Bayesískar aðferðir í hagrannsókum: Saga, heimspeki og aðferðir

Helgi Tómasson
helgito@hi.is
Háskóli Íslands

9. Október 2008

Skipulag Erindis

1. Uppruni tölfræði
2. Klofningur í decision-fræði/mælinga-inference
3. Bayesísk heimspeki og aðferðarfræði
4. Einfalt sýnidæmi
5. Hugleiðing um flóknara dæmi

Hvað er tölfræði

1. Hagnýting á líkindafræði
2. Líkindafræði er stærðfræðigrein. Líkindamálið mælir stærð á mengjum.

$$\mathbb{P} : \mathcal{F} \to [0,1], \quad \mathcal{F} = \text{mengi mælanlegra mengja},$$
$$\boldsymbol{X} : \Omega \to \mathbb{R}, \quad x = \boldsymbol{X}(\omega) = \text{gildi sem}$$
$$\boldsymbol{X} \text{ er fall sem úthlutar útkominni } \omega \text{ gildið } x,$$
$$\Omega = \text{ mengi af mögulegum útkomum},$$
$$\boldsymbol{X} \text{ er kölluð hending eða slembibreyta}$$

3. Tölfræði=Ályktunarfræði (Inference) + Líkindafræði
4. Laplace leggur grunn að hagnýtingum eins og „inverse-probability" (fyrir mælingar) og decicsion-fræði
5. Seinna koma menn eins of Fisher, Pearson-feðgar, Neyman og Jeffrey.

- Ályktunarfræði er að mestu mat (estimation) og prófanir.
- Ýmsir metríu-kúltúrar= Econometrics, biometrics, pschycometrics, sociometrics, chemometrics, technometrics, criminometrics o.s.frv.
- Ýmis talnameðferð er ekki metría í mínum huga. Svo sem bestunarfræði, þ.e. tækni við að hámarka föll. Bestunarfræði byggist oft á rúmfræði en ekki líkindafræði, þ.e. að koma kúrvu „nálægt" einhverjum punktum.
- Tölfræði fjallar um að álykta út frá mælingum.
- Sumir nota orðið tölfræði (statistics) um eitthvað sem ég myndi kalla reikningshald og endurskoðun.
- Rúmfræðigreiningar eins og „support-vector-machines" og að hræra í gögnum, neural-network, data-mining er að mínu viti ekki tölfræði. (Data-mining hægt að gera rétt með einhvers konar AIC/BIC tólum)

Nokkur prinsíp við mat á parametrum í líkönum

- Least-Squares (rúmfræði)
- Method-of-Moments
- Maximum-Likelihood
- Bayes-aðferðir

Regla Bayes fyrir atburði:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

og fyrir þéttiföll (og líkindamassaföll) dreifinga:

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)}$$

Gangur í Bayes-tölfræði

1. Set fram líkan, $f(\boldsymbol{x}|\boldsymbol{\theta})$
2. Set fram fyrirframvissu (a priori) um $\boldsymbol{\theta}$ á formi líkindadreifingar, $\pi(\boldsymbol{\theta})$.
3. Reikna eftirávissu (a posteriori) í formi líkindadreifingar, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$, með reglu Bayes.
4. Tengist decision-teoríu t.d. með því að lágmarka expected-posteriori-loss.

Hvað þýðir að vita ekkert?

1. Hvað er non-informative prior?
   Maximal-Data-Information-Prior-Distribution {Zellner, 1977}
2. Vel upplýsingafall, $I(\theta)$ sem endurspeglar upplýsingar um θ.
   Upplýsingar að lokinni tilraun eru:

$$I_{posterior}(\theta) = I_{data}(\theta) + I_{prior}(\theta)$$

   Leita að prior sem að hámarkar $I_{data}(\theta)$.
3. Hvaða upplýsingafall á að nota?

$$I(\theta) = E(\log(p(\theta)), \quad I(\theta) = E(-\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'})$$

1. Geisser {1984} rökstyður nokkrar a priori dreifingar fyrir Bernoulli módel $P(Y = 1) = \theta, \quad P(Y = 0) = 1 - \theta$.

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1/2},$$
$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2},$$
$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1} \text{ og}$$
$$p(\theta) \propto \text{constant}$$

2. Ef gögnum er safnað, $Y_1, \ldots, Y_n$ til að meta θ þá má til dæmis hugsa sér að ákveða n fyrirfram (binomial sampling) eða að hætta þegar tiltekinn fjöldi af $Y = 1$ hefur verið mældur (negative-binomial sampling).
3. Ekki er æskilegt að non-informative prior sé háður því hvernig mælingum er safnað.
4. Bernardo {1979} rekur að ekki sé til einhlít lausn á að setja fram „non-informative" prior. Hann stingur upp á annari lausn sem hann kallar „reference-prior".

1. Ekki hafa öll líkön reference-prior, en allar exponential-family dreifingar.
2. Non-informative prior ekki vel skilgreinanlegt hugtak.
3. Í umræðum um grein Bernardo {1979} segir A.F.M. Smith: *My own view of "vague" or "improper" prior is that they are simply mathematical artefacts (having no intrinsic interest in their own right) whose justification rests on the fact that their use in Bayes' theorem results in a posterior which is a "good approximation" in some sense,*
4. Í praktískum tilfellum eru oft notaðir „diffuse" priorar, þ.e. dreifingar sem hafa líkindamassa á mjög stóru svæði.

Tæknierfiðleikar

1. Ef mælingar eru $x = (x_1, \ldots, x_n)$ þá er posterior θ

$$f(\theta|x) \propto L(\theta|x)\pi(\theta)$$

2. Erfitt getur verið að átta sig á hvernig $f(\theta|x)$ líkur út, sem og að reikna t.d. $E(\theta|x)$, $V(\theta|x)$, o.s.frv. Sérstaklega ef θ er margvíður vektor.
3. Fyrir „exponential-family-dreifingar" getur verið hentugt að vinna með „conjugate"-par af líkani og prior.

Conjugate priorar

1. Ef prior $\pi(\theta)$ og posterior $\pi(\theta|x)$ tilheyra sömu fjölskyldu er sagt að priorinn sé „closed under sampling"
2. Þó að prior sé conjugate er ekki þar með sagt að t.d. $E(\theta)$, $V(\theta)$, o.s.frv. verði þægilegar formúlur.
3. Ef conjugate-prior er valinn þá er posterior á sama formi. Því eru eftiráupplýsingar úr sömu dreifingu og fyrirframvissan.
4. Conjugate-prior er aðeins til fyrir exponential-family líkön og ekki alltaf meðfærilegur.

Einfalt dæmi

Hugsum okkur Bernoulli-líkan (krónukast),
$P(Y = 1) = \theta$, $P(Y = 0) = 1 - \theta$.
Þá er conjugate-prior fyrir θ beta-dreifing, $Beta(\alpha, \beta)$, þ.e.

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Ef gögn eru random úrtak, $y = (y_1, \ldots, y_n)$ þá er sennileika-fallið,

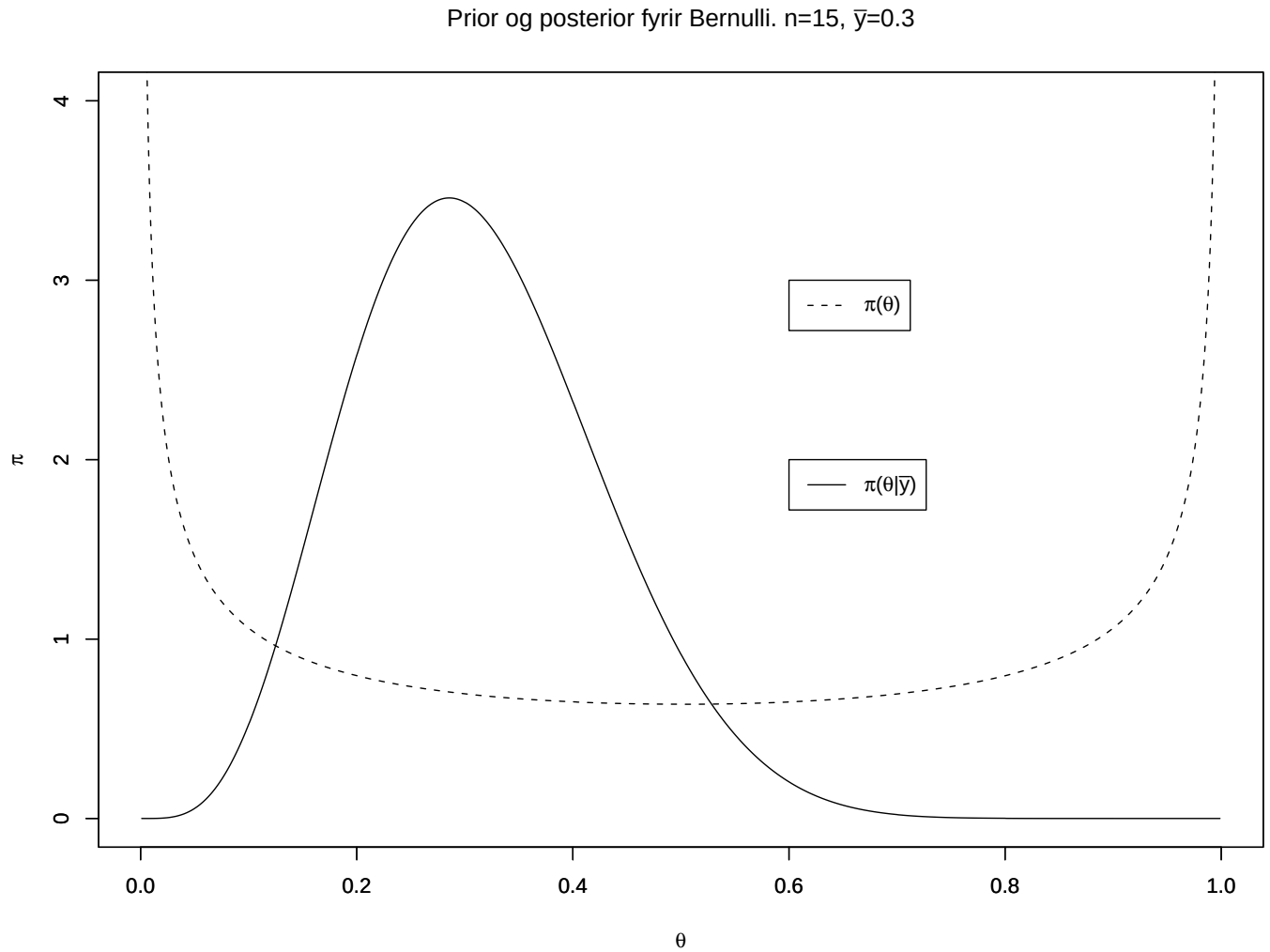
$$L(\theta|y) \propto \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}$$
$$= \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}$$

og posterior því,

$$\pi(\theta|y) \propto \pi(\theta)L(\theta|y) = \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}$$
$$= \theta^{n\bar{y}+\alpha-1}(1-\theta)^{n(1-\bar{y})+\beta-1}.$$

Þ.e. við það að safna gögnum hafa upplýsingar um θ breyst úr því að vera $Beta(\alpha, \beta)$ í að vera $Beta(\alpha + n\bar{y}, \beta + n(1-\bar{y}))$. Ef θ er $Beta(\alpha, \beta)$-dreifð þá er

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad \text{og } V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
$$E(\theta|y) = \frac{\alpha+n\bar{y}}{\alpha+n\bar{y}+\beta+n(1-\bar{y})} = \frac{\alpha+n\bar{y}}{\alpha+\beta+n} \xrightarrow[n\to\infty]{} \bar{y}$$
$$V(\theta|y) = \frac{(\alpha+n\bar{y})(\beta+n(1-\bar{y}))}{(\alpha+\beta+n)^2(\alpha+\beta+1)} \xrightarrow[n\to\infty]{} \frac{\bar{y}(1-\bar{y})}{n}$$



Mynd: Prior og posterior fyrir Bernoulli líkan.

Fyrir normaldreifingu, $N(\mu, \sigma^2)$ er normal conjugate fyrir μ gefið σ og IG=inverse-gamma conjugate fyrir σ^2.

$$\pi(\mu|\sigma) \sim N(m_0, A\sigma^2), \quad \pi(\sigma^2) \sim IG(\alpha, \beta),$$
$$f(y|\mu, \sigma)$$

Formúlur eru aðeins snyrtilegri ef sett er $v = 1/\sigma^2$, þ.e. v gammadreift. Ef $y_1, \ldots, y_n$ er random úrtak þá er

$$\pi(\mu|v) \propto v^{1/2} e^{-(\mu - m_0)^2 v/(2A)}$$
$$\pi(v) \propto v^{\alpha-1} e^{-\beta v}$$
$$\pi(\mu, v|y_1, \ldots, y_n) = \pi(\mu|v)\pi(v)L(\mu, v|y_1, \ldots, y_n)$$
$$\propto v^{1/2} e^{-(\mu - m_0)^2 v/(2A)} v^{\alpha-1} e^{-\beta v} \left(v^{1/2}\right)^n e^{-v \sum_{i=1}^n (y_i - \mu)^2/(2A)}$$

menntaskólaalgebra gefur

$$\propto v^{\alpha + n/2 - 1/2} e^{-v\beta} e^{-v(1+An)(\mu - (m_0 + An\bar{y})/(1+An))^2/(2A)}$$

Þ.e. posterior dreifingin verður:

$$\pi(\mu|v,\bar{y}) = N(\frac{m_0 + An\bar{y}}{1+An}, \frac{1}{v}\frac{A}{1+nA})$$
$$\pi(v) = Gamma(\alpha_1, \beta_1), \quad \alpha_1 = \alpha + (n-1)/2,$$
$$\beta_1 = \beta + (1+An)(\mu - (m_0 + An\bar{y})/(1+An))^2/(2A)$$

Normal-Inverse-Gamma prior í normal líkani er táknað $NIG(m_0, A, \alpha, \beta)$ og er closed under sampling því að við random-sampling er posterior líka NIG.

- Conjugate-prior má túlka sem upplýsingar um sams konar líkan úr eldri mælingum.
- IG-dreifingin er heavy-tail
- $\pi(\mu|\bar{y})$ er t-dreifing (líka heavytail).

Margvítt normal líkan

$\boldsymbol{Y}$ margvítt normal

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-1/2} e^{-(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/2}$$

Conjugate-prior fyrir $\boldsymbol{\mu}$ er normal og fyrir Σ Inverse-Wishart, þ.e. Σ^{-1} er Wishart. Wishart-dreifingin er eins konar margvíð χ^2 dreifing.

Conjugate priorar fyrir nokkur líkön

Líkan	Conjugate prior
Einvítt normal	Normal-Inverse-Gamma, NIG
Margvítt normal	Normal-Inverse-Wishart
Bernoulli	Beta
Margvítt Bernoulli	Dirichlet
Poisson	Gamma
Exponential	Gamma

MCMC hermunaraðferðir

1. Markov-keðja er slembiferli (stochastic-process) sem er þannig að dreifing framtíðar gildis er aðeins háð síðasta gildi.

$$f(y_t|y_{t-1}, y_{t-2}, \ldots) = f(y_t|y_{t-1})$$

2. Hugmyndin að gagnsemi Markov-Chain-Monte-Carlo liggur í þeim eiginleika „time-homogen" Markov-keðju að ef hún er „time-reversible" þá er til „time-invariant" dreifing, jafnvægisdreifing.
3. Gibbs-sampling og Metropolis-Hastings eru þekktar hugmyndir við að herma Markov-keðjur.

Einfalt líkan

Einföld aðhvarfsgreining(regression)

x	y
1	1
2	3
3	3
4	3
5	5

Líkanið er

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \varepsilon \sim N(0, \sigma^2)$$

a priori dreifingar $\alpha \sim N(0, 10000)$ $\beta \sim N(0, 10000)$

$\sigma^{-2} \sim gamma(0.00001, 0.00001)$

Venulegt niðurstaða úr vengulegu tölfræðiforriti er t.d. $\hat{\sigma} = 0.73$

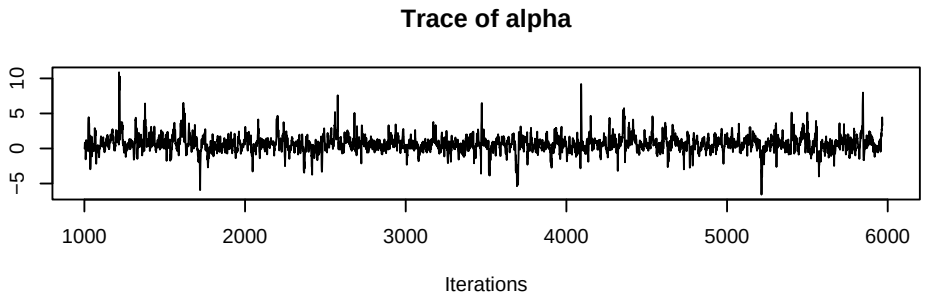
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6000	0.7659	0.78	0.4906
x	0.8000	0.2309	3.46	0.0405

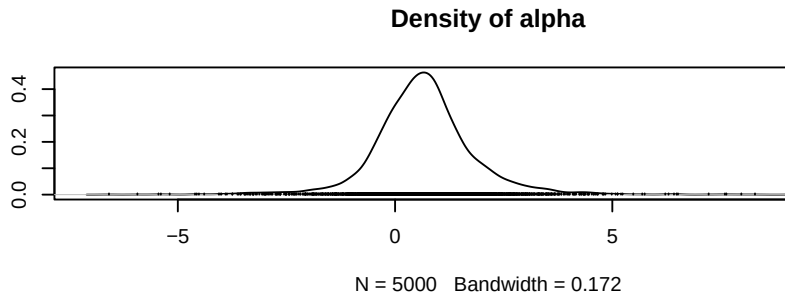
1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

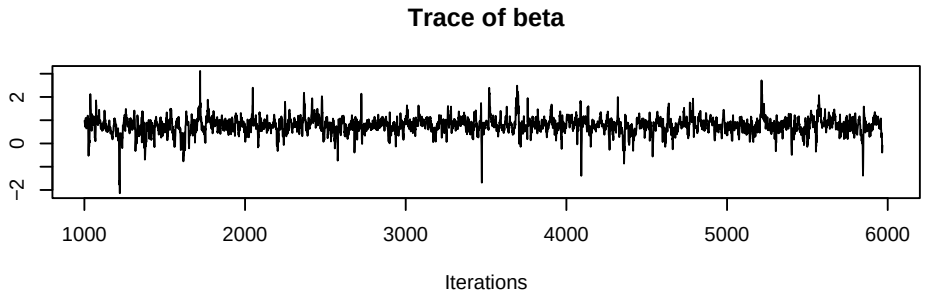
	Mean	SD	Naive SE	Time-series SE
alpha	0.6719	1.2109	0.017124	0.030980
beta	0.7768	0.3605	0.005099	0.009386
sigma	0.9993	0.6285	0.008889	0.013761
tau	1.8749	1.5220	0.021524	0.028852

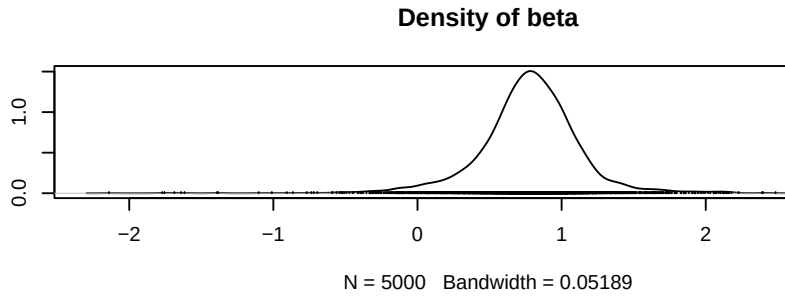
2. Quantiles for each variable:

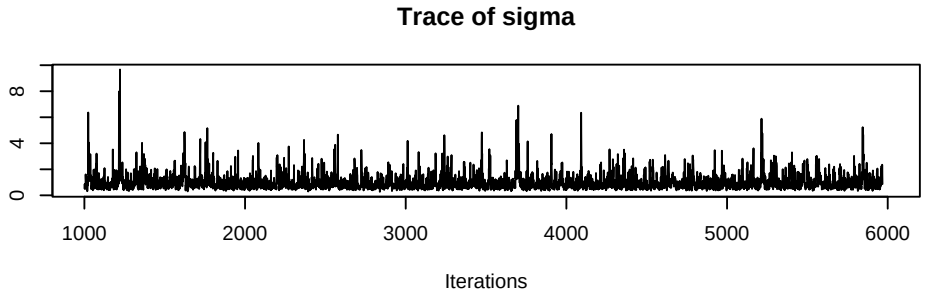
	2.5%	25%	50%	75%	97.5%
alpha	-1.672447	0.03057	0.6327	1.225	3.263
beta	.001338	60776	0.7870	0.968	1.468
sigma	0.415499	0.62246	0.8200	1.171	2.649
tau	0.142518	0.72870	1.4874	2.581	5.792

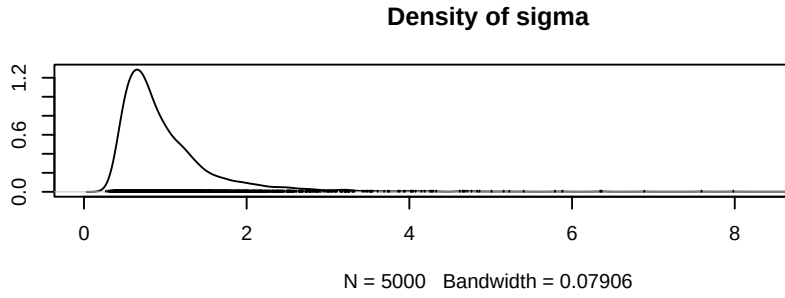

Trace of alpha
−5 0 5 10
1000 2000 3000 4000 5000 6000
Iterations

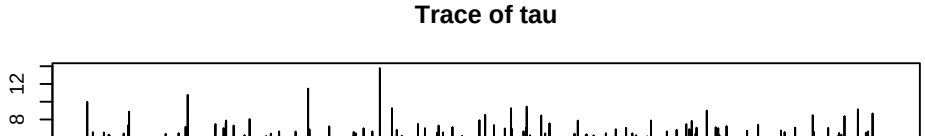

Density of alpha
0.0 0.2 0.4
−5 0 5
N = 5000 Bandwidth = 0.172

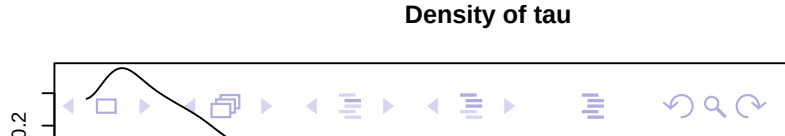

Trace of beta
−2 0 2
1000 2000 3000 4000 5000 6000
Iterations


Density of beta
0.0 1.0
−2 −1 0 1 2
N = 5000 Bandwidth = 0.05189


Trace of sigma
0 4 8
1000 2000 3000 4000 5000 6000
Iterations


Density of sigma
0.0 0.6 1.2
0 2 4 6 8
N = 5000 Bandwidth = 0.07906


Trace of tau
8 12


Density of tau
0.2

Flókin líkön

1. Analýtískt form á posterior bara til í fyrir conjugate-par af líkani og prior (og nokkur einföld sértilfelli)
2. Jafnvel þó að conjugate prior sé til getur verið að hann sé illskiljanlegur og t.d. erfitt að finna $E(\theta)$, $V(\theta)$, o.s.frv.
3. Á tímum tölvutækni hafa menn farið út í að herma postiori dreifingarnar og skoða þær grafískt.
4. Þörf á aðferð sem býr til random tölur úr posterior dreifingunni.
5. Það getur verið erfitt að skrifa forrit sem hermir random-úrtak úr tiltekinni dreifingu.
6. Reyni að ofsampla í hölunum og vigta síðan. „Importance-sampling" (líkt og stratified sampling),
7. Í dag eru aðferðir sem herma Markov-keðjur notaðar.

Um markaðskannanir (ordered-regression)

- Þáttakendur svara spurningum og kvarða frá 1-5. Þýðir þetta eitthvað?

	Númer spurningar					
	1	2	3	4	5	6
Einstaklingur 1	4	5	4	5	4	5
Einstaklingur 2	2	3	2	3	2	5
Einstaklingur 3	1	5	1	5	1	5

Tafla: Sammála einstaklingar.

Líkan fyrir multivariate ordered response

Ómæld J-víð breyta y_i^* fyrir einstakling i,

$$y_i^* \sim N(\mu_i^*, \Sigma_i^*).$$

Mæli skorna útgáfu y_i af y_i^*. vera með einstaklingsbundin μ_i^* og Σ^*. Að sjálfsögðu má ekki slá saman mismunandi einstaklingum. Að sjálfsögðu þarf ýmis hliðarskilyrði til að hægt sé að meta svona. Til dæmis:

$$\Sigma_i^* = \sigma_i \Sigma.$$

Þ.e. að einstaklingarnir hafi sameiginlegt fylgnifylki. Síðan má setja upp Gibbs-sampler með hluta priors að Σ sé Inverse-Wishart. Einnig má nota skýristærðir, t.d.

$$\mu_i^* = \mu_i + \boldsymbol{x}\boldsymbol{\beta}$$

Dæmi um útkomur

	$\hat{\beta}_{ML}$	s.e.-β	t-gildi
kyn	0.15	0.22	0.67
einkunn	0.31	0.11	2.89

Tafla: Niðurstaða ML-mats fyrir eina spurningu í einu námskeiði.

	$E(\beta)$	$\sqrt{V(\beta)}$
kyn	0.15	0.22
einkunn	0.31	0.11

Tafla: Meðaltal og staðalfrávik bayesísks mats (byggt á mjög „diffuse prior", fyrir eina spurningu í einu námskeiði.

1	2	3	4	5	6	7	8	9	10	11
1.00	0.06	0.39	0.01	0.10	0.24	0.04	−0.07	0.10	0.05	0.02
0.06	1.00	0.26	0.14	0.26	0.22	0.29	0.24	0.07	−0.09	−0.01
0.39	0.26	1.00	0.05	0.12	0.35	0.12	0.02	0.06	−0.03	−0.10
0.01	0.14	0.05	1.00	0.09	−0.02	0.13	0.21	0.18	0.23	0.20
0.10	0.26	0.12	0.09	1.00	0.17	0.33	0.32	0.04	−0.11	0.12
0.24	0.22	0.35	−0.02	0.17	1.00	0.40	0.11	−0.05	−0.09	0.02
0.04	0.29	0.12	0.13	0.33	0.40	1.00	0.48	0.05	−0.13	0.01
−0.07	0.24	0.02	0.21	0.32	0.11	0.48	1.00	0.03	0.01	0.06
0.10	0.07	0.06	0.18	0.04	−0.05	0.05	0.03	1.00	0.11	−0.05
0.05	−0.09	−0.03	0.23	−0.11	−0.09	−0.13	0.01	0.11	1.00	0.21
0.02	−0.01	−0.10	0.20	0.12	0.02	0.01	0.06	−0.05	0.21	1.00
−0.14	0.09	−0.15	0.22	0.10	−0.10	0.02	0.12	−0.04	0.14	0.40
−0.05	0.19	−0.11	0.26	0.12	−0.00	0.26	0.25	−0.03	0.14	0.36
0.03	−0.18	0.01	−0.06	−0.21	−0.06	−0.27	−0.17	−0.17	−0.00	−0.05
−0.03	−0.24	−0.02	−0.07	−0.23	−0.19	−0.33	−0.19	−0.07	0.01	−0.06
−0.00	−0.16	0.02	−0.11	−0.27	−0.08	−0.21	−0.15	−0.11	0.03	−0.10

Tafla: Einstaklingshreinsað mat á fylgni svara.

$$y_{ij} = \begin{cases} 1 & -\infty < y_{ij}^* < c_1 \\ \vdots & \vdots \\ 5 & c_4 \leq y_{ij}^* < \infty \end{cases}$$

$$y_i^* = (y_{i1}, \ldots, y_{iJ})', \quad J = 16.$$

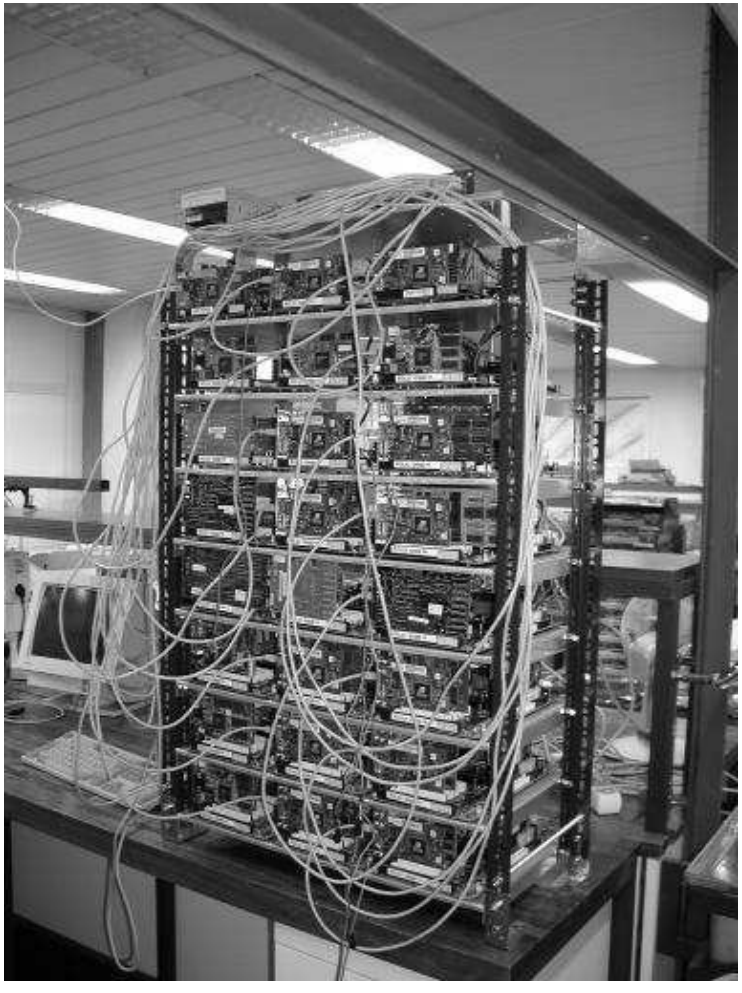
$$y_i^* \sim (\mu_i^*, \Sigma_i^*).$$

Til að gera stika greinanlega eru settar ýmsar skorður, t.d.:

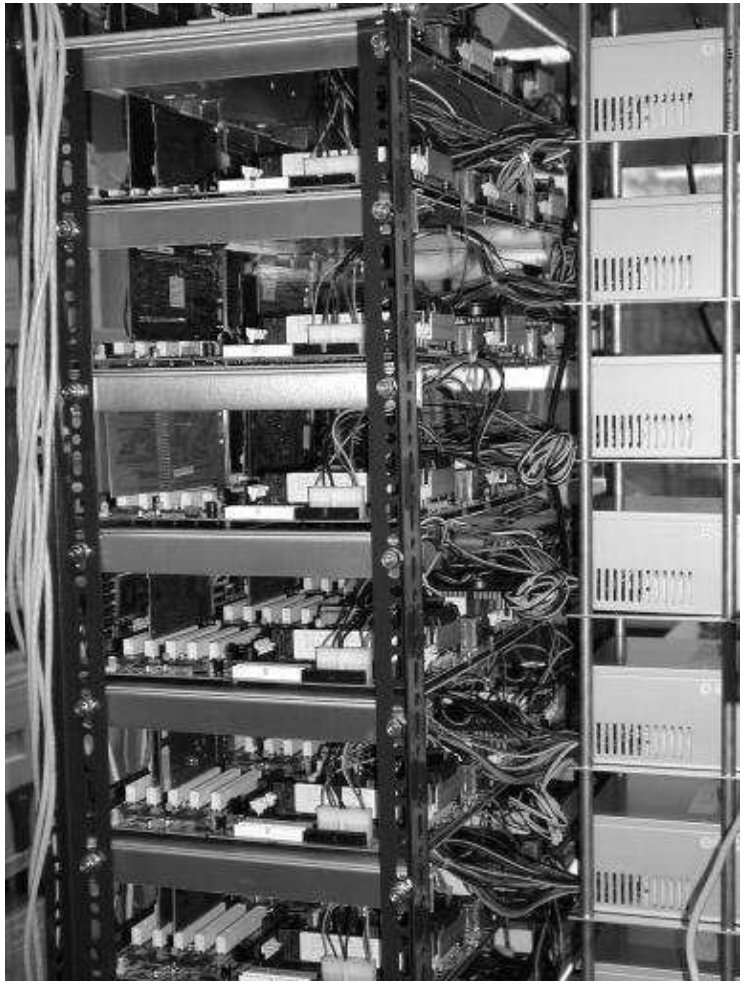
$$\mu_i^* = \mu + \tau_i + \sigma_i z_i, \quad z_i \sim N(0, \Sigma)$$

þar sem z_i er 16 víð ómæld breyta með samdreifnifylki Σ. Gibbs sampler er settur upp í nokkrum þrepum. Þrepin eru valin þannig að skilyrtar dreifingar verði þægilegar. T.d. er gert ráð fyrir að dreifing μ gefið annað í líkani sé normal, dreifing fylkisins Σ sé inverse-Wishart. Inverse-Wishart hendingin er einskonar margvíð χ^2 dreifing og eru gildi hennar fylki.

HIP Software and Physics project



The 24 nodes of the Celeron ATX blade server [1].



Here the power supply tower can be seen.

Lítið kvæði

I thought inference was just a fairy tale,
Confused by stats and probability,
Frequentist approaches (doo-doot doo-doot)
made no sense to me (doo-doot doo-doot)
Summarizing evidence by p?!
Then I saw Tom Bayes – Now I'm a believer,
Without a trace – of doubt in my mind,
[I'm a] Bayesian (ooooh) – Oh, I'm a believer
I couldn't p now if I tried!

Lokaorð

- Það erfiða (vísindalega) er að velja líkan, prior og loss-function.
- Tæknilega hefur verið erfitt að koma við hagnýtri bayesískr tölfræðivinnu. Það hefur nú breyst með tölvutækninni og MCMC hermunaraðferðum.
- Túlkun Bayesista hefur verið önnur en „frequentista," eða „sampling-teoríu" fólks.
- Margir „frequentistar" segja: Mér er alveg sama um heimspekina, MCMC-tæknin býður upp á möguleika á að meta flókin líkön og því vil ég kynna mér hana.
- Trúarbragðadeilur frequentista og Bayesista ligga að mestu niðri.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 113–147.

Geisser, S. (1984). On prior distributions for binary trials. *The American Statistician*, *38*(4), 244–247.

Zellner, A. (1977). Maximal data information prior distributions. In *New Developments in the Application of Bayesian Methods*. Amsterdam: North-Holland.