Statistical inference

or

Some (out-of-)bed time stories

about $\log p(x|\theta)$ and its derivatives

Jean-François Cardoso.
CNRS-LTCI / UP7-APC

.

Outline

- Introduction (selected topics)
- Some simple parametric models, some less simple, all $p(X|\theta)$.
- Parameter estimation:
  estimators, consistency, efficiency, Fisher information, the Cramér-Rao bound, sufficient statistics.
- Exponential families (the wonderful world of).
- A bit of asymptotics
- Information geometry (the big picture)

Warning

Warning:

This is an unfinished set of slides.

And there must be quite a few random bugs.

Your brain on drugs in three easy steps

1) In your computer:

$$X = \begin{bmatrix} .23 & .45 & .67 & .09 & .90 & .32 & .73 & .55 \\ .98 & .11 & .22 & .33 & .41 & .31 & .53 & .03 \end{bmatrix} \in \mathcal{X} = \mathbb{R}^{2\times 8}$$

2) On your screen, look at your data.

In any possible way.

Maybe decide that the data should be modeled as the realization of some random process.

3) In your brain: build a statistical model (or several of them)

$\mathcal{M} = \{p(x)\}$ A set of probability distributions

We will mostly focus on regular statistical models

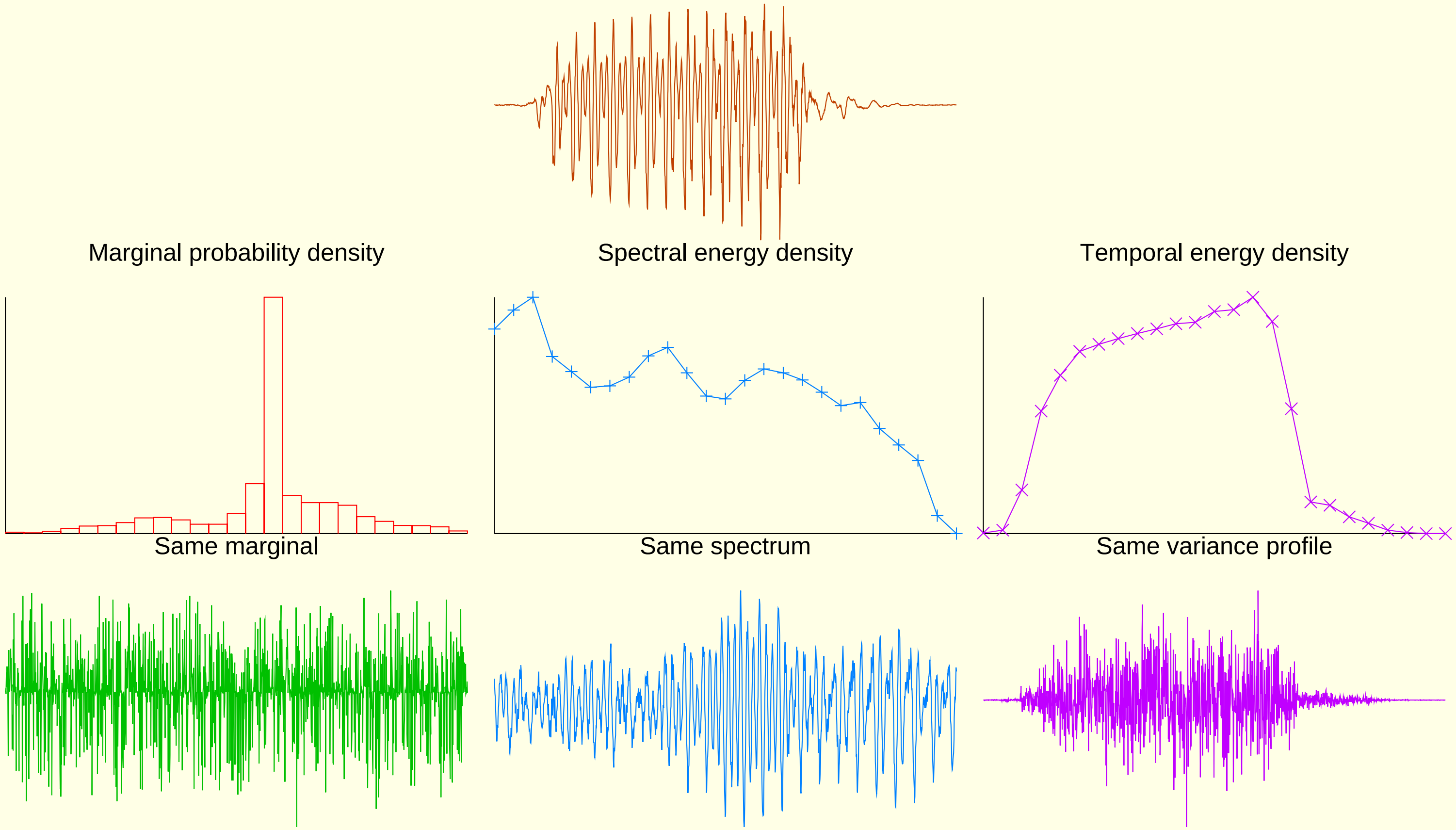
$$\mathcal{M} = \{p(x;\theta),\ \theta \in \Theta \subset \mathbb{R}^d\}$$

where $p(x;\theta)$ is a a smooth function of θ.

The most important slide of this talk

All models are wrong

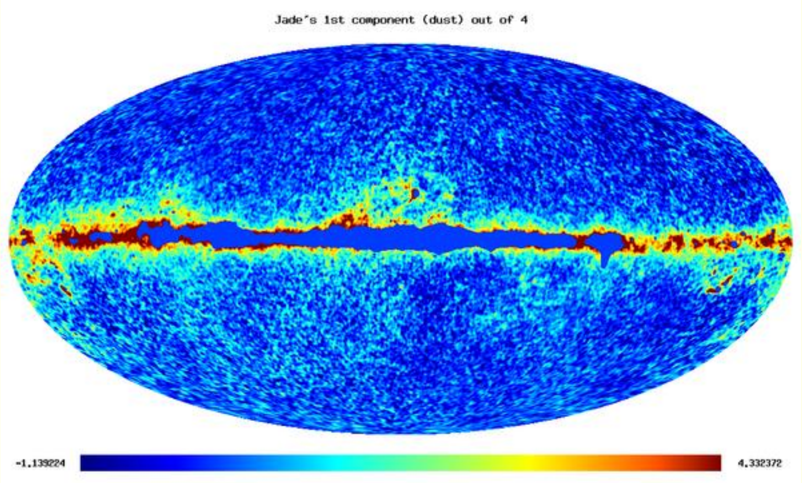
All models are wrong, but some are useful. George Box.

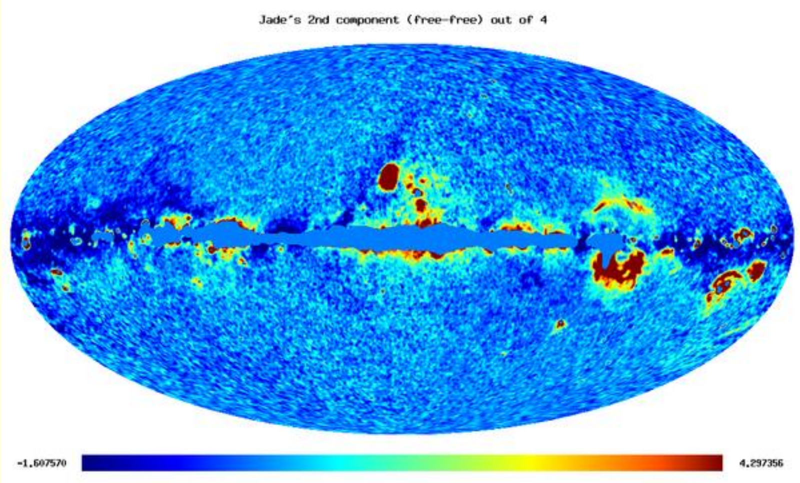


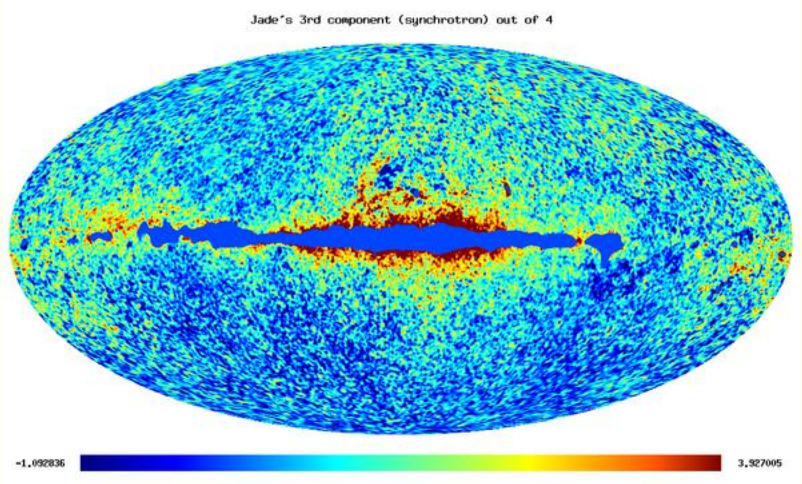
Three points of view on a time series. What is the right statistics?

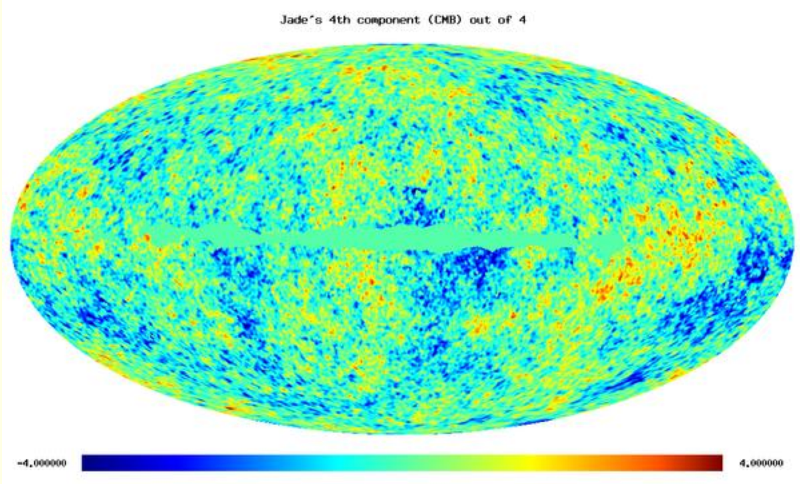
Simple jading of W-MAP 3-year

Blindly looking for four components which are:
1) uncorrelated 2) as independent as possible and 3) modelled as i.i.d.









yields promising results. . .

Numerical work by Frédéric Guilloux.

It goes without saying

Assume basic notions of probability:
probability distributions $p(X)$, expectation $\mathbb{E}X$,
joint and conditional distributions $p(X,Y) = p(X|Y)p(Y) = p(Y|X)p(X)$.

For two random column vectors, define the covariance matrix:

$$\mathsf{Cov}(X,Y) = \mathbb{E}(XY^\dagger) - \mathbb{E}(X)\mathbb{E}(Y)^\dagger \quad \text{Notation: } \mathsf{Cov}(X) = \mathsf{Cov}(X,X)$$

Linearity $\mathsf{Cov}(\mathbf{A}X) = \mathbf{A}\mathsf{Cov}(X)\mathbf{A}^\dagger$.

A symmetric matrix $\mathbf{Q}$ is said to be positive: $\mathbf{Q} \geq 0$ if $Z^\dagger \mathbf{Q} Z \geq 0$ for any Z.

A covariance matrix is positive since $Z^\dagger \mathsf{Cov}(X) Z = \mathsf{Var}(Z^\dagger X) \geq 0$.

For parametric models $p(X|\theta)$:

$$\mathbb{E}_\theta X = \mathbb{E}_\theta(X) = \int X p(X|\theta) dX, \quad \mathsf{Cov}_\theta(X) = \mathbb{E}_\theta X X^\dagger - \mathbb{E}_\theta X \mathbb{E}_\theta X^\dagger.$$

Statistical models

Outline:

- Univariate
- Multivariate
- Time series, parametric
- Stationary fields

Some simple statistical models. 1

Well known families

- Gaussian (or normal) distribution
- χ_p^2
- Exponential
- Poisson
- Multinomial
- $<$Some British guy$>$ distribution . . .

Some simple statistical models. 2

- *Transformation models*

  - Location model: $X = \mu + N \quad \mu \in \mathbb{R}^d$ and $N \sim p_N$

    $$p(X|\theta) = p_N(X - \mu) \quad \theta = \{\mu\}$$

  - Scale model $X = \sigma N \quad \sigma \in \mathbb{R}$

    $$p(X|\theta) = p_N(X/\sigma)/\sigma \quad \theta = \{\sigma\}$$

  - Location-scale model: $X = \mu + \sigma N$

    $$p(X|\theta) = p_N((X - \mu)/\sigma)/\sigma \quad \theta = \{\mu, \sigma\}$$

- Contamination $X = Y + \alpha Z$ for independent Y and Z random variables $\theta = \{\alpha\}$.

- Including the distribution p_N into the unknown parameter yields *semi-parametric* models where θ now is infinite-dimensional *e.g.* $\theta = \{\mu, \sigma, p_N\}$.

Some simple statistical models. 3

Models for an $m \times 1$ vector in terms of q (noisy) factors:
$X = \mathbf{A}S + N$ with an $m \times q$ matrix $\mathbf{A}$ and
(usually) uncorrelated factors: $\text{Cov}(S) = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ and
uncorrelated noise $\text{Cov}(N) = \text{diag}(p_1, \ldots, p_m)$
(but all kinds of perversions are to be found).

• Principal component analysis: Orthogonal factors $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}_m$ and no noise.
So $\theta = (\mathbf{A}, \{\sigma_i^2\})$.

• Factor analysis: Matrix $\mathbf{A}$ is known.
The variances are to be found: $\theta = (\{\sigma_i^2\}, \{p_j^2\})$.
Interesting in conjunction with factor selection.

• (regular) Independent component analysis: matrix $\mathbf{A}$ is unknown but S_i is independent from S_j for all $i \neq j$. Needs non-Gaussianity!

• Direction finding. Uses a physical model to connect the direction α of an impinging wave to the corresponding column so $\theta = (\{\alpha_i\}, \text{Cov}(N))$ with $\mathbf{A} = [a(\alpha_1), \ldots, a(\alpha_q)]$.

Time series 1. Deterministic signal in random noise

We build a model $p(X|\theta)$ for a sequence $X = \{X(1), X(2), \ldots, X(n)\}$ by assuming a determistic signal in noise: $X(i) = S(i) + N(i)$.

- For instance, for the signal part:

$$S(t) = S(t; \theta_S) = \sum_{p=1}^{P} a_p \cos(\omega_p t + \phi_p)$$

Then $\theta_S = \{a_p, \omega_p, \phi_p\}$ (or some subset of it).

- For instance for the noise part:

  - $\{N(t)\}$ is i.i.d. with scale σ: $p_N(N(1), \ldots N(n)) = \prod_{t=1}^{n} \frac{1}{\sigma} q(\frac{N(t)}{\sigma})$. Then $\theta_N = \{\sigma, q\}$ or $\theta_N = \{\sigma\}$.

  - $\{N(t)\}$ is zero-mean Gaussian stationary with correlation $\mathbb{E} N_t N_{t'} = \rho(t' - t)$. Then $\theta_N = \{\rho(\tau)\}$.

- Combining the deterministic and the stochastic parts: $\theta = (\theta_S, \theta_N)$

$$p_X(X|\theta) = p_X(X|\theta_S, \theta_N) = P_N(X - S(\theta_S); \theta_N)$$

Time series 2. Parametric stationary models

- Auto-regressive (AR) model:

$$X(t) = \sum_{\ell=1}^{L} a_\ell X(t-\ell) + \sigma N(t) \quad \theta = \{a_1, a_2, \dots, a_L, \sigma\}$$

where $\{N(t)\}$ an i.i.d. zero-mean unit-variance Gaussian sequence.

- Linear model

$$X(t) = \sum_{t_1 \le t' \le t_2} h(t') N(t-t') \quad \theta = \{\{h(t)\}, p_N(\cdot)\}$$

- A whole zoology

Like, you know, heteroscedastic models (*i.e.* an AR model where $\sigma^2 = \sigma^2(t)$ now is a weigthed average of the past values of $X(t)^2$.

Non parametric stationary models

- Stationary time series: the second-order structure:

$$X = \{X(t)\}_{t=1}^{T} \quad \mathbb{E}x(t) = 0 \quad \mathbb{E}x(t)x(t') = \rho(t'-t) \quad \rho(\tau) = \int e^{i\omega\tau} P(\omega) d\omega$$

- Gaussian stationary field on the sphere $\{X(\xi), \xi \in S^2\}$.

$$X(\xi) = \sum_{\ell \geq 0} X^{(\ell)}(\xi) \quad \widehat{C}_\ell = \frac{\|X^{(\ell)}\|^2}{2\ell+1} \quad C_\ell = \mathbb{E}\widehat{C}_\ell \qquad \text{harmonic spectrum}$$

$p(X)$ depends only on the harmonic spectrum and on its empirical value:

$$p(X|\theta) = \exp -\frac{1}{2} \sum_{\ell \geq 0} (2\ell+1) \left(\frac{\widehat{C}_\ell}{C_\ell} + \log C_\ell \right) + \text{cst} \quad \theta = \{C_\ell, \ell \geq 0\}$$

- Poisson processes, Markov fields, multi-scale models, wavelet models...

The sad truth about parameters

- A vector $\theta \in \mathbb{R}^d$ is just a (continuous) label to a probability distribution $p(\cdot|\theta)$ (think GR).

- The model *is* the manifold.

  $$\mathcal{M} = \{p(\cdot|\theta),\ \theta \in \Theta \in \mathbb{R}^d\}$$

  and it can be smoothly reparameterized in infinitely many ways.

- Hence, parameterization often is arbitrary to some large extent.

- Q: Is there a best parameterization?
  A: Yes, for some models which have canonical parameters.

- Later: The (differential) geometry of statistical models.

Estimation

- Warning: this is mostly the frequentist story.

- Estimation, estimators, estimates
- Method of moments
- Cramér-Rao bound
- Fisher efficiency
- Maximum likelihood
- Sufficient statistics

Parametric estimation

Once we have selected a parametric model $p(X|\theta)$, we need to adjust the model to the data.
Meaning: find the 'best' (?) parameter value $\widehat{\theta}$ given the available data X.

An estimator is a function $T : \mathcal{X} \mapsto \Theta$.
Notation $\widehat{\theta} = T(X)$.

Unbiasedness: $\mathbb{E}_\theta T(X) = \theta$.

Dispersion: $\mathsf{Cov}_\theta(T(X))$.

Important note:
unbiasedness and accuracy should not be taken too seriously on a manifold because parameterization is arbitrary. What do the expectation and the covariance mean when, for instance, θ parameterizes a rotation matrix?

The method of moments / least squares

Let $\widehat{S} = \widehat{S}(x)$ be a q-valued statistic: $\widehat{S} : \mathcal{X} \mapsto \mathbb{R}^q$ whose expected value under $p(x|\theta)$ is a known (meaning: computable) function of θ:

$$\mathbb{E}_\theta \widehat{S}(x) = S_\theta$$

– If $q = \dim(\theta)$, the *method of moments* estimates θ by $\widehat{\theta}$ such that

$$S_{\widehat{\theta}} = \widehat{S}(x)$$

– If $q > \dim(\theta)$, the *method of moments* estimates θ by finding the best match

$$\widehat{\theta} = \arg\min_\theta \phi(x;\theta) \quad \phi(x;\theta) = \|S_\theta - \widehat{S}(x)\|^2$$

A better estimator may be obtained using a (positive) weighting matrix W

$$\phi(x;\theta) = (S_\theta - \widehat{S}(x))^\dagger W (S_\theta - \widehat{S}(x))$$

An even better estimator may be obtained with a parameter dependent weight $W = W_\theta$.

Moment/LS square methods require only $\mathbb{E}_\theta \widehat{S}(X)$ and possibly $\mathrm{Cov}_\theta \widehat{S}(X)$ but *not* the full distribution $p(x|\theta)$.

Can least-squares beat $\phi(x;\theta) = -\log p(x|\theta)$?

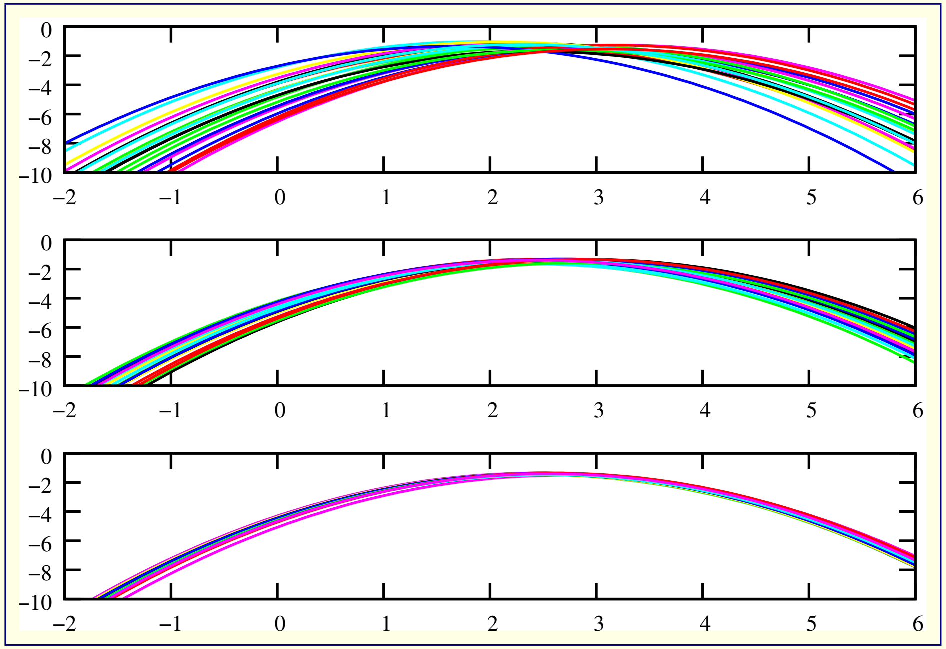
Likelihood

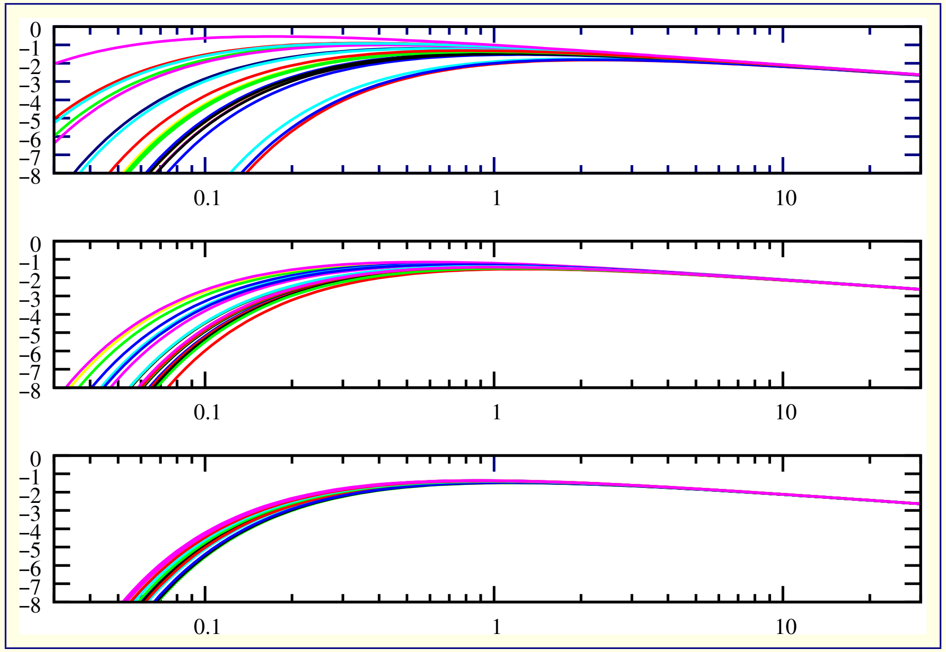
A likelihood:

- Data: X is n i.i.d. $\mathcal{N}(\mu_\star, \sigma^2_\star)$ samples. Top to bottom: $n = 3, 30, 300$

- Model: 'true' model.

– Left: $\phi(\mu) = \frac{1}{n} \log p(X|\mu, \sigma^2 = \sigma^2_\star)$.
– Right: $\phi(\sigma^2) = \frac{1}{n} \log p(X|\mu = \mu_\star, \sigma^2)$.



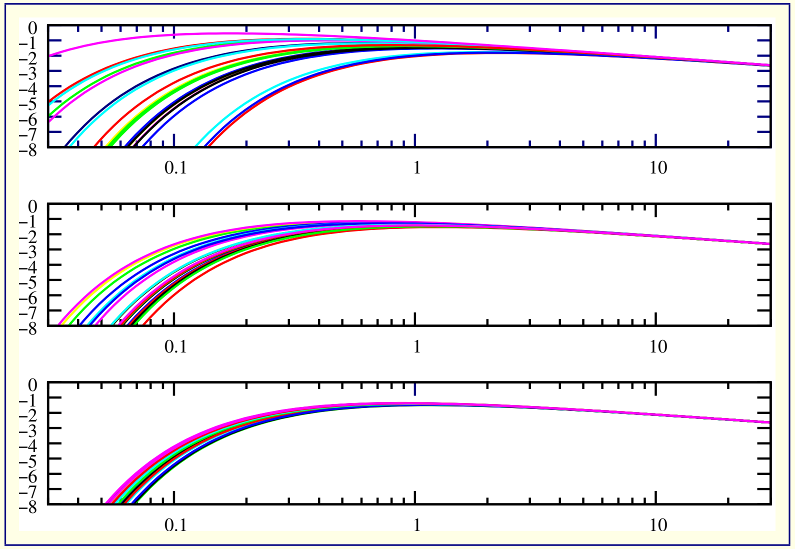


The likelihood is $p(x|\theta)$ seen as a function of the parameter vector.
Given the data x and lacking prior information, *this is all we have.*

Legitimate questions about the (log)-likelihood

What is the meaning of

- the maximum value of the the likelihood,
- the value of θ which maximizes it,
- the dispersion of the latter,
- the width of the likelihood peak,
- the general shape of the likelihood function?



Note on Bayes: there is a transparent interpretation of the likelihood function when a prior distribution $\pi(\theta)$ on θ is available. By the Bayes theorem

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta')\pi(\theta')d\theta'}$$

Hence, the shape of the log-likelihood $\log p(x|\theta)$ is the shape of the log-posterior distribution $\log p(\theta|x)$ if the prior distribution $\pi(\theta)$ is uniform.

1) In spite of the maths, likelihood analysis is *not* Bayesian with a flat prior!
2) WHAT DO YOU MEAN “UNIFORM” ?

Intermezzo

Plausible definitions of the "straight segment" from one probability distribution $p_a(x)$ to another $p_b(x)$?

The *mixture segment* makes some statistical sense:

$$p(x|\alpha) = (1-\alpha)p_a(x) + \alpha\, p_b(x)$$

The *exponential segment* also seems pretty darn reasonable

$$\log p(x|\alpha) = (1-\alpha)\log p_a(x) + \alpha \log p_b(x) - \psi(\alpha)$$

with $\psi(\alpha)$ for normalization.

In 'traditional' exponential form

$$p(x|\alpha) = p_a(x)e^{\alpha S(x) - \psi(\alpha)} \quad S(x) = \log\frac{p_b(x)}{p_a(x)}$$

Score and consequences

The log-likelihood: $\ell(x|\theta) \stackrel{\text{def}}{=} \log p(x|\theta)$ is a very interesting *random function*. Its derivatives with respect to θ, even more so.

For a d-dimensional model ($\theta \in \mathbb{R}^d$), define

$$\partial\ell(x|\theta) = \frac{\partial \log p(x|\theta)}{\partial\theta} \qquad \text{random } d \times 1 \text{ vector: the score}$$

$$\partial^2\ell(x|\theta) = \frac{\partial^2 \log p(x|\theta)}{\partial\theta^2} \qquad \text{random } d \times d \text{ matrix}$$

The score function has zero mean under $p(x|\theta)$:

$$\mathbb{E}_\theta \partial\ell(x|\theta) \stackrel{a}{=} 0$$

and its covariance is called the *Fisher information matrix*

$$\mathbf{F}_\theta \stackrel{\text{def}}{=} \mathbb{E}_\theta \left(\partial\ell(x|\theta)\partial\ell(x|\theta)^\dagger\right) \stackrel{b}{=} -\mathbb{E}_\theta \partial^2\ell(x|\theta)$$

Properties a and b stem from $\int p(x|\theta)dx = 1$.

Unbiased estimation and the score

For an estimator $\widehat{\theta} = T(x)$, differentiating the unbiasedness condition

$$\mathbb{E}_\theta T(x) = \theta$$

with respect to θ yields the covariance between two zero-mean random vectors $T(x) - \theta = \widehat{\theta} - \theta$ and $\partial \ell(x|\theta)$:

$$\mathrm{Cov}_\theta \left(\partial \ell(x|\theta), \widehat{\theta} - \theta \right) = \mathbf{I} \qquad \text{(the identity matrix)}$$

Hence an unbiased estimator $T(x)$ necessarily has a very specific correlation to the score function and we have

$$\mathrm{Cov}_\theta \left(\begin{bmatrix} \widehat{\theta} \\ \partial \ell(x|\theta) \end{bmatrix} \right) = \begin{bmatrix} \mathrm{Cov}(\widehat{\theta}) & \mathbf{I} \\ \mathbf{I} & \mathbf{F}_\theta \end{bmatrix}$$

Since a covariance matrix must be positive, it must hold that

$$\mathrm{Cov}_\theta(\widehat{\theta}) \geq \mathbf{F}_\theta^{-1}$$

The amazing CRB

Covariance matrices are positive. In particular

$$\text{Cov}_\theta\left(\widehat{\theta} - \theta - \mathbf{F}_\theta^{-1}\partial\ell(x|\theta)\right) \geq 0$$

Expand this covariance matrix, recalling that $\text{Cov}_\theta\left(\partial\ell(x|\theta), T(X)\right) = I$:

$$\begin{aligned}
\text{Cov}_\theta\left(\widehat{\theta} - \theta - \mathbf{F}_\theta^{-1}\partial\ell\right) &= \text{Cov}(\widehat{\theta}) + \text{Cov}(\mathbf{F}_\theta^{-1}\partial\ell) - \mathbb{E}\left(\mathbf{F}_\theta^{-1}\partial\ell, \widehat{\theta} - \theta\right) - \text{symm} \\
&= \text{Cov}(\widehat{\theta}) + \mathbf{F}_\theta^{-1}\text{Cov}(\partial\ell)\mathbf{F}_\theta^{-1} - \mathbf{F}_\theta^{-1}\mathbb{E}\left(\partial\ell, \widehat{\theta} - \theta\right) - \text{symm} \\
&= \text{Cov}(\widehat{\theta}) + \mathbf{F}_\theta^{-1}\mathbf{F}_\theta\mathbf{F}_\theta^{-1} - \mathbf{F}_\theta^{-1}\mathbf{I} - \mathbf{I}\mathbf{F}_\theta^{-1} \\
&= \text{Cov}(\widehat{\theta}) - \mathbf{F}_\theta^{-1}
\end{aligned}$$

Therefore an unbiased estimator cannot have arbitrarily small variance:

$$\text{Cov}_\theta(\widehat{\theta}) \geq \mathbf{F}_\theta^{-1} \quad \text{(Fréchet-Darmois)-Cramér-Rao bound}$$

Remember it is a *matrix* inequality. We may look at individual entries:

$$\text{Cov}_\theta(\widehat{\theta}_i) \geq [\mathbf{F}_\theta^{-1}]_{ii} \geq [\mathbf{F}_\theta]_{ii}^{-1}$$

The last inequality is ‘statistically obvious’.

Fisher information and efficiency

- (again) An unbiased estimator cannot have arbitrarily small variance:

  $$\mathrm{Cov}_\theta(\widehat{\theta}) \geq \mathbf{F}_\theta^{-1} \quad \text{Cramér-Rao bound} = \text{CRB}$$

  and also nothing can travel faster than light.

- **Breakthrough**
  – Our statistical model $\mathcal{M} = \{p(x|\theta)\}$ seen as a manifold is given a natural metric by the FIM matrix $\mathbf{F}_\theta$.
  – Even better: it gives the statistical resolution cell.
  – Also, there *does* exist a canonical prior: Jeffreys prior which gives the same prior weight to all resolution cells. This construction is parameter independent.

- **Definition**: An estimator reaching the CRB is called (Fisher)-*efficient*.

- **Question**: Do efficient estimators exist? Model dependent ?

Maximum likelihood

Note An efficient estimator (if it exists) has no choice. It must behave as

$$T(x) = \widehat{\theta} = \theta + \mathbf{F}_\theta^{-1} \partial \ell(x|\theta)$$

Recalling $\mathbf{F}_\theta = -\mathbb{E}_\theta \partial^2 \ell(x|\theta)$, this looks very very much like a Newton step...

This suggests estimating θ as the most likely parameter *i.e.*

$$\widehat{\theta}_{\mathrm{ML}} \stackrel{\mathrm{def}}{=} \arg\max_\theta \ell(x|\theta)$$

This is a solution of

$$\partial \ell(x|\widehat{\theta}_{\mathrm{ML}}) = 0$$

Compare to a key property of the score:

$$\mathbb{E}_\theta \partial \ell(x|\theta) = 0$$

- **Note**. The ML estimate is perfectly invariant under re-parameterization.

- **Question**. The ML estimate is a least-square fit when the model is a deterministic signal in Gaussian noise. How to understand $-\log p(x|\theta)$ as a measure of mismatch between model data *in general*?

A detour

The next two slides give a quick view of likelihood for *discrete* valued data.

Discrete random variables are easy to deal with because the probability distribution π of a d-valued random variable is specified by d numbers $\pi = (\pi_1, \ldots, \pi_d)$.

Hence, we can always picture the set of all probability distributions of a d-valued variable as the simplex:

$$\mathcal{S} = \{\pi = (\pi_1, \ldots, \pi_d),\ \pi_j \geq 0,\ \sum_{j=1}^{d} \pi_j = 1\}$$

In the discrete case, several important concepts show up right away. After enlightenment from the discrete world, we return to the general case.

Likelihood for discrete data

Take x a discrete variable taking d possible values with probability $\pi = (\pi_1, \ldots, \pi_d)$. The probability of a sequence $(x_1, \ldots, x_n)$ modeled as i.i.d. is

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n) = \prod_{j=1}^{d} \pi_j^{n_j}$$

where n_j is the number of occurences of the j-th symbol in the sequence. So

$$\log p(x_1, \ldots, x_n) = \sum_j n_j \log \pi_j = n \sum_j \widehat{\pi}_j \log \pi_j \quad \text{where} \quad \widehat{\pi}_j \stackrel{\text{def}}{=} \frac{n_j}{n}$$

$$= -n \sum_j \widehat{\pi}_j \log \frac{\widehat{\pi}_j}{\pi_j} + n \sum_j \widehat{\pi}_j \log \widehat{\pi}_j$$

Hence

$$p(x_1, \ldots, x_n | \pi) = e^{-nK[\widehat{\pi}, \pi]} e^{-nH[\widehat{\pi}]}$$

with

$$K[p, q] \stackrel{\text{def}}{=} \sum_j p_j \log \frac{p_j}{q_j} \quad \text{Kullback divergence from } p \text{ to } q$$

$$H[p] \stackrel{\text{def}}{=} -\sum_j p_j \log p_j \quad \text{(Shannon) entropy of } q$$

Likelihood for discrete data (cont.)

x a discrete variable taking d possible values with probability $\pi = (\pi_1, \ldots, \pi_d)$.

Again, the probability of an i.i.d. n-sequence depends only on $\widehat{\pi} = [\frac{n_1}{n}, \ldots, \frac{n_d}{n}]$:

$$-\frac{1}{n} \log p(x_1, \ldots, x_n) = K[\widehat{\pi}, \pi] + H[\widehat{\pi}] \quad \text{with} \begin{cases} K[p,q] & = \sum_j p_j \log \frac{p_j}{q_j} \\ H[q] & = -\sum_j q_j \log q_j \end{cases}$$

→ Note: the number of sequences with $\widehat{\pi}$ roughly is $\exp nH[\widehat{\pi}]$.

The empirical distribution $\widehat{\pi}$ is an *exhaustive statistic*:
→ Statistical compression, a.k.a. "Keep $\widehat{\pi}$, trash your data."

The Kullback divergence $K[p, q]$ is positive unless $p = q$. It is a (non-symmetric) measure of mismatch between two probability distributions.

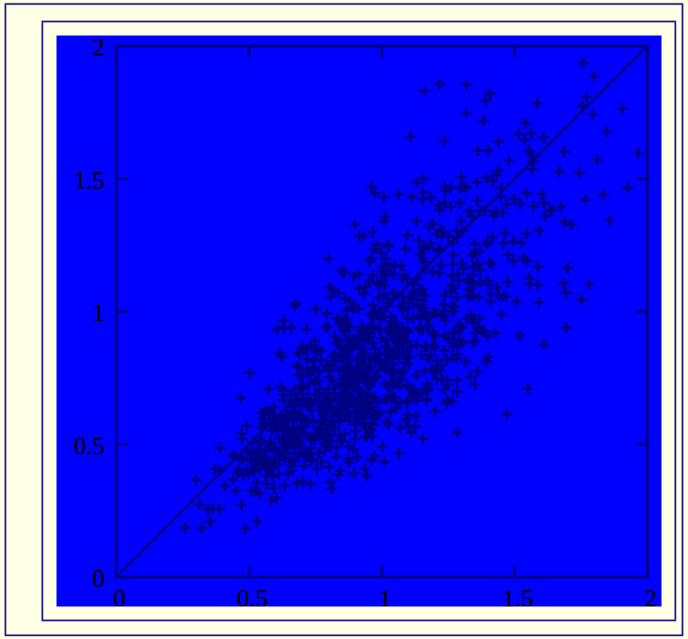
For a parametric model $\mathcal{M} = \{\pi = \pi(\theta), \theta \in \Theta\}$, nothing really changes:

$$-\frac{1}{n} \log p(x_1, \ldots, x_n | \theta) = K[\widehat{\pi}, \pi(\theta)] + H[\widehat{\pi}] \qquad \text{ML = Kullback matching!}$$

Sufficiency. 1

Let T values $x_1, \ldots, x_T$ be modeled as i.i.d. Laplace:

$$p(x|\theta) = \frac{1}{\theta} \exp -\frac{x}{\theta} \quad x \geq 0$$



Since $\mathbb{E}_\theta x = \theta$ and $\mathbb{E}_\theta x^2 = 2\theta^2$, two possible estimates of θ are

$$\widehat{\theta}_1 = \frac{1}{T} \sum_{t=1}^{T} x_t$$

$$\widehat{\theta}_2 = \left(\frac{1}{2T} \sum_{t=1}^{T} x_t^2 \right)^{\frac{1}{2}}$$

Plot: many realizations of $\widehat{\theta}_1$ versus $\widehat{\theta}_2$.

Question: Among all these statistics

$$\widehat{\theta}_1, \quad \widehat{\theta}_2, \quad \widehat{\theta}_3 = (\widehat{\theta}_1 + \widehat{\theta}_2)/2, \quad \widehat{\theta}_4 = \frac{\sigma_1^{-2}\widehat{\theta}_1 + \sigma_2^{-2}\widehat{\theta}_2}{\sigma_1^{-2} + \sigma_2^{-2}} \quad \text{with} \quad \sigma_i^2 = \mathrm{Var}(\widehat{\theta}_i)$$

(which also are estimates of θ) which one (or which combination thereof) contains the most information about the scale parameter θ?

Sufficiency. (cont. example)

If T values $x_1, \ldots, x_T$ are modeled as i.i.d. realizations of an exponential distribution $p(x|\theta) = \frac{1}{\theta} \exp -\frac{x}{\theta}$ and we define the statistics

$$\widehat{\theta}_1 = \frac{1}{T} \sum_{t=1}^{T} x_t \quad \widehat{\theta}_2 = \left(\frac{1}{2T} \sum_{t=1}^{T} x_t^2 \right)^{\frac{1}{2}}$$

then, there is *no information* about θ in $\widehat{\theta}_2$ in addition to $\widehat{\theta}_1$!

$$\begin{aligned} p(\widehat{\theta}_1, \widehat{\theta}_2|\theta) &= p(\widehat{\theta}_2|\widehat{\theta}_1, \theta) p(\widehat{\theta}_1|\theta) && \text{Conditioning, always true} \\ &= p(\widehat{\theta}_2|\widehat{\theta}_1) p(\widehat{\theta}_1|\theta) && \text{Factorization theorem (next slide)} \end{aligned}$$

i.e. the distribution of $\widehat{\theta}_2$ given that $\widehat{\theta}_1$ has been observed does *not* depend on the unknown parameter θ. This is because $\widehat{\theta}_1$ is a *sufficient statistic*.

If you know how to extract optimally information from $\widehat{\theta}_1$, there is no need to involve $\widehat{\theta}_2$.

The French call it "statistique exhaustive" which is pas mal non plus.

$\log p(x_1, \ldots, x_T|\theta) = -T(\frac{\widehat{\theta}_1}{\theta} + \log \theta)$

Sufficiency

Official definition: $S(x)$ is a sufficient statistic for the model $p(x|\theta)$ if the distribution of x conditionned on the observation of $S(x)$ does not depend on θ:

$$p(x|S(x), \theta) = p(x|S(x))$$

This is equivalent (theorem) to the factorization property:
there exist some functions g and h such that:

$$p(x|\theta) = g(x)\, h(S(x); \theta)$$

In effect, $S(x)$ exhausts the information in x since the likelihood of θ depends on x only through $S(x)$.

Of course, $S(x) = x$ always is a sufficient statistic. A sufficient statistic is interesting if it does compress the data in the sense that $\dim(S(x)) < \dim(x)\ldots$

. . . or maybe also if it makes our life easier. As in exponential models.

Exponential models

Outline

- Definition
- Examples
- Some stupendous properties
- Maximum likelihood estimation within exponential models
- Convexity and duality
- More stupendous properties
- Connection to maximum entropy

Exponential families: Informal definition

When a statistical model $\mathcal{M} = \{p(x|\theta);\ \theta \in \Theta\}$ admits a sufficent statistic $S(x)$, one has (by definition) the form

$$\log p(x|\theta) = g(x) + h(S(x); \theta)$$

"Exponential families" have an even more favorable form. By definition, an exponential model has, possibly after some serious massaging of both x and θ, the structure

$$\log p(x|\theta) = g(x) + h(\theta) + S(x)^{\dagger}\theta$$

that is: the part of the log-density which connects the variable x and the parameter θ is just their scalar product.

Exponential families

If a family $\mathcal{M}$ of probability distributions can be parameterized by a d-dimensional vector $\theta \in \Theta \subset \mathbb{R}^d$ in such a way that

$$p(x|\theta) = g(x)\, e^{S(x)^{\dagger}\theta - \psi(\theta)}$$

i) using a function $S : \mathcal{X} \mapsto \mathbb{R}^d$,
ii) a function $\psi : \Theta \subset \mathbb{R}^d \mapsto \mathbb{R}$
iii) using a measure $g(x)$
then,

- $\mathcal{M}$ is said to be an *exponential family* (of probability distributions),
- $S(x)$ is a sufficient statistic,
- θ is a canonical parameter,
- we are happy.

Is that too much too ask?

Let's see why we are happy and when such a happiness is possible.

Who's exponential?

- Many 'standard' families can be massaged into exponential form.

- Many other families are 'curved' exponential families, *i.e.* can be naturally embedded into exponential families as $\mathcal{M} = \{p(x|\theta); \theta = \theta(\alpha)\}$.

- Asymptotically (in the number of samples), all regular families are exponential.

- **Note**: Exponentiality is a property of a *family* of distributions; it is *not* the property of a single given distribution. Any single distribution is part of (infinitely many) exponential families.

Some examples

- Example 1: Laplace
- Example 2: Multinomial
- Example 3: Location scale normal
- Example 4: Multivariate Gaussian
- Example 5: Poisson
- Example 6: Beta
- Example 7: Gamma
- Example 8: Dirichlet. . .

Recipe for generating an exponential family

Try this at home:

1) Pick a probability distribution with density $p(x)$

2) Pick a function $S : \mathcal{X} \mapsto \mathbb{R}^p$, $x \rightarrow S(x)$.

3) Define $\psi(\theta) = \log \int p(x) e^{S(x)^\dagger \theta}$ for all $\theta \in \Theta = \{\theta \in \mathbb{R}^q | \int p(x) e^{S(x)^\dagger \theta} < \infty\}$.

4) Enjoy your own home-made, probably p-dimensional, exponential family

$$p(x; \theta) = p(x) e^{S(x)^\dagger \theta - \psi(\theta)} \quad \theta \in \Theta \subset \mathbb{R}^p$$

– Simple example 1: $p(x) = \mathcal{N}(0, \mathbf{I}_m)$ and $S(x) = [\ldots, x_i, \ldots]_{1 \le i \le m}$.

– Simple example 2: $p(x) = \mathcal{N}(0, \mathbf{I}_m)$ and $S(x) = [\ldots, x_i^2 - 1, \ldots]_{1 \le i \le m}$.

– Simple example 3: $p(x) = \mathcal{N}(0, \mathbf{I}_m)$ and $S(x) = [\ldots, x_i x_j, \ldots]_{1 \le i < j \le m}$.

Uniqueness

An exponential family with sufficient statistic $S(x) \in \mathbb{R}^p$ and canonical parameter $\theta \in \mathbb{R}^p$:

$$p(x|\theta) = g(x)\, e^{S(x)^\dagger \theta - \psi(\theta)}$$

If T is an invertible $p \times p$ matrix: $TT^{-1} = I_p$, then the exponential family

$$q(x|\theta) = g(x)\, e^{S(x)^\dagger \theta - \psi(\theta)}$$

Centering

1) Rescaling: for any $\alpha > 0$, if $\bar{g} = \alpha g$ and $\bar{\psi} = \psi + \log\alpha$, then

$$p(x|\theta) = g(x)\, e^{S(x)^\dagger\theta - \psi(\theta)} = \bar{g}(x)\, e^{S(x)^\dagger\theta - \bar{\psi}(\theta)}$$

Thus, if $\int g(x) < \infty$, then $g(x)$ can always be rescaled to sum to be a pdf.

2) Centering the statistic: for any fixed $S_\star$, let $\bar{S}(x) = S(x) - S_\star$ and $\bar{\psi} = \psi - S_\star^\dagger\theta$

$$p(x|\theta) = g(x)\, e^{S(x)^\dagger\theta - \psi(\theta)} = g(x)\, e^{\bar{S}(x)^\dagger\theta - \bar{\psi}(\theta)}$$

Thus, we can shift the statistic as we please and, in particular, we may ensure $E_\theta S(x) = 0$ for some fixed θ.

3) Centering the parameter vector: for any fixed point $\theta_\star$, define $\bar{\theta} = \theta - \theta_\star$, $\bar{g}(x) = g(x) \exp S(x)^\dagger\bar{\theta}$, $\bar{\psi}(\bar{\theta}) = \psi(\bar{\theta} + \theta_\star)$ and check that

$$p(x|\theta) = g(x)\, e^{S(x)^\dagger\theta - \psi(\theta)} = \bar{g}(x)\, e^{S(x)^\dagger\bar{\theta} - \bar{\psi}(\bar{\theta})}$$

Thus, any point can be used as the origin.

Subfamilies

Start with some exponential family $\mathcal{M}$

$$p(x|\theta) = g(x)\, e^{S(x)^\dagger \theta - \psi(\theta)}$$

A q-dimensional subset of Θ defined by some mapping $\theta(\eta) : \mathbb{R}^q \mapsto \mathbb{R}^p$ defines a subfamily

$$q(x|\eta) = p(x|\theta(\eta))$$

If $q > p$, over-parameterization. What's the point?

If $q = p$ and the mapping is invertible, what's the point?

If $q < p$ this defines a *curved family* embedded in the ambient exponential family.

This sub-family is an exponential family itself only when $\theta(\eta) = \mathbf{T}\eta$ for some fixed $p \times q$ matrix $\mathbf{T}$. It then has obviously sufficient statistic $\mathbf{T}^\dagger S(x)$.

Example: binned spectra.

Partition function. First derivative

Consider an exponential family $\mathcal{M}$ parameterized as

$$p(x;\theta) = p(x)e^{S(x)^\dagger\theta - \psi(\theta)}$$

Then the score function splits additively as

$$\frac{\partial \log p(x;\theta)}{\partial\theta} = S(x) - \frac{\partial\psi(\theta)}{\partial\theta}$$

Remember that the score has zero mean under θ. Therefore

$$\frac{\partial\psi(\theta)}{\partial\theta} = \mathbb{E}_\theta S(x)$$

i.e. the first derivative of ψ is the mean value of the sufficient statistic.

We will see shortly that there is a one-to-one mapping between θ and $\mathbb{E}_\theta S(x)$. Thus, we can use it to label any distribution in the family. This is the *dual parameterization* using

$$\eta = \eta(\theta) = \mathbb{E}_\theta S(x) = \frac{\partial\psi(\theta)}{\partial\theta}$$

Partition function. Second derivative

The second derivative of the log-likelihood is

$$\frac{\partial^2 \log p(x;\theta)}{\partial\theta^2} = -\frac{\partial^2\psi(\theta)}{\partial\theta^2}$$

which is *not* random. Thus

$$0 \le \mathbf{F}_\theta = -\mathbb{E}_\theta \partial^2 \ell(x|\theta) = \frac{\partial^2\psi(\theta)}{\partial\theta^2}$$

i.e. the second derivative of ψ is the (positive) Fisher information matrix.

Hence ψ is a convex fonction. Thus, there is a unique point θ where $\psi(\theta)$ takes a given value η of the gradient.

Therefore we can label any distribution in an exponential family either by θ or by $\eta = \mathbb{E}_\theta S(x)$. The two labels θ and η are related by $\eta = \frac{\partial\psi(\theta)}{\partial\theta}$.

Repeat

In an exponential family $\mathcal{M}$

$$p(x|\theta) = g(x)\, e^{S(x)^{\dagger}\theta - \psi(\theta)}$$

- the gradient of the partition function can be used as a dual parameter:

$$\eta = \frac{\partial \psi(\theta)}{\partial \theta}$$

- the Hessian of the partition function is the Fisher information matrix:

$$\frac{\partial^2 \psi(\theta)}{\partial \theta^2} = \mathbf{F}_{\theta}$$

Maximum likelihood estimation in exponential models

Maximum likelihood in the canonical parameter. Recall

$$\frac{\partial \log p(x;\theta)}{\partial \theta} = S(x) - \frac{\partial \psi(\theta)}{\partial \theta}$$

Hence the ML estimate $\widehat{\theta}_{\mathsf{ML}}$ of θ given data x is the solution of

$$\frac{\partial \psi(\widehat{\theta}_{\mathsf{ML}})}{\partial \theta} = S(x)$$

This is not a statistical problem any longer. It is just a matter of inverting the mapping $\theta \to \frac{\partial \psi(\theta)}{\partial \theta}$

This is trivial (void) in the dual parameterization

Recall $\quad \eta \stackrel{\mathsf{def}}{=} \frac{\partial \psi(\theta)}{\partial \theta} \quad$ so that $\quad \widehat{\eta}_{\mathsf{ML}} = S(x)$

Even more obviously

$$\mathbb{E}_{\widehat{\eta}_{\mathsf{ML}}} S(x) = S(x)$$

That is *Under the likeliest distribution, the mean value of the sufficient statistic is equal to the observed value.*

Inverse problems, MaxEnt and Kullback

We want to estimate (the distribution p of) a variable x based on the sole knowledge of the value s of a statistic $S(x)$.

Problem:
if $\dim(x) < \dim(S(x))$, then x is not uniquely determined from $s = S(x)$. Even if $s = S(x)$ is invertible, it may be an ill-posed problem.

The "Maximum entropy on the mean" proposal:
– Select a prior a reference distribution $q(x)$ and estimate p as

$$p_\star = \arg\min_p K[p|q] \quad \text{subject to } \mathbb{E}_p S(x) = s$$

– Optionally, estimate x as $\mathbb{E}_{p_\star} x$.

The solution in terms of an d-dimensional Lagrange multiplier θ:

$$p(x|\theta) = q(x) e^{\theta^\dagger S(x) - \psi(\theta)}$$

The observations define mixture families $\mathcal{M}(s) = \{p(x) \,|\, \mathbb{E}_p S(x) = s\}$.
→ Exponential/mixture foliation.

Asymptotics

- Simple i.i.d. asymptotics makes your life easy
- The two basic convergence theorems for large sample size
- Influence function
- Asymptotic covariance
- Asymptotics for the likelihood
- The MLE is asymptotically efficient

Asymptotics

Somehow, statistics is all about asymptotics because we need repetition. Actually, *non* asymptotics are difficult. Asymptotics makes our life easy.

We consider only simple asymptotics: the observation of n samples $X^n = \{x_1, x_2, \ldots, x_n\}$ assumed to be independently and identically distributed (i.i.d.) according to some distribution in a parametric family $p(x|\theta)$:

$$p(X^n|\theta) = p(x_1, \ldots, x_n|\theta) = \prod_{t=1}^{n} p(x_t|\theta)$$

Denote $\widehat{\theta}^n = \widehat{\theta}^n(X^n)$ an estimate of θ based on n samples.

We hope for, at least, asymptotic unbiasedness: $\mathbb{E}\widehat{\theta}_n \xrightarrow{n\to\infty} \theta$

But we may expect better. *Consistency*: $\widehat{\theta}_n \xrightarrow{n\to\infty} \theta$.

In which sense? At which rate? Asymptotic behavior of the estimate.

Two big shots in asymptopia: LLN and CLT

Big question: what happens to an average

$$\bar{X}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} x_t$$

of n i.i.d. samples $\{x_1, x_2, \ldots, x_n\}$ when the sample size n goes to infinity?

- **LLN** Law of large numbers: $\bar{X}_n \xrightarrow{\mathcal{P}} \mathbb{E}x$ meaning

$$\forall \epsilon > 0 \quad \text{Prob}\left(|\bar{X}_n - \mathbb{E}x| < \epsilon\right) \xrightarrow{n \to \infty} 1$$

- **CLT** Central limit theorem. Zooming in on the convergence

$$\sqrt{n}(\bar{X}_n - \mathbb{E}x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \text{Cov}(x))$$

So, *we have a rate*! Namely: the regular “square root” rate.

Square root consistency.

Influence function

Influence function: A key tool for asymptotics.

Again consider $\widehat{\theta}^n$ an estimator of θ based on n samples $\{x_1, \ldots, x_n\}$.

An *influence function* for the estimator is $f(x;\theta)$ such that

$$\mathbb{E}_\theta f(x;\theta) = 0 \quad \text{and} \quad \widehat{\theta} = \theta + \frac{1}{n}\sum_{t=1}^{n} f(x_t;\theta) + \text{hot}$$

It tells how much each data point is "perturbating" the estimation.

If the data points are i.i.d. then, by the CLT

$$\sqrt{n}(\widehat{\theta}^n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \text{Cov}_\theta(f(x;\theta)))$$

Loosely speaking: the estimation error is asympotically Gaussian with (asymptotic) covariance matrix $\frac{1}{n}\text{Cov}(f)$.

Example

Estimation of the variance σ^2 of a zero mean data set:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_t x_t^2$$

Finding the estimating function

$$\hat{\sigma}^2 = \frac{1}{n}\sum_t (x_t^2 - \sigma^2 + \sigma^2) = \sigma^2 + \frac{1}{n}\sum_t (x_t^2 - \sigma^2)$$

so the influence of x is $f(x;\sigma^2) = x^2 - \sigma^2$.

$\mathrm{Cov}(f) = \mathbb{E}(x^2 - \sigma^2)^2 = \mathbb{E}x^4 - 2\sigma^2\mathbb{E}x^2 + \sigma^4 = 2\sigma^4 + k$ where $k = \mathbb{E}x^4 - 3\mathbb{E}^2x^2$ is the *kurtosis* of x.

The (asymptotic) covariance of the estimation error is

$$\mathrm{Cov}(\hat{\sigma}^2) = \frac{2\sigma^4 + k}{n}$$

It boils down to σ^4/n for Gaussian variables for which $k = 0$ but it can be much larger (how much?) or much smaller (how much?).

Influence of the MLE

The i.i.d. model: $\log p(X^n|\theta) = \sum_{t=1}^n \log p(x_t|\theta)$

The MLE $\widehat{\theta}^n = \arg\max \log p(X^n|\theta)$ is characterized by

$$0 = \sum_{t=1}^n \frac{\partial \log(x_t|\widehat{\theta})}{\partial\theta}$$

First order expansion denoting $\partial\ell(x|\theta) \stackrel{\text{def}}{=} \frac{\partial \log(x|\theta)}{\partial\theta}$ and $\partial^2\ell(x|\theta) \stackrel{\text{def}}{=} \frac{\partial^2 \log(x|\theta)}{\partial\theta^2}$

$$0 = \sum_{t=1}^n \partial\ell(x_t|\theta + \widehat{\theta} - \theta) \approx \sum_{t=1}^n \partial\ell(x_t|\theta) + \sum_{t=1}^n \partial^2\ell(x_t|\theta)(\widehat{\theta} - \theta)$$

But (LLN)

$$\sum_{t=1}^n \partial^2\ell(x_t|\theta) \approx n\mathbb{E}\partial^2\ell(x|\theta) = -n\mathbb{E}\partial\ell(x|\theta)\partial\ell(x|\theta)^\dagger = -n\mathbf{F}_\theta$$

Putting all together gives us the influence function

$$\widehat{\theta} \approx \theta + \frac{1}{n}\sum_{t=1}^n f(x_t;\theta) \quad \text{for} \quad f(x;\theta) = \mathbf{F}_\theta^{-1}\partial\ell(x|\theta)$$

MLE asymptotics

Therefore

$$\mathsf{Cov}(\widehat{\theta}_{\mathsf{ML}}) \approx \frac{1}{n}\mathsf{Cov}(f) = \frac{1}{n}\mathsf{Cov}(\mathbf{F}^{-1}\partial\ell(x|\theta)) = \frac{1}{n}\mathbf{F}^{-1}\mathsf{Cov}(\partial\ell(x|\theta))\mathbf{F}^{-1}$$

But $\mathsf{Cov}_\theta(\partial\ell(x|\theta)) \stackrel{\mathsf{def}}{=} \mathbf{F}_\theta$ so $\mathsf{Cov}_\theta(\widehat{\theta}_{\mathsf{ML}}) = \frac{1}{n}\mathbf{F}_\theta^{-1}\mathbf{F}_\theta\mathbf{F}_\theta^{-1} = \frac{1}{n}\mathbf{F}_\theta^{-1}$.

The FIM for n independent samples is $n\mathbf{F}_\theta$.

Therefore, the MLE is asymptotically efficient!

When the model does not hold

We still assume i.i.d. samples but $p(x) \neq p(x|\theta)$ for all θ.

The MLE is still defined by

$$0 = \sum_{t=1}^{n} \frac{\partial \log(x_t|\widehat{\theta})}{\partial \theta}$$

As the sample size n grows to ∞, the MLE $\widehat{\theta}^n$ tends to the solution $\theta_\star$ of

$$0 = \mathbb{E}_p \frac{\partial \log(x|\theta_\star)}{\partial \theta}$$

We will see later that $p(x|\theta_\star)$ is the best approx. to the true distribution $p(x)$.
Meaning: *the MLE does something meaningful even for wrong models.*

The MLE is thus "‘biased" but $\widehat{\theta}_{\mathsf{ML}} \approx \theta_\star + \frac{1}{n}\sum_t f(x_t)$ with influence function

$$f(x) = -\left(\mathbb{E}_p \partial^2 \ell(x|\theta_\star)\right)^{-1} \partial \ell(x|\theta_\star)$$

and thus has, around $\theta_\star$, the asymptotic covariance matrix

$$\mathsf{Cov}(\widehat{\theta}_{\mathsf{ML}}) = \left(\mathbb{E}_p \partial^2 \ell(x|\theta_\star)\right)^{-1} \mathsf{Cov}_p\left(\partial \ell(x|\theta_\star)\right) \left(\mathbb{E}_p \partial^2 \ell(x|\theta_\star)\right)^{-1}$$