LECTURE 5: Information Theory, Entropy, Experiment Design

- The concept of information theory and entropy appears in many statistical problems.
- Here we will develop some basic theory and show how it can be applied to questions such as how to compute statistical (in)dependence and how to design experiments such that we achieve the goals
- In principle the development mirrors classical statistical mechanics
- There is a corresponding concept of quantum information theory, which we will not cover here (see e.g. Preskill's book)

Information Theory

- Suppose we have a random variable with a discrete set of outcomes p_i, for $i=1, \ldots, M$
- We construct a message from N independent outcomes of this variable
- We need $N\log_2 M$ bits to transmit this information
- But what if some are more likely than others: for large N we expect $N_i=Np_i$ events for each I
- Number of typical events is given by multinomial coefficient $g = N!/(\prod_{i=1}^{M} N_i!) << M^N$
- Remember Stirling formula $x! = x^x e^{-x}(2\pi x)^{1/2} + \ldots$
- The number of bits needed to specify one of g events in large N limit is $\log_2 g = -N\Sigma_i p_i \log_2 p_i << N\log_2 M$: Shannon's theorem proves that in large N limit error with this number of bits vanishes
- Information content of p is $I(p_i) = \log_2 M + \Sigma_i p_i \log_2 p_i$

Entropy and information: discrete case

- Shannon information (Shannon 1948):
  $h(x) = -log_2 p(x)$
- Its average is called Shannon entropy:
  $H(X)=-\Sigma_i p_i log_2 p_i$
- $log_2 g = log_2 N!/(\Pi_{i=1}^{M} N_i!)$ is known as entropy of mixing in the context of mixing of M components
- Example: English alphabet has information content of 4.7+4.1 bit
  ($p(x)=0$: $0*log_2 0=0$) ($log_2 27=4.7$)
- Entropy is minimized at 0 for $p_i=\delta_{i,j}$ and maximized for $p_i=1/M$: it is a measure of disorder

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

Relation between entropy and likelihood

- Instead of actual data likelihood we can replace it with its ensemble average
- Suppose we have N measurements x_i

$$L = \prod_i p(x_i)$$

$$\ln L = \sum_i \ln p(x_i)$$

$$\Big\langle \ln L \Big\rangle = \Big\langle \sum_i \ln p(x_i) \Big\rangle = N \Big\langle \ln p(X) \Big\rangle$$

$$\Big\langle \ln p(X) \Big\rangle \ (\text{or } E\{\ln p(X)\}) = \int dx \cdot p(X) \ln p(X) = -H(X)$$

$$\Big\langle \ln L \Big\rangle = -NH(X)$$

Entropy for Continuous Distribution

$$H(X) = -\int p(x)\, log\, p(x) dx = E\{-log\, p(X)\}$$

- Not invariant under reparametrization: if we change x to $F(x)$ entropy changes by $<|F'(x)|>$, so absolute value is meaningless. Not always positive definite. We will not distinguish $\log_2$ vs log/ln.
- In statistical mechanics this is solved by canonical conjugate pairs whose Jacobian is unity or if states are discretized (quantum statistics): no such concept in statistics

Entropy for Continuous Distribution

- Joint entropy of X and Y

$$H(X,Y) = -\int p(x,y)\ log\ p(x,y)dx\ dy = E\{-log\ p(X,Y)\}$$

- Conditional entropy of X given y

$$H(X|y) = -\int p(x|y)\ log\ p(x|y)dx = E\{-log\ p(X|Y)\mid Y = y\}$$

- Conditional of X given Y

$$\begin{aligned} H(X|Y) \ &: \int p(y) H(X|y)dy = -\int p(y)\int p(x|y)\ log\ p(x|y)dxdy \\ &: -\int\int p(x,y)\ log\ p(x|y)dxdy = E\{E\{-log\ p(X|Y)\mid Y\}\} \end{aligned}$$

Maximum Entropy

- For a bounded interval $a < x < b$ find p with maximum entropy given the normalization constraint: use Lagrange multiplier method

$$H(p) \triangleq -\int_a^b p(x)\, \lg p(x)\, dx$$

$$\begin{aligned} p(x) &\geq 0 \\ \int_a^b p(x)\, dx &= 1. \end{aligned}$$

$$J(p) \triangleq -\int_a^b p(x)\, \ln p(x)\, dx + \lambda_0 \left(\int_a^b p(x)\, dx - 1 \right)$$

$$\frac{\partial}{\partial p(x)\, dx} J(p) = -\ln p(x) - 1 + \lambda_0.$$

$$p(x) = e^{\lambda_0 - 1}.$$

$$\lambda_0 = 1 - \ln(b-a)$$

$$p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

Uniform distribution maximizes entropy

Maximum Entropy for Semi-unbounded Distributions

- If we are given mean on a semi-unbounded range from 0 to infinity, $p(x) = 0$ for $x < 0$

$$\int_{-\infty}^{\infty} x\, p(x)\, dx = \mu < \infty$$

$$\begin{aligned} J(p) \;\; &\triangleq \;\; -\int_0^{\infty} p(x)\, \ln p(x)\, dx + \lambda_0 \left(\int_0^{\infty} p(x)\, dx - 1 \right) . \\ &\quad\;\; + \lambda_1 \left(\int_0^{\infty} x\, p(x)\, dx - \mu \right) \end{aligned}$$

$$p(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note the "Boltzmann" factor $e^{-\beta x}$

Maximum Entropy for Unbounded Distributions

- If we are given mean μ and variance s on an unbounded range from – to + infinity

$$
\begin{aligned}
J(p) \triangleq & -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx + \lambda_0 \left(\int_{-\infty}^{\infty} p(x)\, dx - 1 \right) \\
& + \lambda_1 \left(\int_{-\infty}^{\infty} x\, p(x)\, dx - \mu \right) + \lambda_2 \left(\int_{-\infty}^{\infty} x^2\, p(x)\, dx - \sigma^2 \right)
\end{aligned}
$$

- $p(x) = e^{(\lambda_0 - 1) + \lambda_1 x + \lambda_2 x^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$ if $\mu=0$

- Can be further generalized: if we have constraints on first $n = 2k$ cumulants we obtain exponential of n-th order polynomial
- These "Boltzmann factors" have direct analogy with statistical mechanics

Kullback-Leibler (KL) divergence

- KL divergence is a relative entropy between two distributions (discrete or continuous)

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}. \qquad D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\,\log\frac{p(x)}{q(x)}\,dx.$$

- Satisfies Gibbs inequality KL $\geq$ 0: proof using Jensen inequality for convex functions (see e.g. MacKay 2.7) or:

$$\ln x \leq x - 1 \qquad \begin{aligned} -\sum_{i\in I} p_i \ln\frac{q_i}{p_i} &\geq -\sum_{i\in I} p_i\left(\frac{q_i}{p_i} - 1\right) \\ &= -\sum_{i\in I} q_i + \sum_{i\in I} p_i \end{aligned}$$

- This is 0 since probabilities are normalized
- It is not a distance: KL(*p*,*q*) is not KL(*q*,*p*)

KL Divergence for Gaussians

- Always positive
- Increases as the two distributions differ from each other
- Only zero when the two distributions are equal
- Good way to probe how similar are two distributions: starting point for Variational Inference/Variational Bayes methods

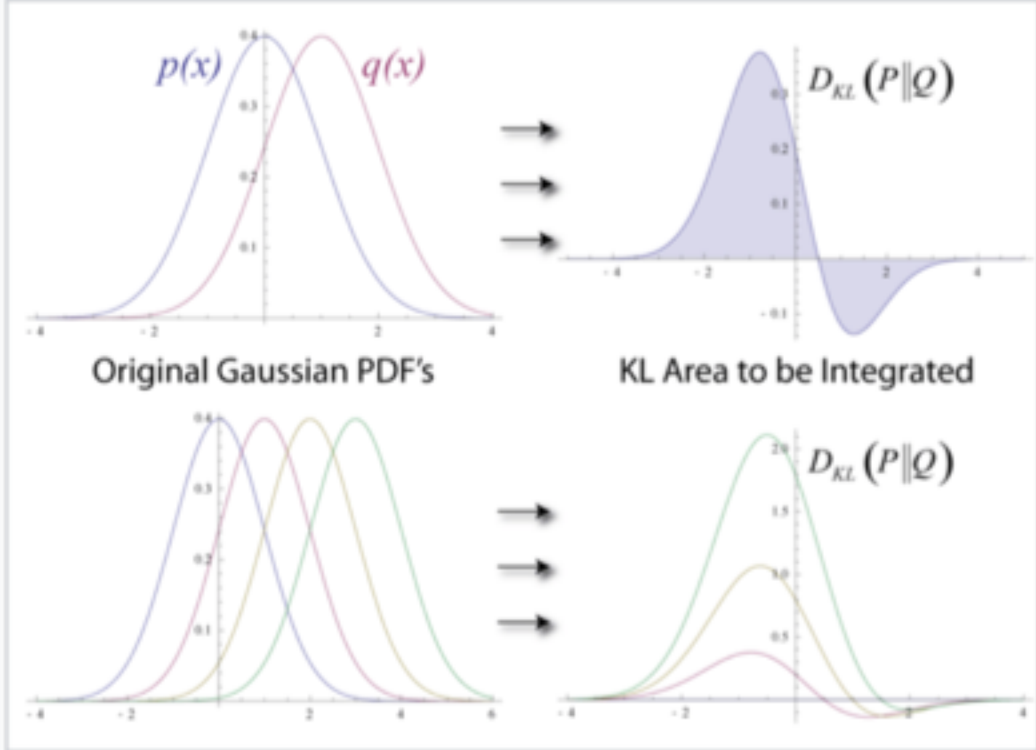


Illustration of the Kullback–Leibler (KL) divergence for two normal Gaussian distributions. Note the typical asymmetry for the Kullback–Leibler divergence is clearly visible.

Exercise: KL Divergence for Gaussians

- Assume p = gauss(μ_1, σ_1) and q = gauss(μ_2, σ_2)
- Evaluate KL(p||q) and show KL > 0
- Evaluate KL(q||p) and show it differs from KL(p||q)

KL Divergence and Negentropy

- Negentropy: KL divergence, i.e. relative entropy, against a Gaussian with equal variance

$$J(y) = H(y_G) - H(y) \geq 0$$

- Measures deviation of a distribution from gaussian. Can be approximated as

$$J(y) \approx \frac{1}{12}E\{y^3\}^2 + \frac{1}{48}kurt(y)^2$$

$kurt(y)=E(y^4)-3E(y^2)^2$

- But other approximations may work better:

$$J(y) = [E\{G(y)\} - E\{G(g)\}]^2$$

$$G_1(y) = \frac{1}{a}log\ cosh\ (a\ y), \qquad G_2(y) = -exp(-y^2/2)$$

Mutual Information

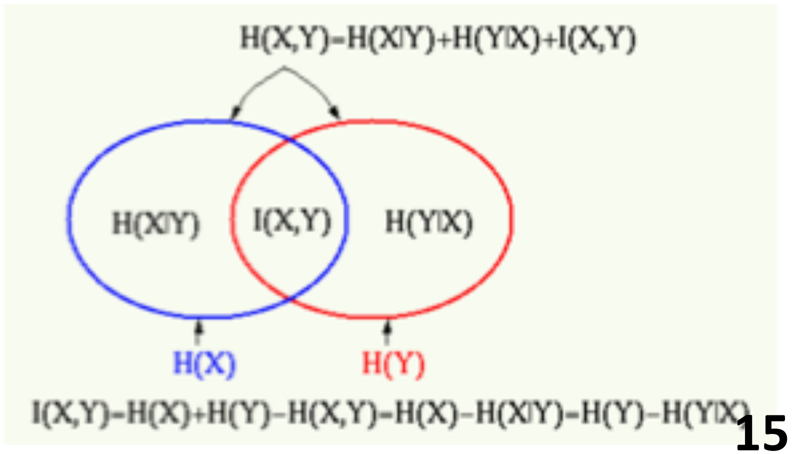
- Defined as amount of information shared between *X* and *Y*

$$
\begin{aligned}
I(X,Y) &= H(X) + H(Y) - H(X,Y) \\
&= E\{-log\ \ p(X)\} + E\{-log\ \ p(Y)\} + E\{-log\ \ p(X,Y)\} \\
&= E\{log\ \frac{p(X,Y)}{p(X)\,p(Y)}\}
\end{aligned}
$$

$$
\begin{aligned}
I(X,Y) &= E\{log\ \frac{p(X,Y)}{p(X)\,p(Y)}\} \\
&= E\{log\ \frac{p(X|Y)}{p(X)}\} = H(X) - H(X|Y) \\
&= E\{log\ \frac{p(Y|X)}{p(Y)}\} = H(Y) - H(Y|X)
\end{aligned}
$$

$$I(X;Y) = D_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y))$$

- Minimizing $I(X,Y)$ is a good way to define independence: $I(X,Y)=0$ if $H(X|Y)=H(X)$ or $H(Y|X)=H(Y)$ and is positive (KL divergence)



Multi-Information

- Generalization of mutual $I(\mathbf{y}) = \int P(\mathbf{y}) \log_2 \frac{P(\mathbf{y})}{\prod_i P(y_i)} d\mathbf{y}$
- Information $I(X,Y)$ to several variables y (or s)
- Multi-information is 0 if y statistically independent

$$
\begin{aligned}
I(\hat{\mathbf{s}}) &= \sum_i H\left[(\mathbf{V}\mathbf{x}_w)_i\right] - H\left[\mathbf{V}\mathbf{x}_w\right] \\
&= \sum_i H\left[(\mathbf{V}\mathbf{x}_w)_i\right] - \left(H\left[\mathbf{x}_w\right] + \log_2 |\mathbf{V}| \right)
\end{aligned}
$$

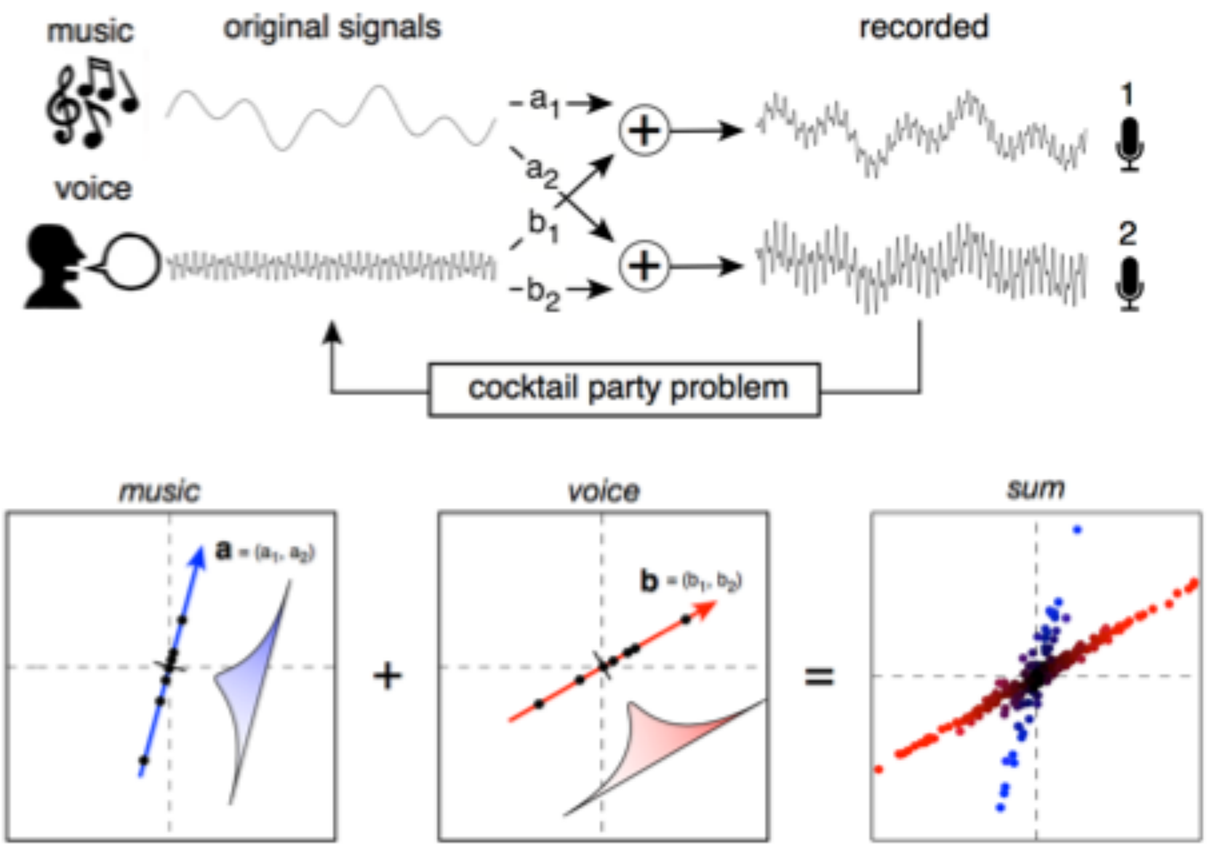
- ICA: we want to know V that minimizes $I(s)$, $s = Vx_w$

$$
\mathbf{V} = \arg\min_{\mathbf{V}} \sum_i H\left[(\mathbf{V}\mathbf{x}_w)_i\right]
$$

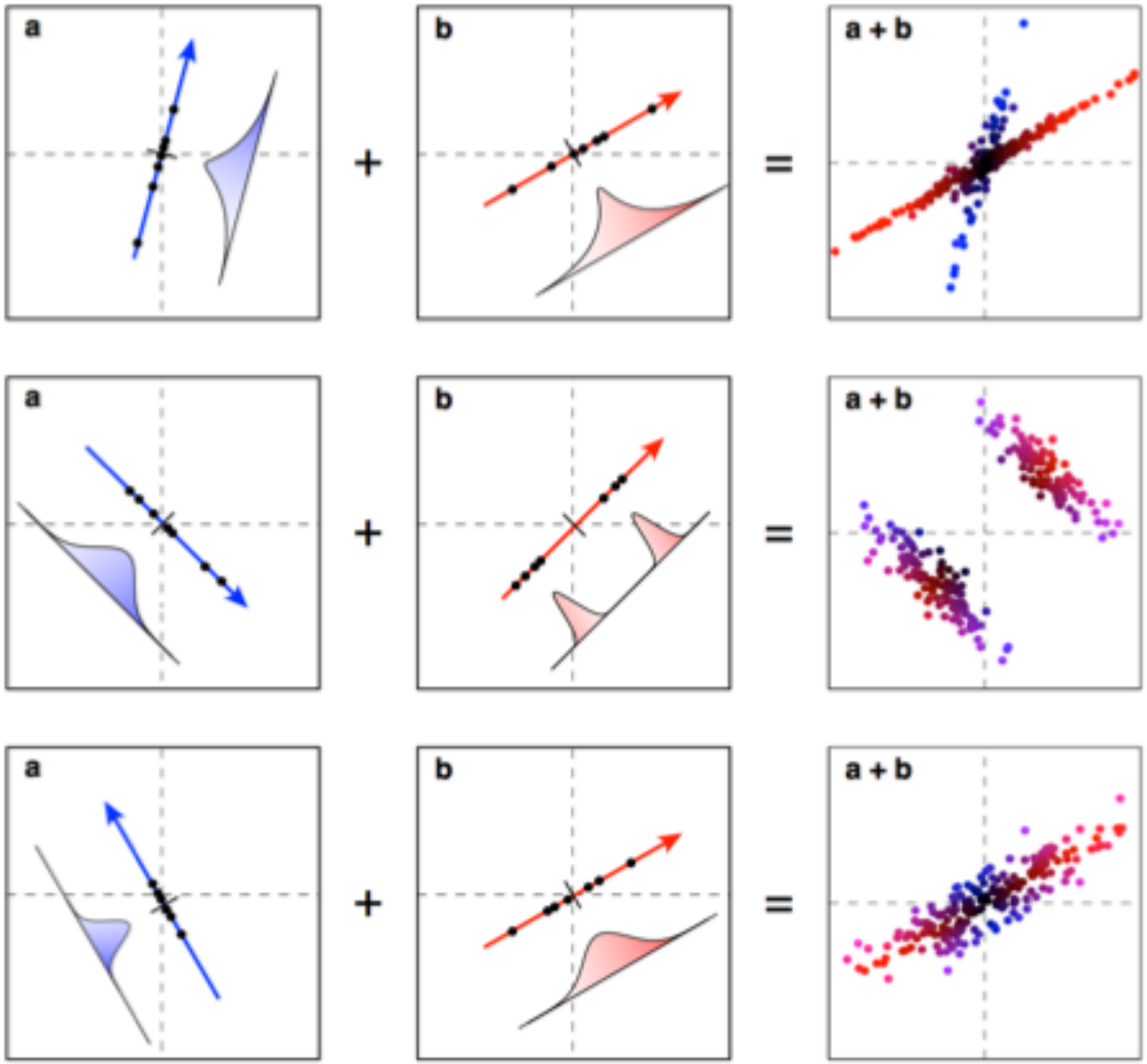
- This is equivalent to maximizing negentropy $J = \Sigma_i\, [H(s_{gi})\text{-}H(s_i)]$ where $s = Vx_w$
- We do not know $P(s)$ so we need to use some approximation to evaluate relative entropy

Independent Component Analysis (ICA): cocktail party problem

- We have 2 sources of sound and 2 microphones and we would like to separate the two

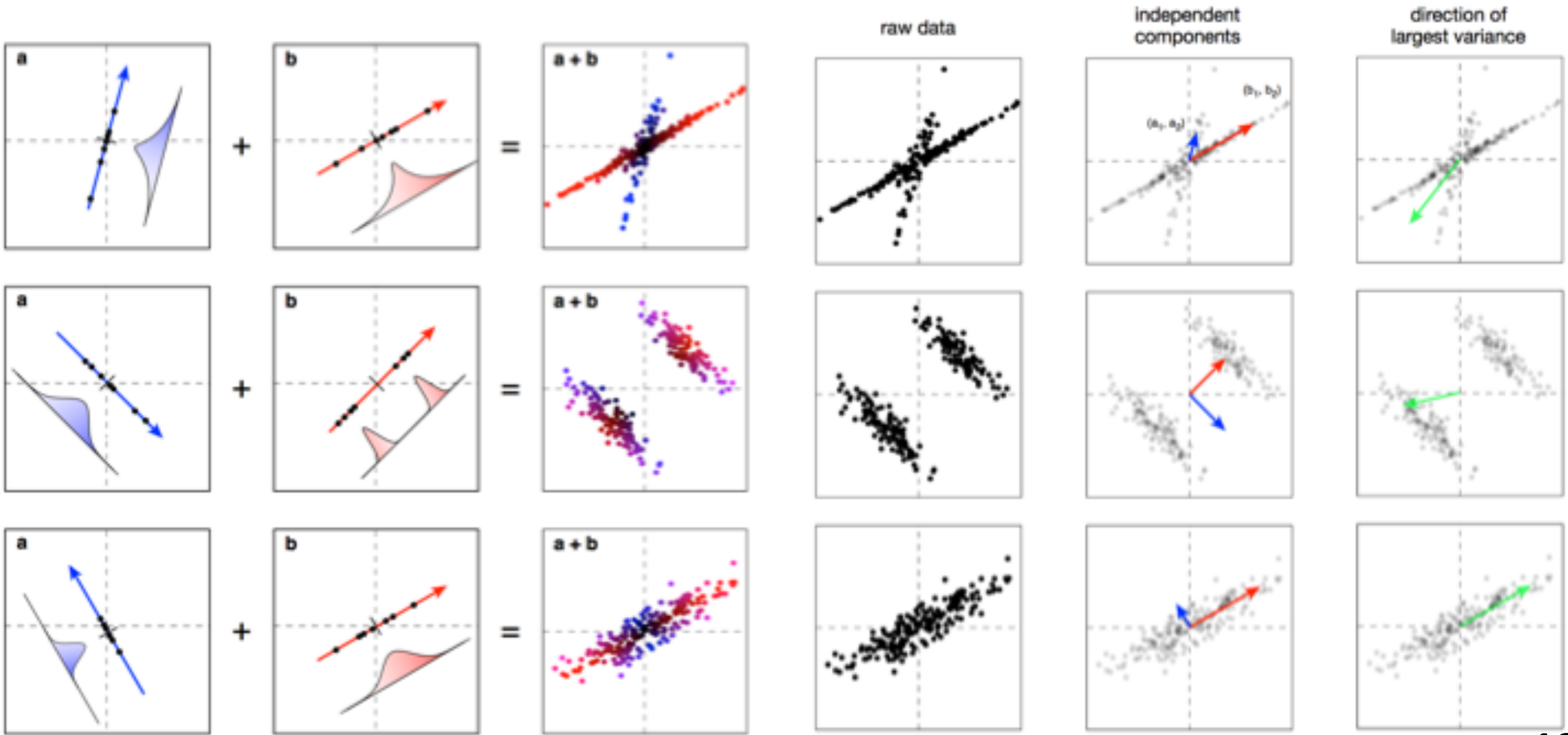


Many possible forms of independent sources



PCA vs. ICA

- They are different in general
- Mean is always subtracted
- PCA is not very meaningful for non-Gaussian distributions

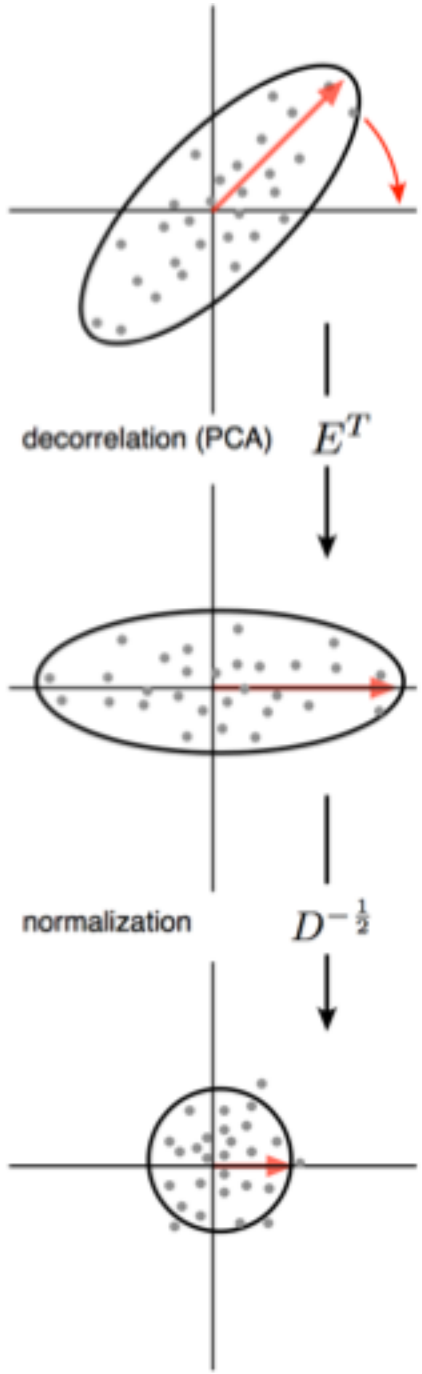


ICA Setup

- We have data $\mathbf{x}$ and would like to reconstruct individual sources $\mathbf{s}$ assuming they are statistically independent. We subtract the mean and assume linear relation $\mathbf{x}=\mathbf{A}\mathbf{s}$ but we do not know $\mathbf{A}$ or $\mathbf{s}$. We assume linear reconstruction $\mathbf{s}'=\mathbf{W}\mathbf{x}$. Our task is to determine $\mathbf{W}$.
- We can assume SVD of $\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$
- $\mathbf{W}=\mathbf{A}^{-1}=\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{T}$
- Let us whiten $\mathbf{s}$, $\mathbf{s}\mathbf{s}^{\mathrm{T}}=\mathbf{I}$
- Then $\mathbf{x}\mathbf{x}^{\mathrm{T}}=\mathbf{A}\mathbf{s}\mathbf{s}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{s}\mathbf{s}^{\mathrm{T}}\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}=\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}=\mathbf{U}\mathbf{\Sigma}^{2}\mathbf{U}^{\mathrm{T}}$
- We can diagonalize $\mathbf{x}\mathbf{x}^{\mathrm{T}}=\mathbf{E}\mathbf{D}\mathbf{E}^{\mathrm{T}}$, so we find SVD solution $\mathbf{W}=\mathbf{V}\mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}}$. We know $\mathbf{E}$ and $\mathbf{D}$.

ICA Continued

- We have identified **U** and **Σ** of **A**, without knowing **A**. We first decorrelated **A** via PCA (**E**) and then whitened **A** via $\mathbf{D}^{-1/2}$:
  $\mathbf{x_w}=\mathbf{D}^{-1/2}\mathbf{E}^T$ and $\mathbf{x_w x_w}^T=\mathbf{I}$
- We need to find $\mathbf{s'=Vx_w}$. Since V is orthonormal this can be viewed as another rotation that does not change $\mathbf{s's'}^T=\mathbf{I}$.
- If the problem was Gaussian uncorrelated and statistically independent are the same: any **V** would be as good as any other: rotation of a circle is still a circle.
- If it is non-Gaussian we can impose some other statistic to rotate $\mathbf{x_w}$ into statistically independent components

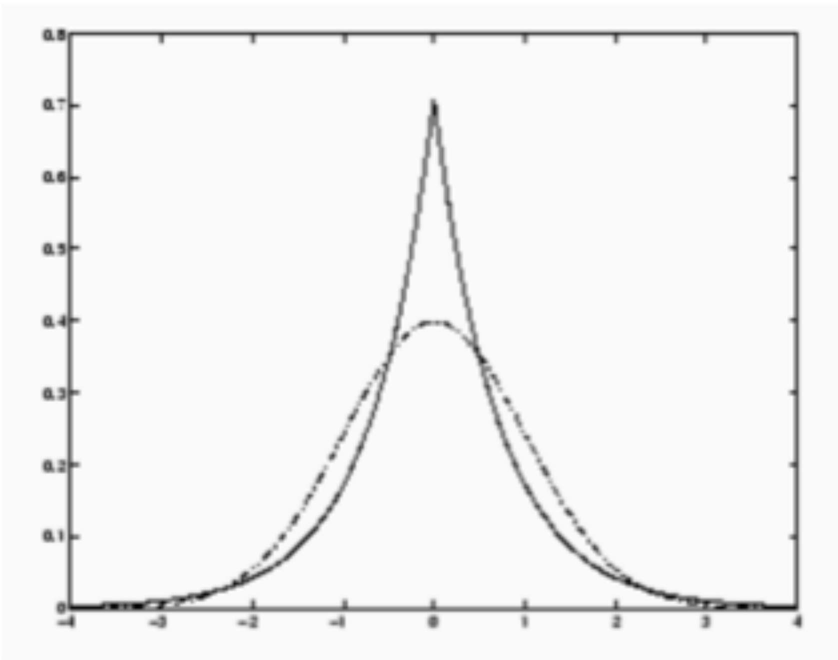


Measures of non-Gaussianity

- Moments: skewness, kurtosis, etc.
- Since we want independence we want to set to 0 cross-terms, e.g.
- Kurtosis Kurt$_{12}$=$<\mathbf{x}_{w1}^2\, \mathbf{x}_{w2}^2>$ -1
- In 2d **V** can be represented as a rotation angle

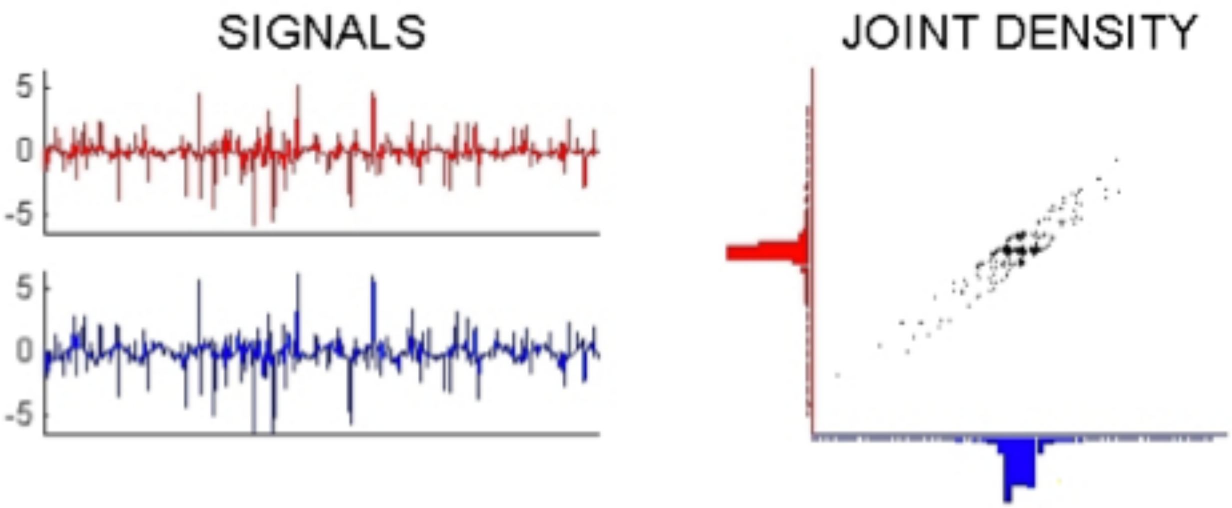
$$\mathbf{V} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

- Vary θ until Kurt$_{12}$ is 0.

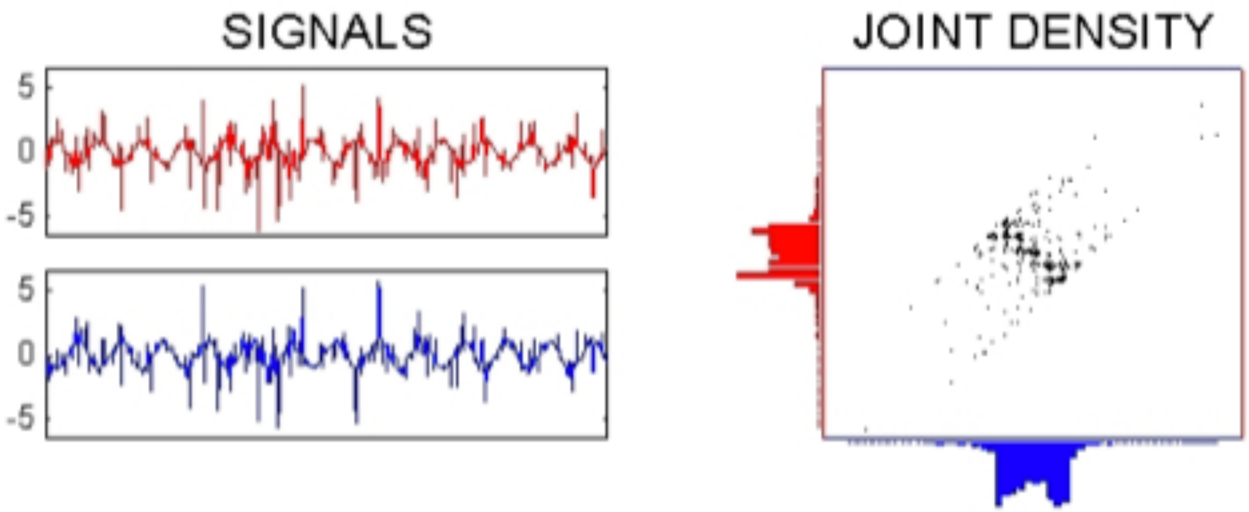


Fast ICA

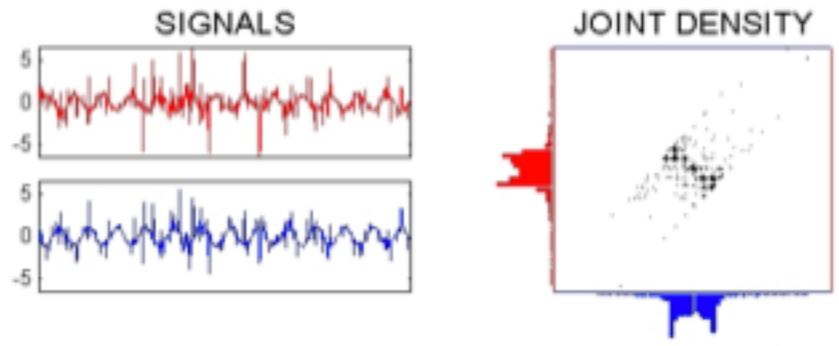
- Approximation for negentropy $J = [E(G(s))\text{-}E(G(g))]^2$
- Maximize $\Sigma_i E(G(s_i)) = \Sigma_i E(G(V^t{}_i x_w)) = E(G(V^t x_w))$
- Subject to normalization for V: Lagrange multiplier β
  $O(V) = E(G(V^t x_w))\text{-}\ \beta(V^T V\text{-}I)$
- This is optimization problem
- We will discuss how to solve optimization problems next, but typically this requires iterations, hence more complicated than linear algebra
- For large dimensions iterative methods are faster than linear algebra and even linear algebra problems are solved iteratively



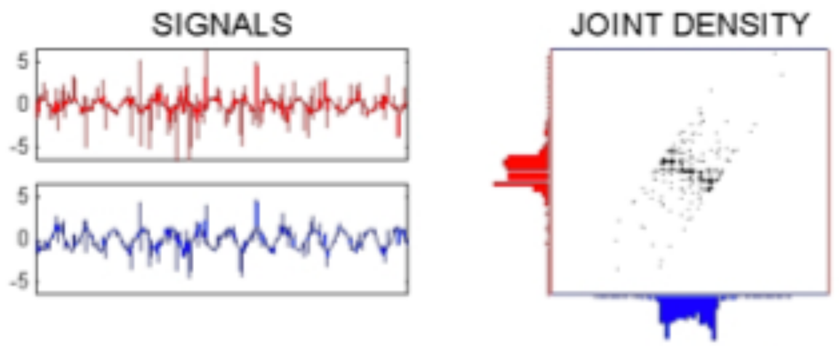
Input signals and density



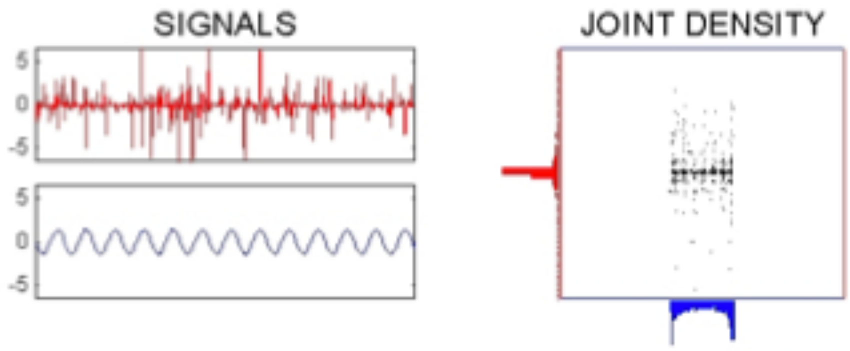
Whitened signals and density



Separated signals after 1 step of FastICA



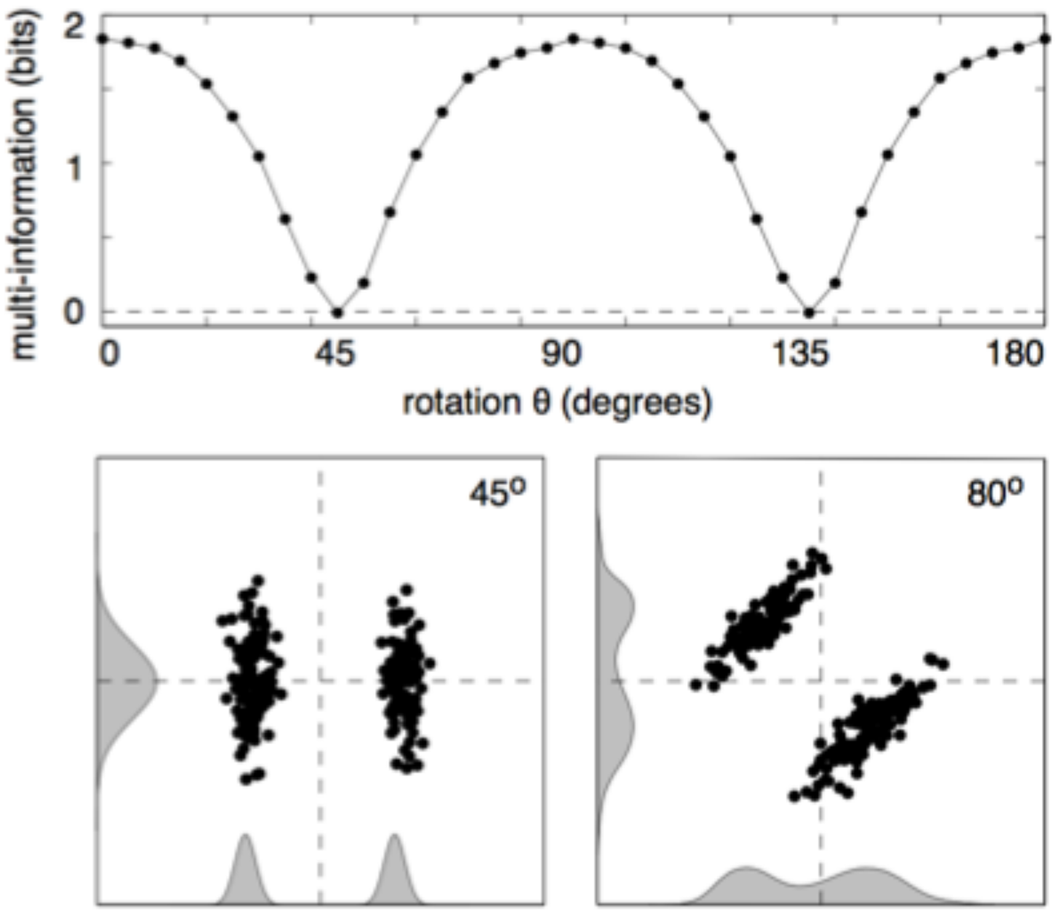
Separated signals after 2 steps of FastICA



Separated signals after 4 steps of FastICA

Subsequent iterations rotate V until it decorrelates the two signals

Doubly Peaked Example



This example uses multi-information: information theory (next lecture)

Fisher Information Matrix (Metric)

- Quantify the power of future experiments
- Instead of actual data likelihood we can replace it with its ensemble average
- Suppose we have N measurements x_i

$$L = \prod_i p(x_i)$$

$$\ln L = \sum_i \ln p(x_i)$$

$$\left\langle \ln L \right\rangle = \left\langle \sum_i \ln p(x_i) \right\rangle = N \left\langle \ln p(X) \right\rangle$$

$$\left\langle \ln p(X) \right\rangle \text{ (or } E\{\ln p(X)\}) = \int dx \cdot p(X) \ln p(X) = -H(X)$$

Ensemble Averaging: Precision matrix becomes Fisher matrix.

- we can Taylor expand around a fiducial model in terms of parameters Θ we wish to measure

$$\ln L(\vec{\theta}_{\rm fid} + \Delta\vec{\theta}) = \ln L(\vec{\theta}_{\rm fid}) + \sum_i \frac{\partial \ln L}{\partial \theta_i}\Big|_{\vec{\theta}_{\rm fid}} \delta\theta_i + \frac{1}{2}\sum_{ij} \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\Big|_{\vec{\theta}_{\rm fid}} \delta\theta_i \delta\theta_j$$

MLE: $\left\langle \frac{\partial \ln L}{\partial \theta_i} \right\rangle = E\left(\frac{\partial \ln p}{\partial \theta_i}\right) = 0$

Maximized at fiducial model: $\theta_i = \theta_{i,\rm fid}$

Fisher Matrix:

$$F_{ij} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle_{\theta_{\rm fid}} = \frac{\partial^2 H}{\partial\theta_i \partial\theta_j}$$

$$F_{ij} = \int dx \frac{\partial \ln p(x,\vec{\theta})}{\partial\theta_i} \frac{\partial \ln p(x,\vec{\theta})}{\partial\theta_j} p(x,\vec{\theta})$$

$$-\int p(x,\vec{\theta}dx = \int e^{-\ln p}dx = 1$$

$$\int e^{-\ln p}\frac{\partial \ln p}{\partial \theta_i}dx = 0$$

$$\int \left[e^{-\ln p}\frac{\partial^2 \ln p}{\partial \theta_i \theta_j} - e^{-\ln p}\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\partial \theta_j}\right]dx = 0$$

$$E\left(\frac{\partial^2 \ln p}{\partial \theta_i \theta_j}\right) = \int p\frac{\partial^2 \ln p}{\partial \theta_i \theta_j}dx = E\left(\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\partial \theta_j}\right) = \int p\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\partial \theta_j}dx$$

Back to Linear Least Squares

$$-\ln L = \sum_i \frac{\left(y_i - y(x_i|\vec{\theta})\right)^2}{2\sigma_i^2}$$

$$\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle = \left\langle \sum_k \frac{\frac{\partial y(x_k)}{\partial\theta_i} \frac{\partial y(x_k)}{\partial\theta_j}}{2\sigma_i^2} \right\rangle = F_{ij}$$

Posterior: $p(\vec{\theta}) \propto e^{-\delta\theta_i F_{ij} \delta\theta_j / 2}$

Covariance Matrix: $\left\langle \theta_i \theta_j \right\rangle = \left\langle \theta_i \right\rangle \left\langle \theta_j \right\rangle = F_{ij}^{-1}$

Experiment Design

- When we design an experiment we may be able to choose several parameters: sampling of points x_i where we measure data y_i, noise level σ_i, number of data points x_i etc.
- At a given x_i information on parameter Θ_j is given by $(dy_i/d\Theta_j)^2/\sigma_i^2$: this suggests choosing x_i where this is maximized. Note that this can be computed at the fiducial model without actually taking any data
- If we have several parameters we need to break their degeneracies: this is not possible if we only observe at a single x_i: we need to compute full Fisher matrix and invert it to obtain the final error estimate
- By varying the design of the experiment we can predict what the expected error on any given parameter will be: this enables us to design experiment to reach the goals we wish to achieve

Literature

- D. Mackay, *Information Theory, Inference, and Learning Algorithms* (See course website), Chapter 2
- M. Kardar, *Statistical Physics of Particles*, Chapter 2
- ICA: J. Shlens, *A Tutorial on Independent Component Analysis*, https://arxiv.org/pdf/1404.2986.pdf