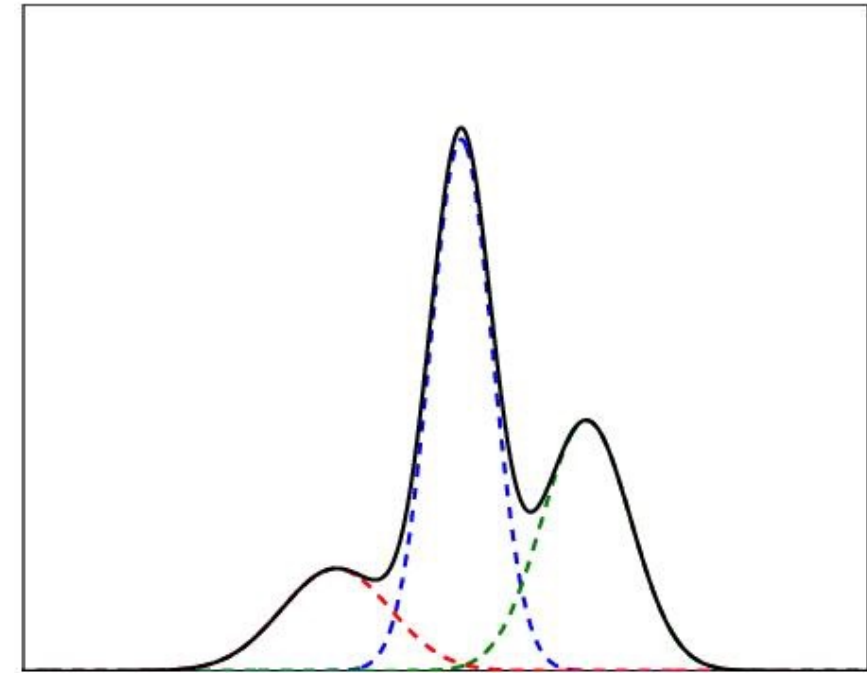
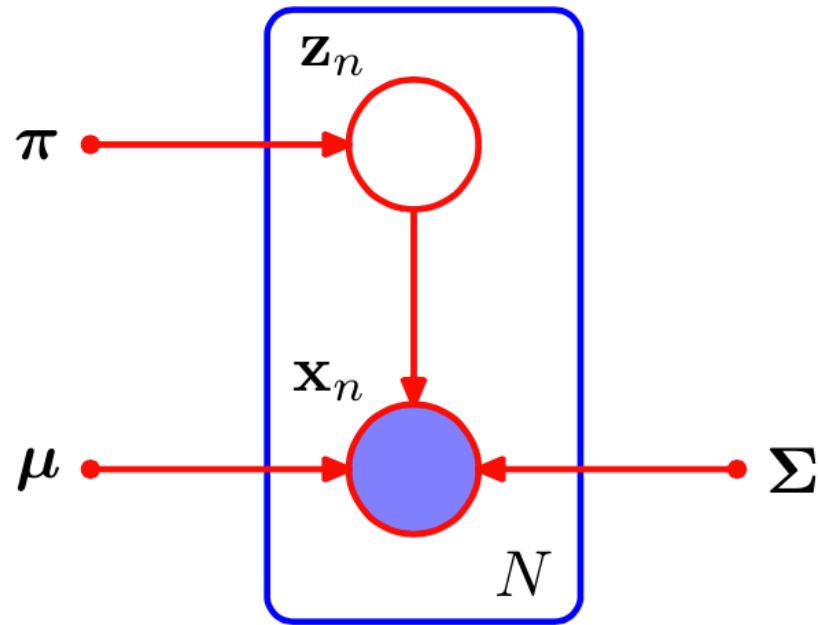
- a PCA egy bázistranszformáció, ahol a legnagyobb szórás irányába szeretnénk beforgatni a tengelyeket
- maximum likelihood megoldás az A keverőmátrix elemeire fix szórás mellett
- ekvivalens a megfigyelések kovarianciamátrixának sajátvektor-keresési problémájával



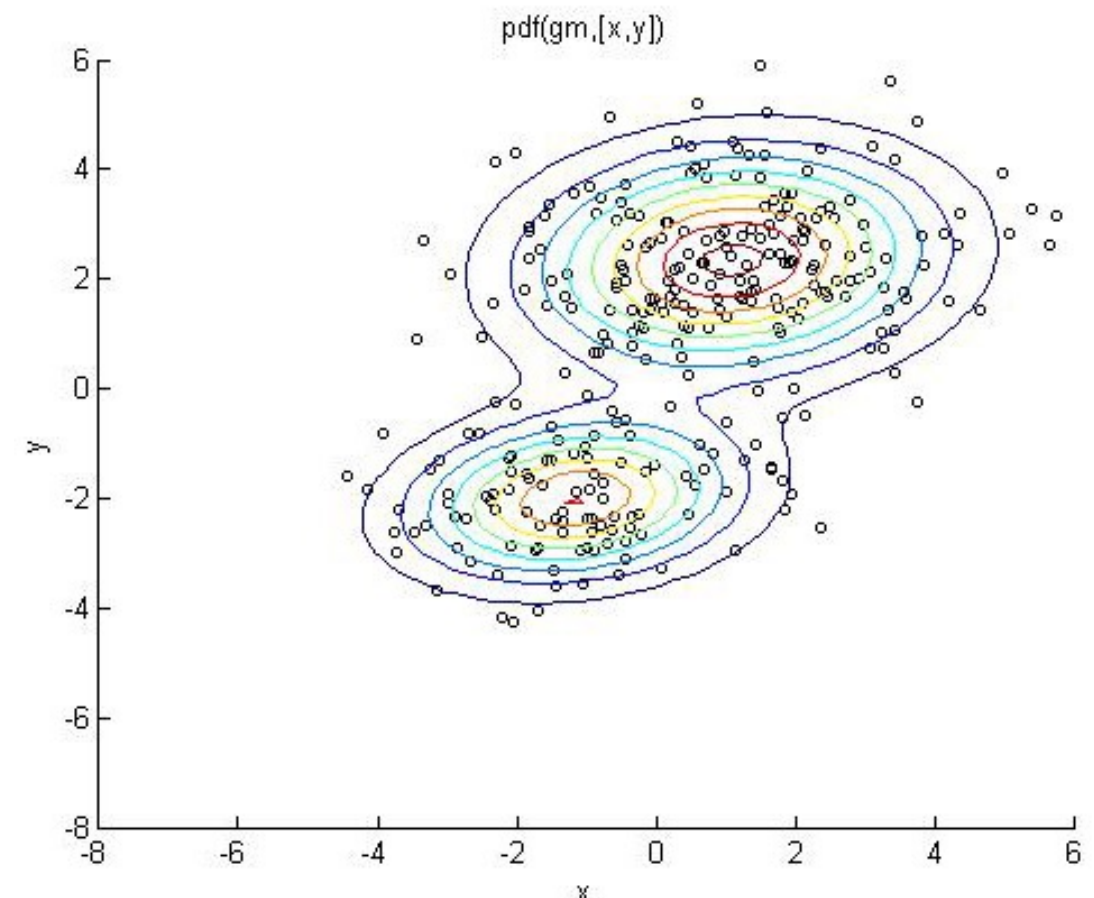
Amikor a pontbecslés is nehéz

- Általában itt is igaz, hogy vannak olyan láttens változók amelyek felett marginalizálni akarunk, aminek lehet, hogy nincs zárt alakja
- vagy az *argmax* nem fejezhető ki zárt alakban

Keverékmodellek



$$p(z) = Mult(z; \pi)$$
$$p(x | z) = \mathcal{N}(x; \mu_z, \Sigma_z)$$



- Inferencia valószínűségi modellekben
- Közelítő inferencia
- **Algoritmusok pontbecslésekhez**
- Inferencia az agyban

Algoritmikus becslés

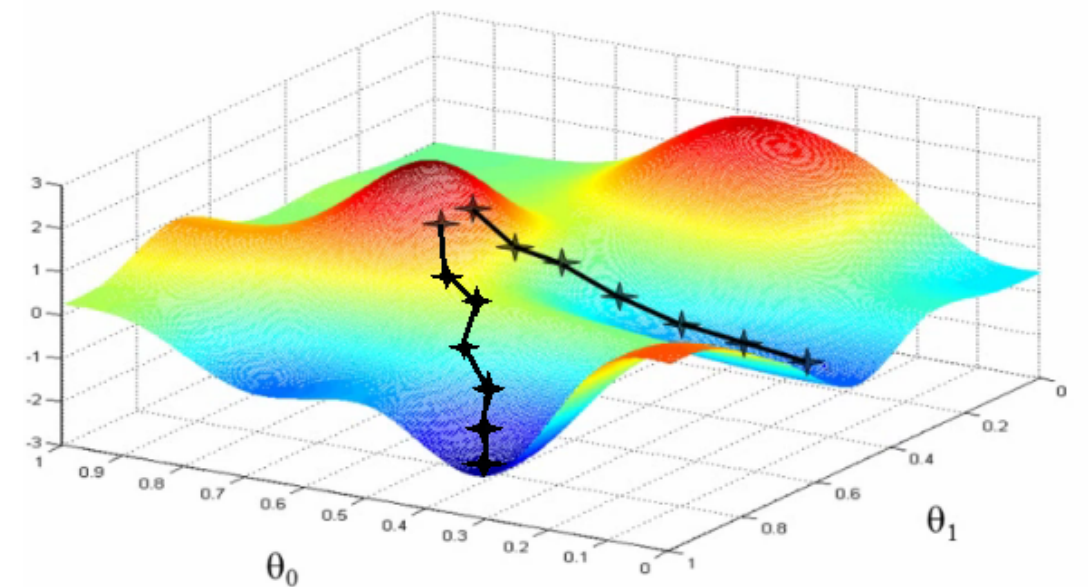
- ha a pontbecslés nem adható meg egzaktul, megfogalmazhatjuk optimalizációs problémaként, amelyben a hibafüggvény pl. a negatív log-likelihood
- a becslőalgoritmusok nem (csak) a valószínűségszámítás szabályait használják
- sokszor nagyon hatékonyak az iteratív algoritmusok, akár zárt formájú megoldásoknál is gyorsabbak lehetnek nagy dimenziójú modellekre
- viszont sokszor nem tudjuk levezetni, hogy mennyire pontosan találják meg a globális maximumot

Gradiens-módszer

- használjuk ki azt az információt, hogy adott pontban merre emelkedik a célfüggvény (vagy csökken a hibafüggvény), pl likelihood

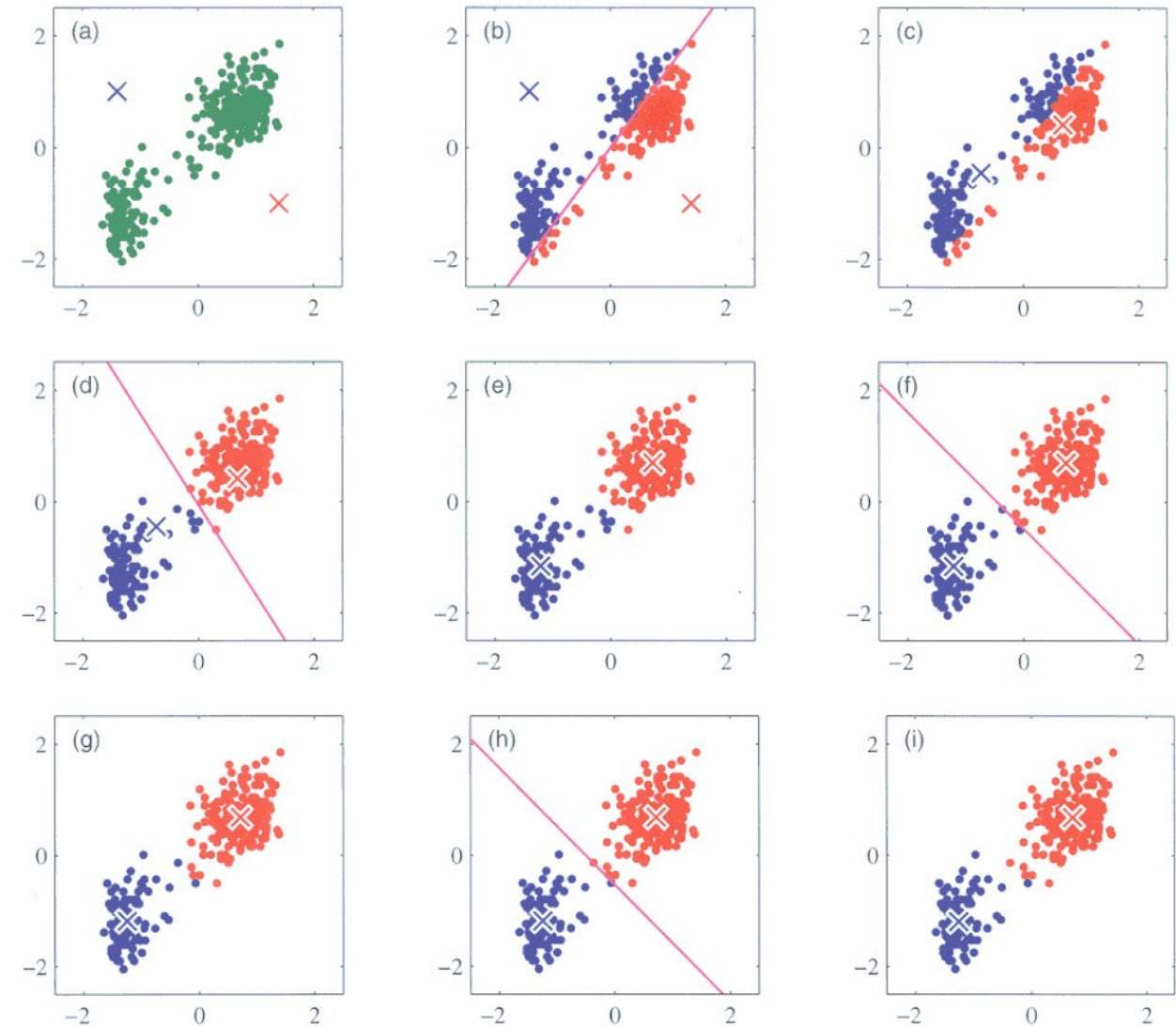
$$\theta_{t+1} = \theta_t + \varepsilon \frac{\partial}{\partial \theta} \ln p(X | \theta_t)$$

- lokális szélsőértéket talál
- a tanulási rátát megfelelően be kell állítani
- kiterjesztések
 - második derivált, Hessian-based módszerek
 - lendületet is definiálhatunk a paraméterterbeli mozgáshoz



K-means klaszterezés

- Klaszterközéppontokat keresünk
- Kiválasztjuk a klaszterek számát
- Véletlenszerűen inicializáljuk a középpontokat
- Hozzárendeljük a megfigyeléseket a legközelebbi klaszterközépponthoz
- Elmozgatjuk a klaszterközéppontokat úgy, hogy a négyzetes távolság a hozzárendelt pontoktól
- Addig ismételjük, amíg átsorolás történik
- Mi az ekvivalens valószínűségi modell?



Expectation Maximization

- Általánosítsuk a k-means ötletét
 - complete-data likelihood: mintha minden változó megfigyelt lenne

$$p(X, Z \mid \theta)$$

- Az algoritmus kétfajta lépést váltogat
 - E: megbecsüljük a látens változók poszterior eloszlását a paramétereket fixen tartva

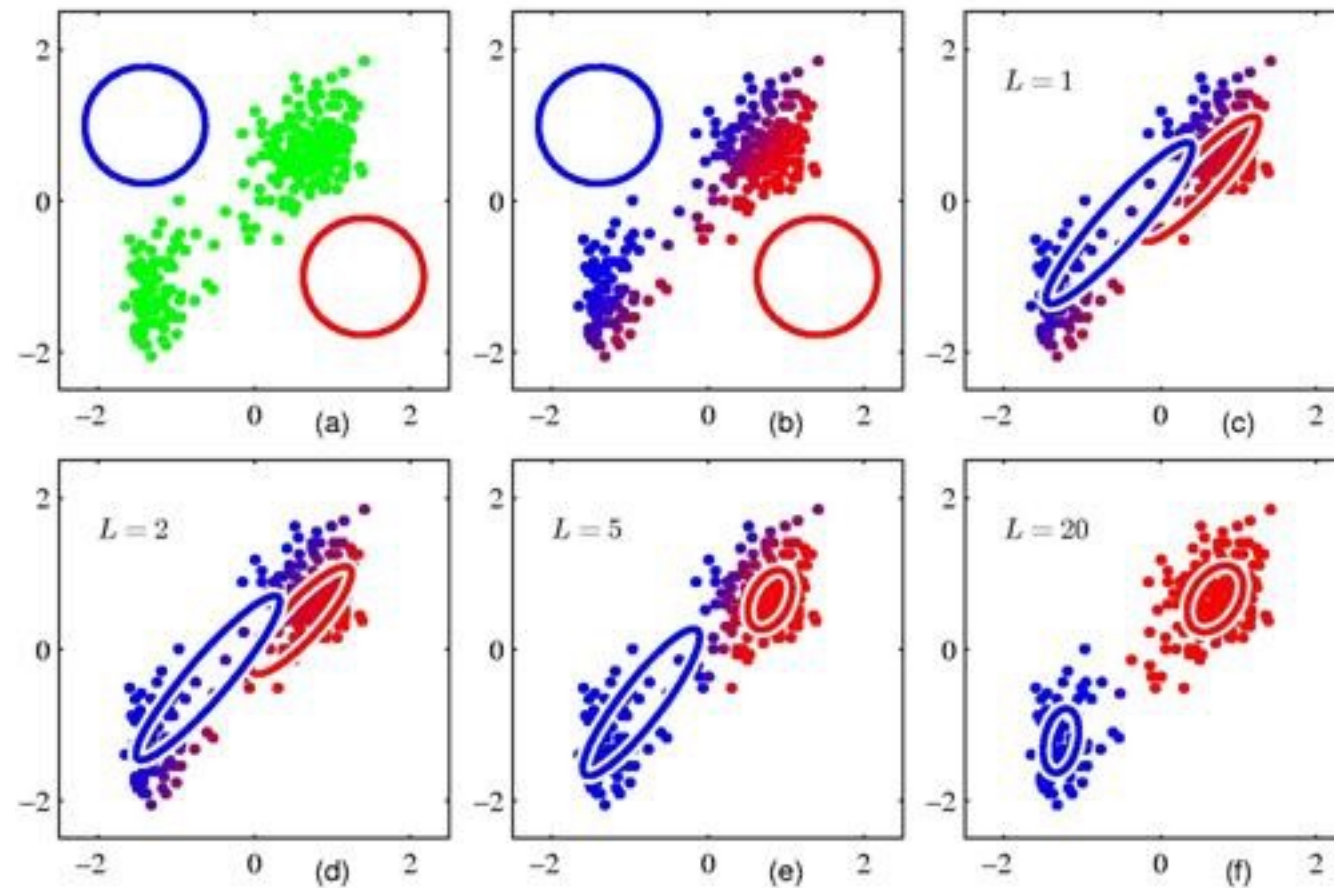
$$p(Z \mid X, \theta_t)$$

- M: megbecsüljük a paraméterek értékét az becsült poszterior alapján
 - a CDL logarimusának vesszük a poszterior feletti várható értékét, és ezt maximalizáljuk a likelihood helyett

$$\theta_{t+1} = \operatorname{argmax}_{\theta} \int_{-\infty}^{\infty} p(Z \mid X, \theta_t) \ln p(X, Z \mid \theta) dZ$$



EM keverékmmodellre



- E: mi a poszterior valószínűsége, hogy egy adott pontot egy adott komponens generált a jelenleg becsült paraméterekkel
 - a komponensek pontokért viselt “felelőssége”
- M: a felelősségekkel súlyozott pontokra mi lenne a legjobb mean és kovariancia
 - keverési együtthatók: a felelősségek összegének aránya a pontok számához

Általánosított EM

- Mi van, ha nem tudjuk zárt alakban megadni az M-lépésben keresett szélsőértéket?
 - gradiens módszer
- Mi van, ha nem tudjuk zárt alakban megadni a CDLL poszterior szerinti várható értékét?
 - mintavételezés a poszteriorból, az integrál közelítése véges számú minta feletti átlaggal

Honnan tudjuk, hogy az algoritmus eredménye hasznos?

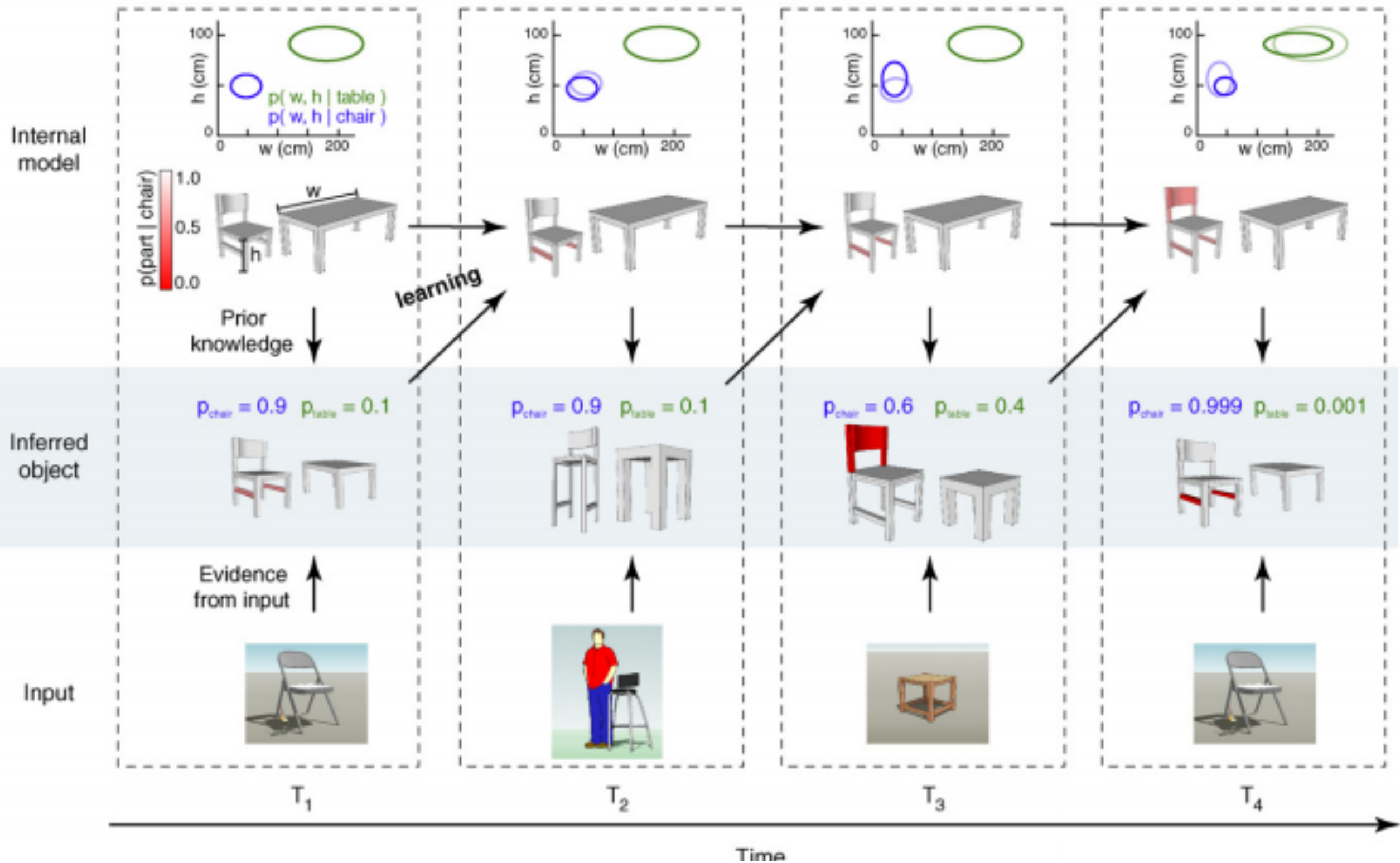
- Konvergencia
 - likelihoodban
 - paraméterek értékeiben
- Az illesztett modell predikciói beválnak
- Honnan tudjuk, hogy jól választottuk meg az olyan hiperparamétereket, mint pl. a komponensek száma?

- Inferencia valószínűségi modellekben
- Közelítő inferencia
- Algoritmusok pontbecslésekhez
- **Inferencia az agyban**

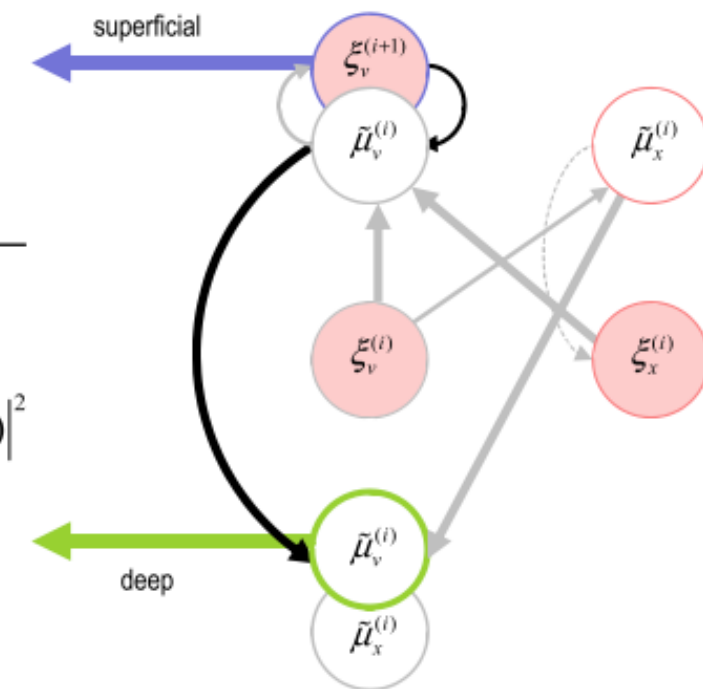
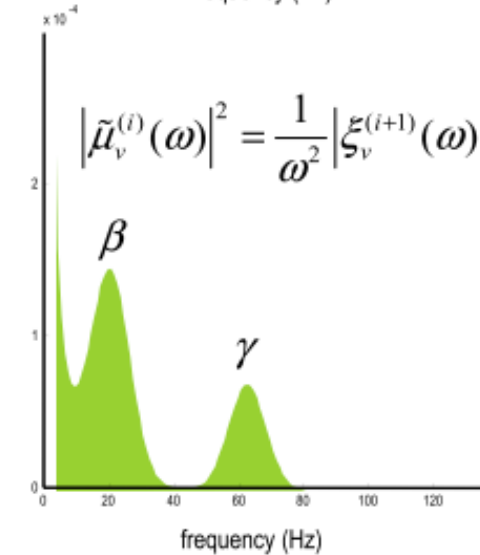
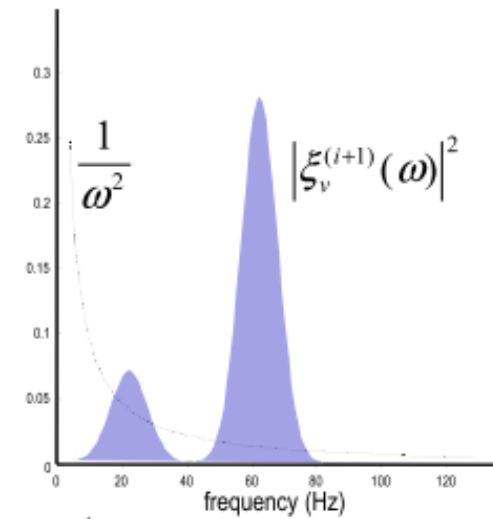
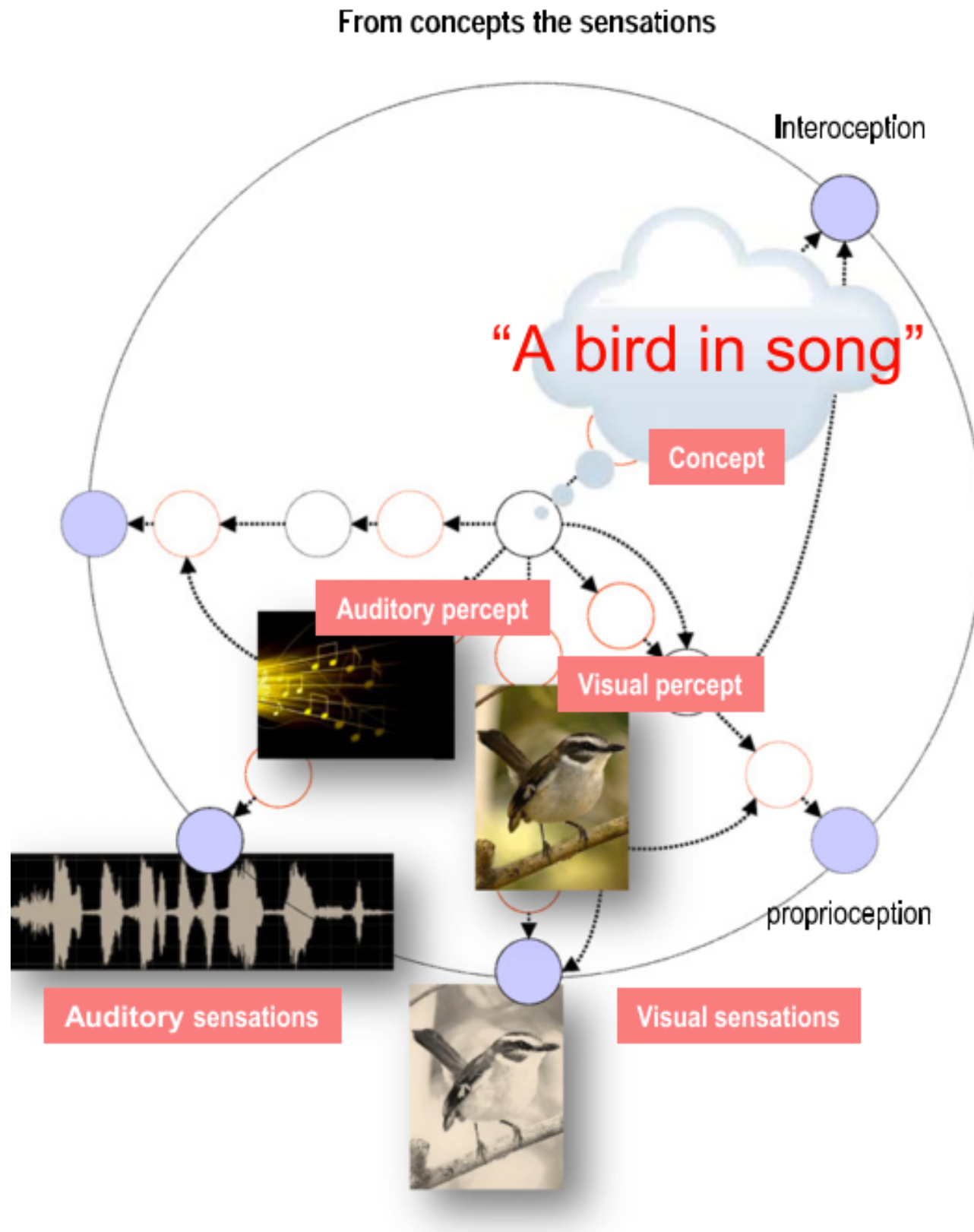
Hogyan tehetünk predikciókat az illesztett modellel?

- A jövőre a prediktív eloszlásból
- A látens változók értékeire a poszteriorból
- Ha agykérgi tanulórendszer modelljét alkotjuk, akkor el kell döntenünk, hogy
 - a paraméterillesztési algoritmust biofizikai folyamatok predikciójára akarjuk használni
 - vagy az algoritmust csak arra használjuk, hogy eljussunk egy optimális modellig, és csak azt feltételezzük, hogy az agy is megteszi ezt valahogy, de nem feltétlenül így

Iteratív becslés az agyban - viselkedési szinten

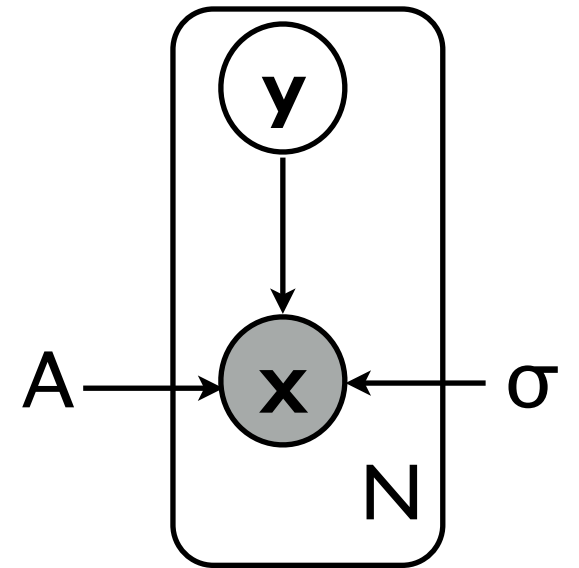


Iteratív becslés az agyban - fiziológiai szinten



Elsődleges látókéreg generatív modellje

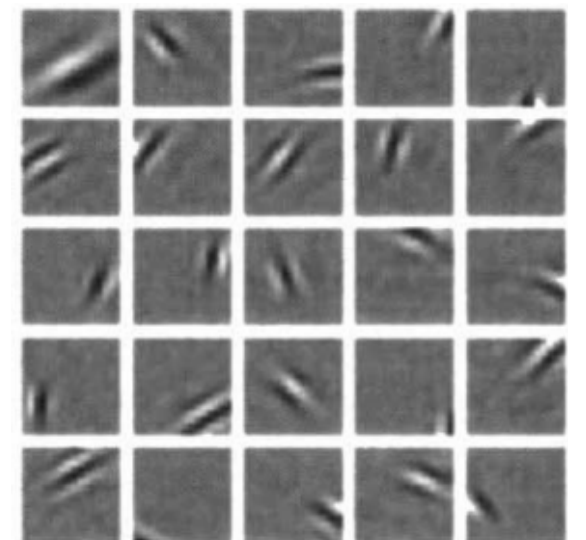
- Olshausen-Field modell: természetes képek független komponensei
- pontbecslés, kombinálva a maximum likelihood és a sparsity kritériumokat
- log-poszterior = kvadratikus log-likelihood + sparse prior
- Egyszerre optimalizáljuk az adatra illeszkedést és a látens aktivációk ritkaságát
- Gradiens-módszerrel lépkedünk az így konstruált hibafüggvényen



$$p(x | y) = \mathcal{N}(x; Ay, \sigma I)$$



sparsity



Házi feladat

- Töltsd le a tárgyhonlapról az em_hf.txt file-t
- Az adatokra illessz 3-komponensű Gaussian Mixture modellt
- Eredményedet illusztráld