Information geometry

J.-F. Cardoso, C.N.R.S.
Institut d'Astrophysique de Paris

Peyresq 2018

Note

These slides are supporting material for the lecture.

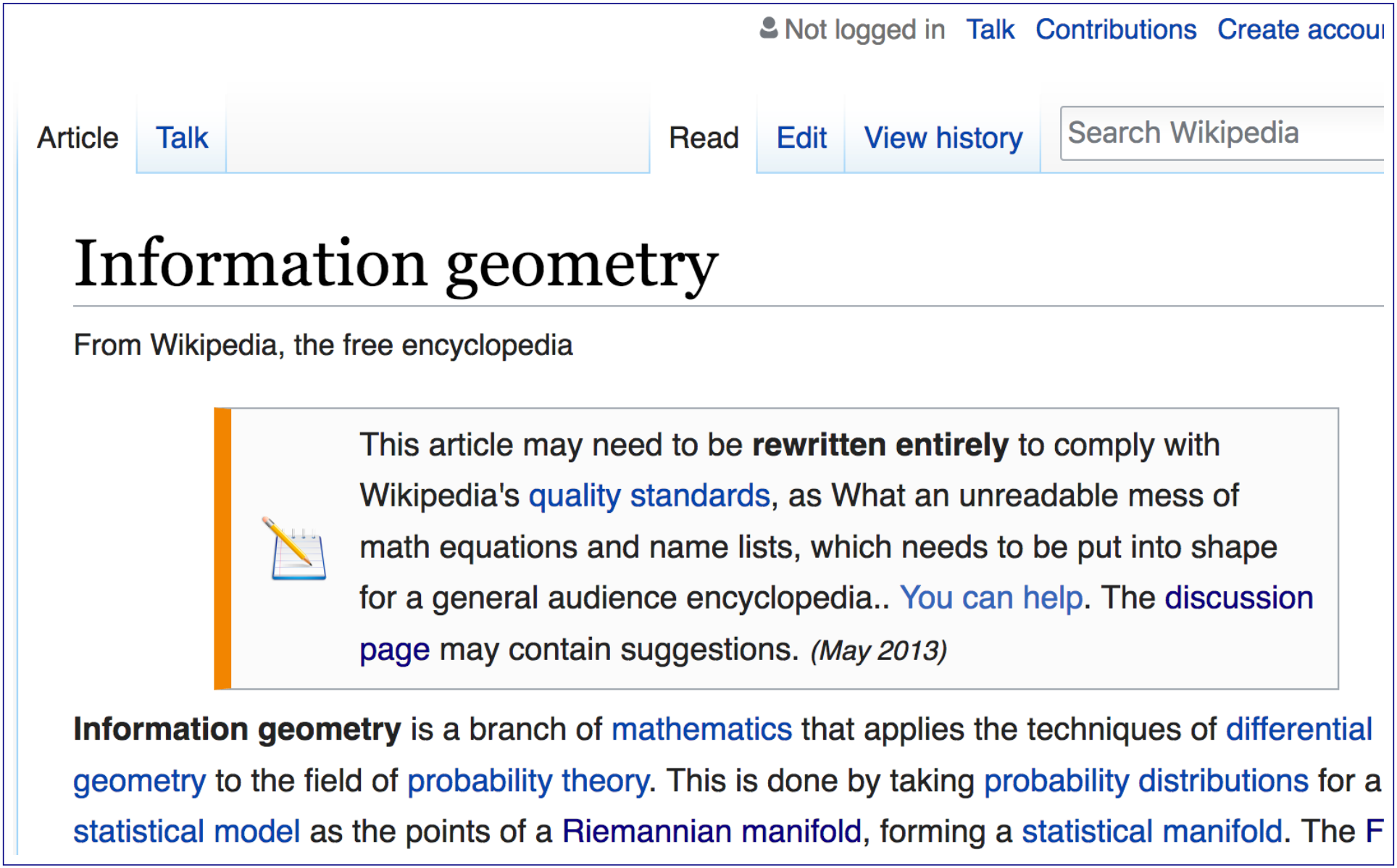
They do not provide a self-contained exposition.

Intro: Geometry in spaces of probability distributions

From Shun-ichi Amari, *Differential-geometrical methods in statistics.*

> One may ask why geometry, in particular differential geometry, is useful for statistics. The reason seems very simple and strong. A statistical model is a set of probability distributions. In particular, a parametric model usually forms a finite-dimensional manifold embedded in the set of all possible probability distributions. [. . .] One often uses a statistical model to carry out statistical inference, assuming that the true distribution is included in the model. However, a model is merely a hypothesis. The true distribution may not be in the model but be only close to it. Therefore, in order to evaluate statistical inference procedures, it is important to know what part the statistical model occupies in the entire set of probability distributions and what shape the statistical model has in the entire set. This is the problem of geometry of statistical models. It is therefore expected that a fundamental role is played in statistics by geometrical quantities such as the distance or divergence between two probability distributions, the flatness or curvature of a statistical model, etc...

Bad news from Wikipedia



Not logged in Talk Contributions Create accou

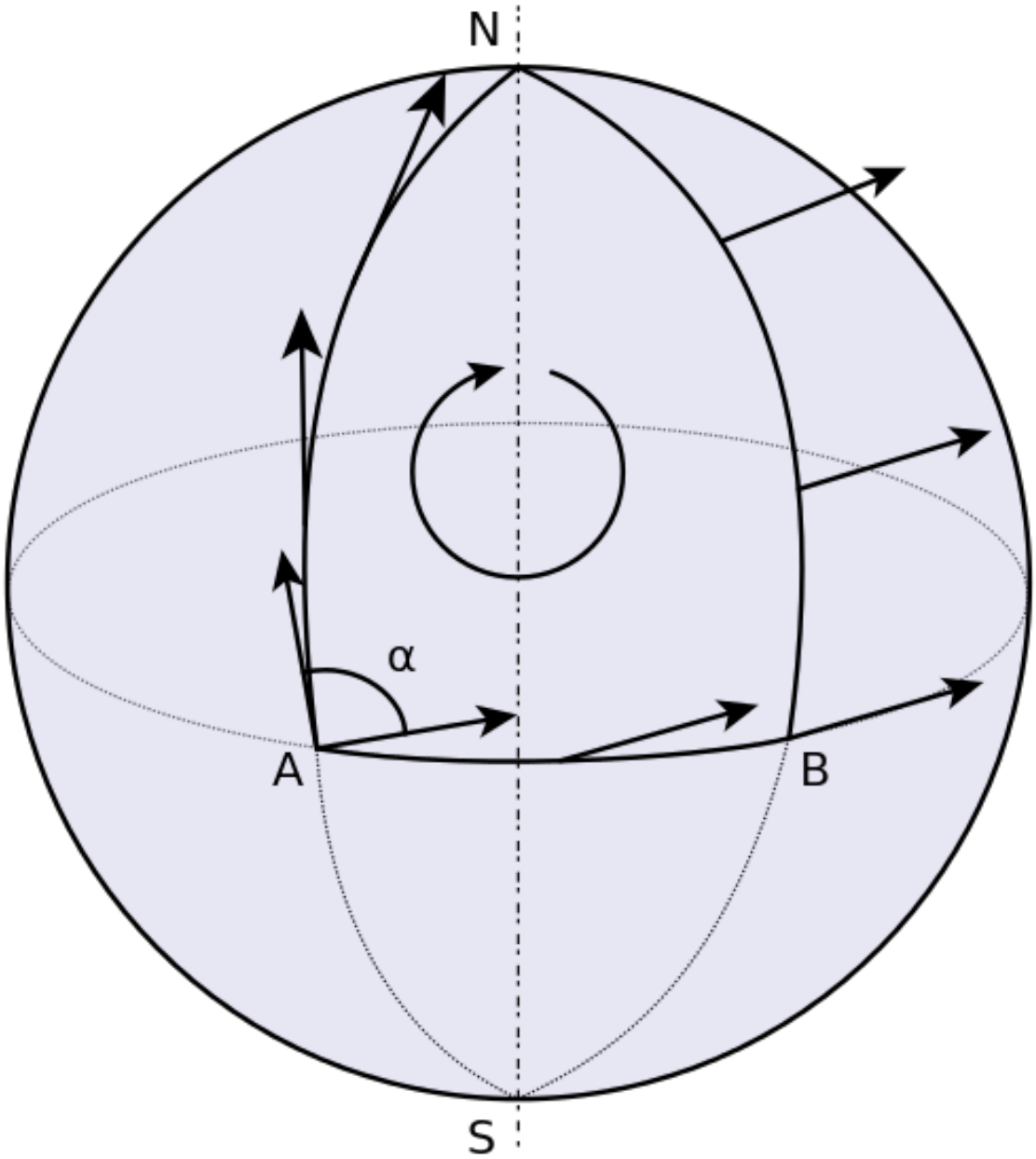
Article Talk Read Edit View history Search Wikipedia

Information geometry

From Wikipedia, the free encyclopedia

This article may need to be **rewritten entirely** to comply with Wikipedia's quality standards, as What an unreadable mess of math equations and name lists, which needs to be put into shape for a general audience encyclopedia.. You can help. The discussion page may contain suggestions. *(May 2013)*

Information geometry is a branch of mathematics that applies the techniques of differential geometry to the field of probability theory. This is done by taking probability distributions for a statistical model as the points of a Riemannian manifold, forming a statistical manifold. The F

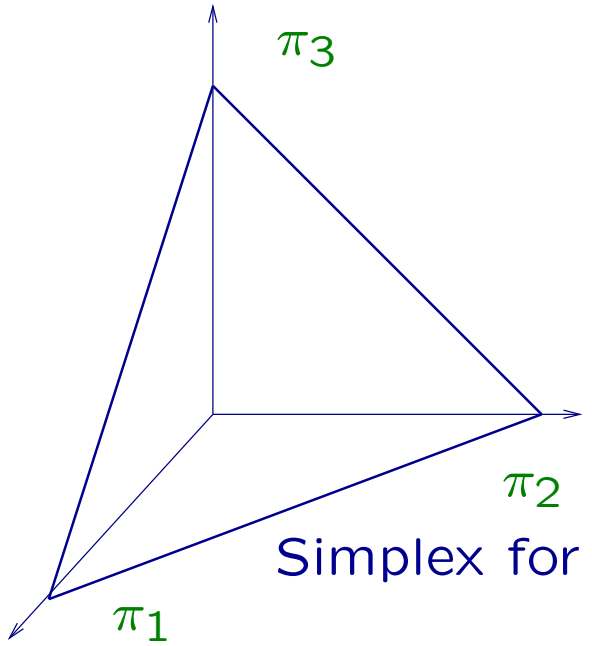

N
α
A
B
S

Points, manifolds and curves

Points I: Probability distributions for categorical variables

A random variable taking only X a discrete variable taking m distinct values $X \in \mathcal{X} = \{S_1, \ldots, S_m\}$ with probability $\pi = (\pi_1, \ldots, \pi_m)$.

Categorical distributions span the probability simplex: the subset of $\mathbb{R}^m$ such that $\sum_i \pi_i = 1$ and $0 \leq \pi_i \leq 1$.



Simplex for $m = 3$

Multinomial: n independent realizations of a categorical variable.
An n-sample projects right away on the simplex by the sample frequencies:

$$\widehat{\pi} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_m) = \left(\frac{n_1}{n}, \ldots, \frac{n_m}{n}\right)$$

Points II: Probability distributions for continuous variables

Continuous: a random variable taking values in a measurable space.

We shall only consider families of probability distributions having a density with respect to a common (dominating) measure. And we further require common support (or trouble will happen).

Parametric families $\{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}^p\}$ as p-dimensional manifolds.

The empirical distribution $\widehat{p}$ of n samples

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

does not live on the manifold.

Curves in distribution space: one-dimensional families

Such as simple transformation models:

- Location $p(x;\theta) = q(x-\theta)$
- Scale $p(x;\theta) = \frac{1}{\theta} q(\frac{x}{\theta})$

But such curves do not exist for a general random variable,
for which addition or multiplication are not defined.

We are looking for a **generic**, way of drawing a 'line' between two distributions.

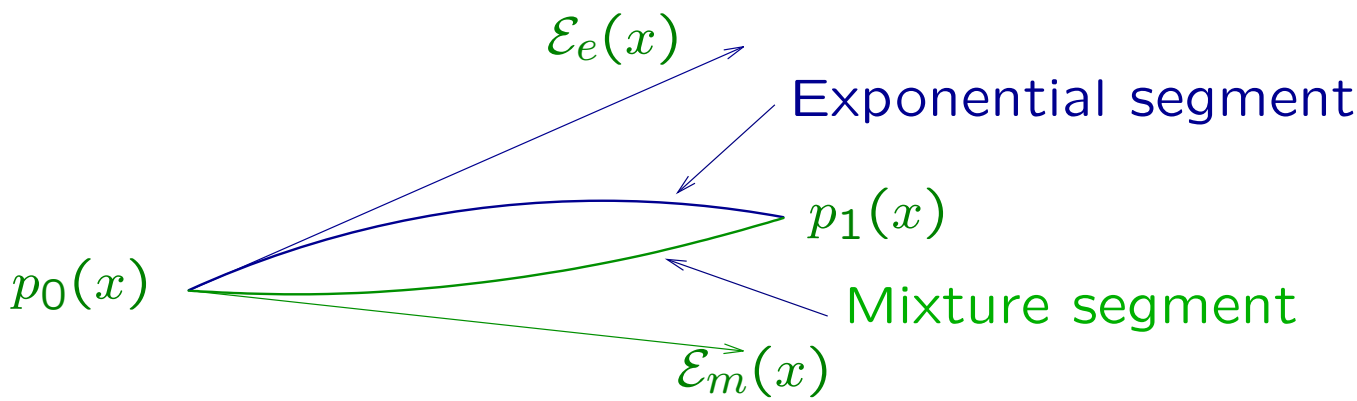
Statistics suggest at least two segments:

- Mixture segment
- Exponential segment

See next slide

Straight segment from p_0 to p_1 ? Two statistical ideas

- Defining 'segments' between two points $p_0(x)$ et $p_1(x)$.



- A 'mixture segment' from $p_0(x)$ to $p_1(x)$:

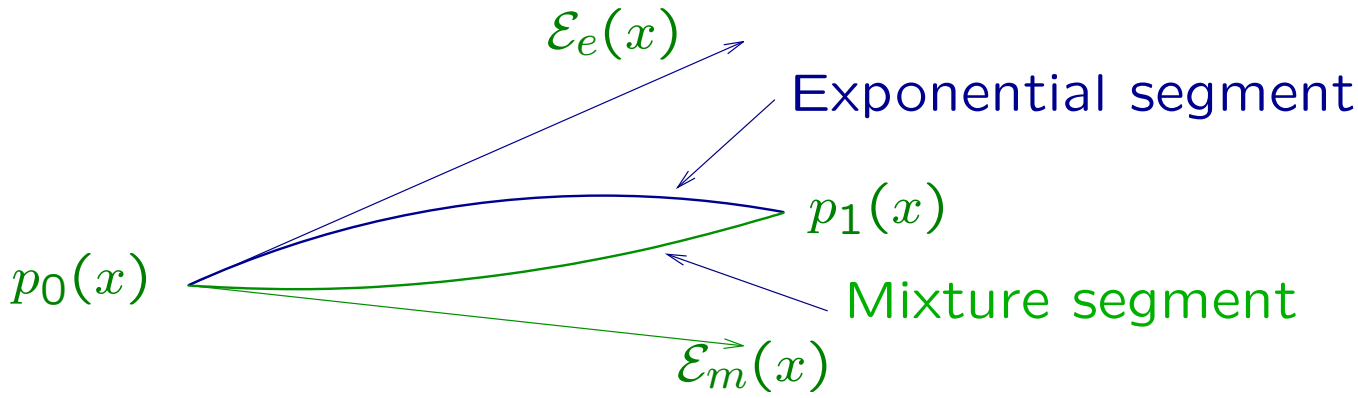
$$p_m(x;t) = (1-t)p_0(x) + tp_1(x), \qquad t \in [0,1]$$

- An 'exponential segment' from $p_0(x)$ to $p_1(x)$:

$$\log p_e(x;t) = (1-t)\log p_0(x) + t\log p_1(x) - \log \phi(t), \qquad t \in [0,1]$$

Do they look good in a simple Gaussian case ?

Tangent vectors (?)



The mixture and exponential segments from $p_0(x)$ to $p_1(x)$ also read

$$p_m(x;t) = p_0(x)\ (1 + t\ \mathcal{E}_m(x)) \qquad \mathcal{E}_m(x) = \frac{p_1(x) - p_0(x)}{p_0(x)}$$

$$p_e(x;t) = p_0(x)\ \exp\Big(t\ \mathcal{E}_e(X) - \psi(t)\Big) \qquad \mathcal{E}_e(x) = \log\frac{p_1(x)}{p_0(x)} - \mathbf{E}_0 \log\frac{p_1(x)}{p_0(x)}$$

We may interpret $\mathcal{E}_e(X)$ and $\mathcal{E}_m(X)$ as tangent vectors at point p_0.

Note that they are zero-mean there:

$$\mathbf{E}_0\ \mathcal{E}_e(X)\ =\ \mathbf{E}_0\ \mathcal{E}_m(X)\ =\ 0$$

Affine space

Affine space: a vector space which has forgotten its origin (and does not care).

In Euclidean space, 3 points A, B, C define a plane $\mathcal{P}$.

Pick up an origin, say point C, and represent any point M in the plane $\mathcal{P}$ as

$$M = M(\alpha, \beta): \quad C\vec{M} = \alpha C\vec{A} + \beta C\vec{B}$$

From segments to affine manifolds

Def. An exponential (resp. mixture) family contains all the exponential (resp. mixture) segments between any two of its points.

So we start with $n+1$ distributions $p_0, \ldots, p_n$ and build the n-dimensional model 'spanned' by them.

In the exponential case, that would be:

$$\log p(x;\theta) = \log p_0(x;\theta) \; + \; \sum_{i=1}^{n} \theta_i \log \frac{p_i(x)}{p_0(x)} - \psi(\theta) \qquad \theta \in \Theta \subset \mathbb{R}^n$$

This is an exponential family. The usual, more general (?), defintion:

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

where $g(x)$ is some positive measure and $S(x)$ an $n \times 1$ function of x, called the *sufficient statistic*. Then

$$\psi(\theta) = \log \int g(x)\, e^{\theta^\dagger S(x)}\, dx$$

This means bliss. Note that this is an 'affine' model. More about it soon.

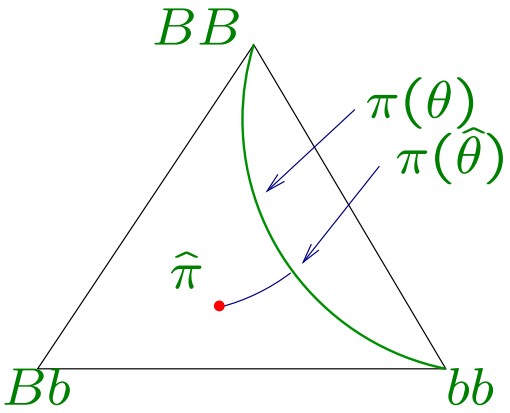
Statistical manifolds

Hardy-Weinberg equilibrium: a manifold for a 3-categorical variable

In a population, a gene exists in two forms B and b
in proportions θ and $1-\theta$ respectively.

Individuals can carry pairs (BB), (Bb), (bb).
At equilibrium, these are found with probabilities:

(BB)	(Bb)	(bb)
θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

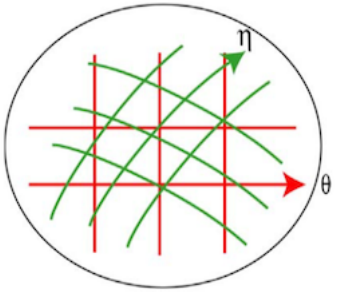


With just a little bit of work,
the Hardy-Weinberg model is seen to be
an exponential family. Hint:
$\log P_{BB} + \log P_{bb} - 2\log p_{Bb} = -\log 4$

Statistical manifolds: smooth parametric families of distributions

Parametric model: $p(x;\theta)$ for the distribution of X. $X \in \mathcal{X}$, $\theta \in \Theta \subset \mathbb{R}^p$.
Assume smoothness and a dominating measure.

The parameter θ is used to label the distributions in the parametric model but there are (infinitely) many ways to do it.
For instance, the normal family $\mathcal{N}(\mu, \sigma^2)$ may be parameterized by



$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \quad \theta = \begin{bmatrix} \mu \\ \log \sigma \end{bmatrix} \quad \theta = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix} \quad \theta = \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix}$$

Statistical inference should be invariant with respect to parameterization
$\rightarrow$ we mean *geometry*.

Still, geometry may help us pick 'preferred' parameterizations. . .

In the above example, which one do you think is 'best' ?

And for the Hardy-Weinberg model ?

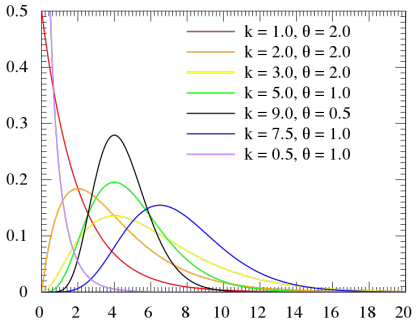
Some parametric models. I: Generative and transformation models

- Location $p(x;\theta) = q(x-\theta)$
- Scale $p(x;\theta) = \frac{1}{\theta} q(\frac{x}{\theta})$, $\theta \in \mathbb{R}^+$
- Location-scale: $p(x;\theta) = \frac{1}{\theta_1} q(\frac{x-\theta_2}{\theta_1})$
- Multi-dimensional scale: $p(x;\theta) = \frac{1}{|\det\theta|} q(\theta^{-1}x)$, with$x \in \mathbb{R}^n$, $\theta \in \mathsf{GL}(n)$.
- Contamination: the distribution of $\frac{X+\lambda Y}{\sqrt{\sigma_Y^2+\lambda^2\sigma_Y^2}}$ when λ is varied over $[0,+\infty]$
- Mixture of Gaussians
- Graphical models
- From statistical physics
- Any stochastic physical model with adjustable parameters. . .

Some parametric models. II: From the big book

Such as, say, the Gamma distribution, with support on R^+. Scale parameter θ and shape parameter k:

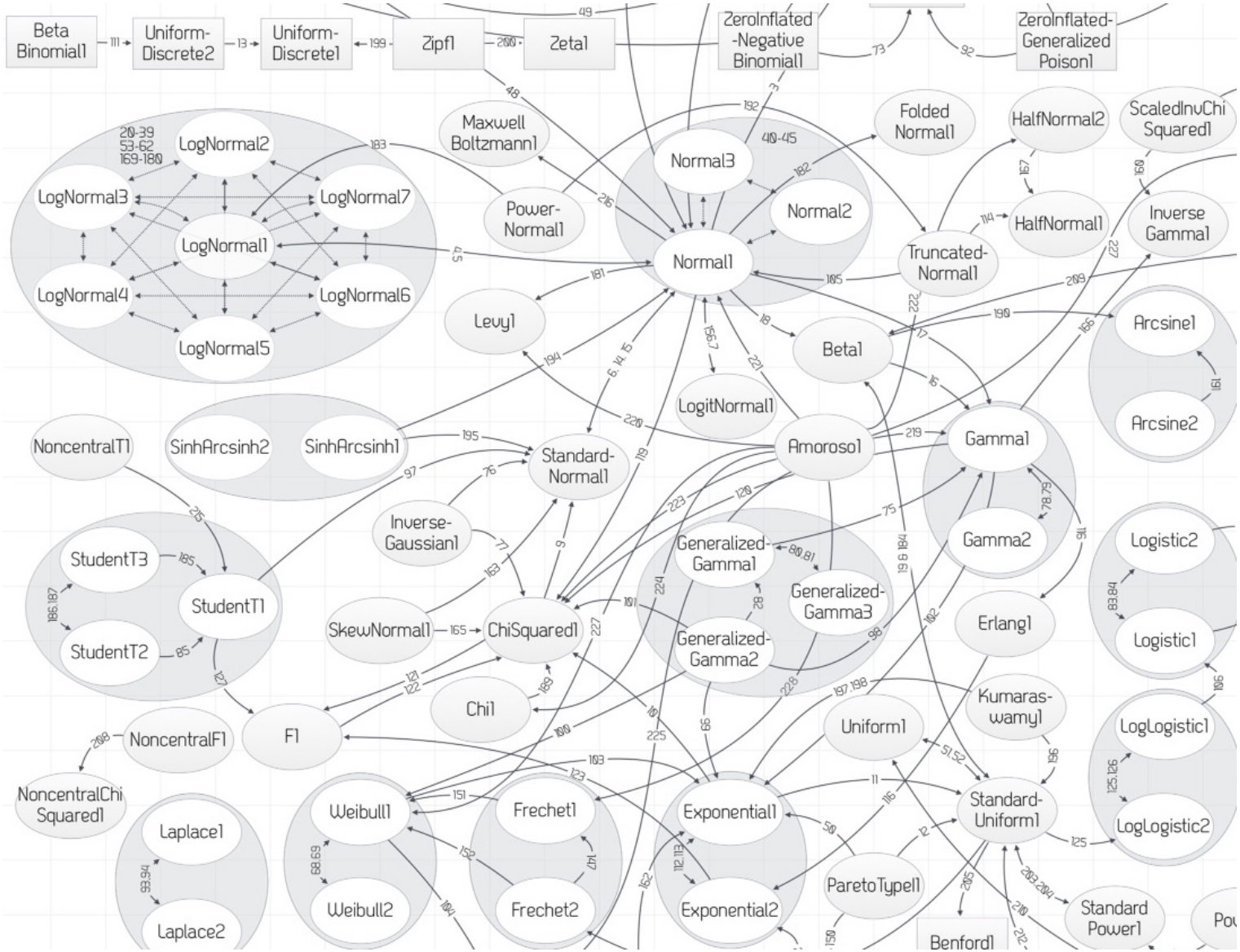
$$p(x;\theta,k) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$$



For integer k, it is the distribution of the sum of k independent exponentially distributed random variables, each with mean θ (Erlang distribution).

And many, many more. . .

Some parametric models. II: From the big book



Overview of supported distributions in ProbOnto, and relationships between univariate probability distributions (cropped). Credits: probonto.org

Some parametric models. III: Exponential families. Oh so nice!

- A p-dimensional family of distributions for a random X (of any kind):

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

where $g(x)$ is some positive measure, $S(x)$ is a $p \times 1$ function of x.

$S(x)$: sufficient statistic, $\qquad \psi(\theta) = \log \int g(x)\, e^{\theta^\dagger S(x)}\, dx$: potential function.

- Differentiate $\mathbf{E}_\theta 1 = 1$ twice to get:

$$\frac{\partial \psi}{\partial \theta} = \mathbf{E}_\theta S(x), \qquad \frac{\partial^2 \psi}{\partial \theta^2} = \mathbf{Cov}_\theta S(x) \text{ Convexity!}$$

Differentiate further: ψ is the cumulant (of $S(X)$) generating function.

- Optional normalizations:
  - Take $\int g(x) = 1$ to have $p(x;0) = g(x)$ then ... $\qquad \psi(0) = 0.$
  - Replace $S(x)$ by $S(x) - \mathbf{E}_0 S(x)$ then ... $\qquad \frac{\partial \psi}{\partial \theta}(0) = 0.$
  - Replace $S(x)$ by $[\mathbf{Cov}_0 S(x)]^{-1/2} S(x)$ then ... $\qquad \frac{\partial^2 \psi}{\partial \theta^2}(0) = I_p.$
  - Change the origin and the basis vectors as you please.

- See the affine structure for the **canonical parameter** θ.

Who is exponential ?

The key thing in the exponential structure is the sufficient statistic

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

and that data and parameters interact only in the log through $\theta^\dagger S(x)$.

Is that so common ? Maybe after some massaging. . .

- The normal family $\mathcal{N}(\mu, R)$ for $X \in \mathbb{R}^d$ is a p-dimensional exponential family $p = d + d(d+1)/2$ since $-(x-\mu)R^{-1}(x-\mu)/2 = \theta^\dagger S(x) - \mu^\dagger R^{-1}\mu/2$ with

$$S(x) = [x_i,\ x_i x_j\,]_{1\le i\le j\le d} \quad \theta = [(R^{-1}\mu)_i,\ -(R^{-1})_{ij}/2\,]_{1\le i\le j\le d}$$

- The set of all multinomial distributions.

- Look in the big book: Normal, Gamma, Wishart, Poisson, . . .

- *Maxent* distributions: distributions of maximal entropy with given moments.

- Curved exponential families as universal approximants to smooth models.

Kullback-Leibler divergence and the likelihood

Likelihood for multinomials

Data $X^n = [X_1, \ldots, X_n]$ where $X_i \in \mathcal{X} = [S_1, \ldots, S_p]$ modelled as n independent samples, each drawn with probability $\pi = [\pi_1, \ldots, \pi_p]$ *i.e.* $\mathsf{Prob}(X_i = S_k) = \pi_k$.

Probability of an n-sequence under a distribution categorical distribution π.

$$\begin{aligned}
\log p(X^n; \pi) &= \sum_{k=1}^{n} \log \mathsf{Prob}(X_k = S_k) \\
&= \sum_{i=1}^{p} \#(X_k = S_i) \log \pi_i \\
&= n \sum_{i=1}^{p} \widehat{\pi}_i \log \pi_i
\end{aligned}$$

where $\widehat{\pi}_i = \#(X_k = S_i)/n$ is the frequency of S_i in the sequence X^n.

Key point: the probability of the sample depends only on the empirical distribution $\widehat{\pi} = [\widehat{\pi}_1, \ldots, \widehat{\pi}_p]$ which thus is a *sufficient statistic*.

Likelihood and Kulback matching

An enlightning decomposition:

$$-\frac{1}{n}\log p(X^n;\pi) = -\sum_{i=1}^{p}\widehat{\pi}_i \log \pi_i = \sum_{i=1}^{p}\Big[\widehat{\pi}_i \log\frac{\widehat{\pi}_i}{\pi_i} - \widehat{\pi}_i \log \widehat{\pi}_i\Big]$$

For two categorical distributions p and q, define the KL divergence:

$$K\left[p\,|\,q\right] \stackrel{\text{def}}{=} \sum_i p_i \log\frac{p_i}{q_i} \quad \text{and} \quad H(p) \stackrel{\text{def}}{=} -\sum_i p_i \log p_i \quad \text{Shannon entropy.}$$

The Kullback-Leibler divergence from p to q also reads

$$K\left[p\,|\,q\right] = \sum_i p_i \; k\Big(\frac{q_i}{p_i}\Big) \qquad k(u) \stackrel{\text{def}}{=} u - 1 - \log u \geq 0$$

This (non symmetric) measure of discrepancy between discrete distributions allows us to write the likelihood as

$$p(X^n;\pi) = e^{-n(K[\widehat{\pi}\,|\,\pi]+H(\widehat{\pi}))}$$

The likelihood measures goodness of fit in the Kullback sense.

Maximum likelihood estimation is a Kullback-projection onto the model.

Distribution of the empirical distribution

If n samples are i.i.d. distributed according to π and if X^n has type $\hat{\pi}$.

$$p(X^n; \pi) = e^{-n\{K[\hat{\pi}\,|\,\pi]+H(\hat{\pi})\}}$$

The set $\mathcal{T}(\hat{\pi})$ of n-samples with an empirical distribution $\hat{\pi}$ has cardinality:

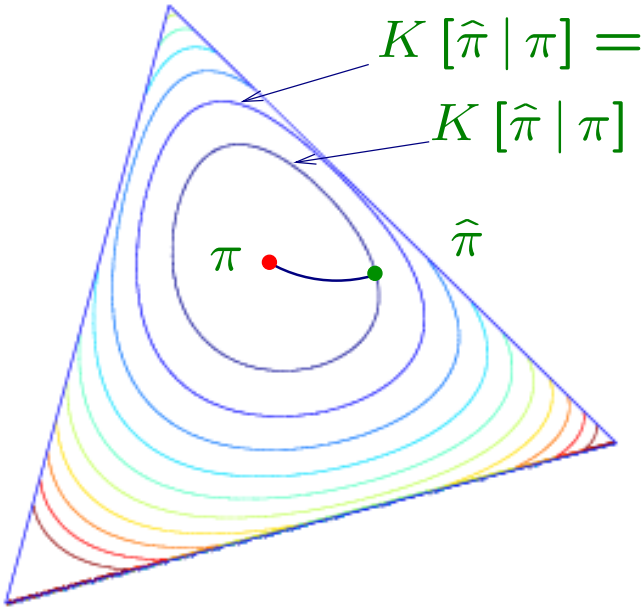
$$|\mathcal{T}(\hat{\pi})| = \frac{n!}{\prod_i n_i!} \approx e^{n\,H(\hat{\pi})} \qquad \text{(via Stirling approx.)}$$

Then, the probability of observing a particular type $\hat{\pi}$ is

$$p(\hat{\pi}; \pi) = \sum_{X^n \in \mathcal{T}(\hat{\pi})} p(X^n; \pi) = |\mathcal{T}(\hat{\pi})|\; e^{-n\{K[\hat{\pi}\,|\,\pi]+H(\hat{\pi})\}} \approx e^{-n\,K[\hat{\pi}\,|\,\pi]}$$

Thus, the sample distribution $\hat{\pi}$ is concentrated in a 'Kullback ball' centered at the true distribution π with a 'radius' of order $1/n$.

Distribution of sample distributions (ctd)



- $p(\hat{\pi}) \approx e^{-n\,K[\hat{\pi}\,|\,\pi]}$
- Most of the probability is concentrated in a Kullback-ball of center π and of radius $1/n$
- or is it $1/\sqrt{n}$?

For $\hat{\pi}$ close to π, the KLD reduces to the (symmetric) quadratic χ^2 distance:

$$K[\hat{\pi}\,|\,\pi] \approx \frac{1}{2}\sum_i \frac{(\hat{\pi}_i - \pi_i)^2}{\pi_i} \approx K[\pi\,|\,\hat{\pi}]$$

and small Kullback balls become ellipsoids with covariance matrices:

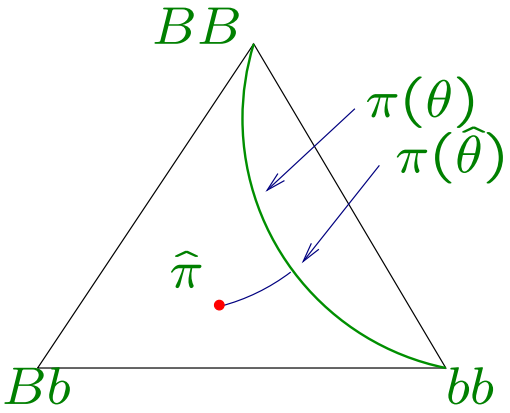
$$\mathbf{Cov}(\hat{\pi}_i, \hat{\pi}_j) = \frac{1}{n}(\delta_{ij}\pi_i - \pi_i\pi_j)$$

Can you see why?

Back to Hardy-Weinberg

If $\pi = \pi(\theta)$, the maximum likelihood estimate of θ is the minimizer of $K\left[\widehat{\pi} \,|\, \pi(\theta)\right]$. *Kullback matching*.

Remember Hardy-Weinberg



(BB)	(Bb)	(bb)
θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Finding the maximum likelihood solution is Kullback-projecting the sample distribution onto the model.

We'll see later why this picture is a bit wrong.

Kullback matching for continuous variables. I

- Kullback-Leibler divergence *from* a distribution P *to* a distribution Q:

$$K\left[P \,|\, Q\right] = \mathbf{E}_P \log \frac{P(x)}{Q(x)} = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

- It is *not* symmetric: in general, $K\left[P \,|\, Q\right] \neq K\left[Q \,|\, P\right]$

- It is strictly positive unless P and Q agree P-almost everywhere because $K\left[P \,|\, Q\right] = \mathbf{E}_P\left[k\left(Q(x)/P(x)\right)\right]$ with $k(u) = u - 1 - \log(u) \geq 0$.

- It is invariant: under any invertible transform $f(\cdot)$,

$$K\left[P_X \,|\, P_Y\right] = K\left[P_{f(X)} \,|\, P_{f(Y)}\right]$$

Kullback matching for continuous variables. II

Question: With a continuous variable X and a parametric model $p(x;\theta)$, can we play the same trick : maximizing the likelihood is finding the model distribution which is the closest in the Kullback sense to the sample distribution ?

No, in general: the sample distribution for n independent realizations $x_1, \ldots, x_n$:

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

is at infinite distance from the model $K\left[\widehat{p}(x) \,|\, p(x;\theta)\right] = +\infty$.

But considering the likelihood loss:

$$\widehat{\mathcal{L}}(\theta) = -\frac{1}{n} \sum_{i} \log p(x_i;\theta)$$

as the estimate of the expected loss, with $p_*(x)$ the true distribution of X:

$$\mathcal{L}(\theta) = -\mathbf{E}_* \widehat{\mathcal{L}}(\theta) = -\int p_*(x) \log p(x;\theta) = K\left[p_*(x) \,|\, p(x;\theta)\right] + H(p_*)$$

gives us the **expected shape of the likelihood**.

Kullback matching in exponential families

Recall the generic exponential family:

$$p(x;\theta) = g(x)\, e^{\theta^\dagger S(x) - \psi(\theta)} \qquad \psi(\theta) = \log \int g(x)\, e^{\theta^\dagger S(x)}\, dx$$

$$\frac{\partial \psi(\theta)}{\partial \theta} = \mathbf{E}_\theta S(X) \stackrel{\text{def}}{=} \eta(\theta) \qquad \frac{\partial^2 \psi(\theta)}{\partial \theta^2} = \mathbf{Cov}_\theta S(X)$$

Since $\psi(\theta)$ is convex, there is a 1-to-1 mapping $\theta \longleftrightarrow \eta = \frac{\partial \psi(\theta)}{\partial \theta}$.
Any distribution can be uniquely labeled with $\eta = \eta(\theta) = \mathbf{E}_\theta S(x)$.

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

So the maximum likelihood estimate of η based on x is $\widehat{\eta}_{\mathsf{ML}} = S(x)$.

Based on an n-sample, it is $\widehat{\eta}_{\mathsf{ML}} = \widehat{\mathbf{E}} S(x)$.

With $\widehat{\theta}_{\mathsf{ML}} = \theta(\widehat{\eta}_{\mathsf{ML}})$, one has exactly

$$\widehat{\mathcal{L}}(\theta) = -\frac{1}{n} \sum_i \log p(x_i;\theta) = K\left[\widehat{\theta}_{\mathsf{ML}} \,|\, \theta\right] + \text{data only term}$$

yielding the exact shape of the likelihood in geometric terms.

Summary

- There is a natural divergence in spaces of probability distributions.
- Maximum likelihood is Kullback matching.
- So we found a (non)-distance for information geometry.
- But differential geometry needs
  a **local (infinitesimal) quadratic distance**.

The matrix and the metric

Estimation theory gives us a metric

Parametric model $p(x;\theta)$ for the distribution of $X \in \mathcal{X}$ with $\theta \in \Theta \subset \mathbb{R}^p$.

Estimator: a function $\mathcal{T} : \mathcal{X} \mapsto \mathbb{R}^p$ of the data, returns an estimate $\hat{\theta} = \mathcal{T}(X)$.

Cramér-Rao bound: The variance of an unbiased estimator is lower bounded:

$$\mathbf{Cov}_\theta(\mathcal{T}(X)) \geq F(\theta)^{-1} \qquad F(\theta)\text{: the } p \times p \text{ Fisher information matrix}$$

Hence, in essence, $p(x;\theta)$ and $p(x;\theta+\Delta\theta)$ are undistinguishable if, say,

$$\Delta\theta^\dagger\, F(\theta)\, \Delta\theta < 1 \qquad \text{A resolution cell.}$$

Infinitesimal distance between $p(x;\theta)$ and $p(x;\theta+d\theta)$?
The CR bound makes it statistically unavoidable (Rao 1945) to define it as

$$\sqrt{d\theta^\dagger F(\theta) d\theta}$$

This is parameterization invariant, of course.

Note: for n indepedent observations, the Fisher matrix is $nF(\theta)$.
The resolution cell shrinks accordingly.

Score function and Fisher matrix

The score function is the vector-valued function

$$\varphi(x;\theta) = \frac{\partial \log p(x;\theta)}{\partial \theta}$$

and the score $\varphi(X;\theta)$ is a random vector which depends on θ.

A trivial but striking property: for any smooth function $h(x)$,

$$\frac{\partial}{\partial \theta}\, \mathbf{E}_\theta\Big(h(X)\Big) = \mathbf{E}_\theta\Big(h(X)\varphi(X;\theta)\Big).$$

In particular, the score $\varphi(X;\theta)$ has zero mean when $X \sim p(x;\theta)$:

$$\mathbf{E}_\theta\ \varphi(X;\theta) = 0$$

The 'Fisher information matrix' is the covariance matrix of the score:

$$F(\theta) = \mathbf{Cov}_\theta(\varphi(X;\theta)) = \mathbf{E}_\theta\ \varphi(X;\theta)\varphi(X;\theta)^\dagger$$

Reminder: some (statistical) Euclidean geometry

• In a vector space $\mathbb{R}^n$: consider the approximation of a vector y as a linear combination of p vectors $\sum_{i=1}^{p} \Phi_i \alpha_i = \Phi\alpha$. Best ($\min_\alpha \|y - \Phi\alpha\|^2$) solution:

$$y^\star = \Phi\alpha^\star = \Phi(\Phi^\dagger\Phi)^{-1}\Phi^\dagger y = \Pi_\Phi y \qquad \Pi_\Phi \stackrel{\text{def}}{=} \Phi(\Phi^\dagger\Phi)^{-1}\Phi^\dagger = \mathsf{Proj}(\mathsf{Span}(\Phi))$$

Basis-free characterization: $\Pi_\Phi y^\star = y^\star$ and $\Pi_\Phi(y - y^\star) = 0$ (orthogonality).

Reminder: some (statistical) Euclidean geometry

• In a vector space $\mathbb{R}^n$: consider the approximation of a vector y as a linear combination of p vectors $\sum_{i=1}^p \Phi_i \alpha_i = \Phi\alpha$. Best ($\min_\alpha \|y - \Phi\alpha\|^2$) solution:

$$y^\star = \Phi\alpha^\star = \Phi(\Phi^\dagger\Phi)^{-1}\Phi^\dagger y = \Pi_\Phi y \qquad \Pi_\Phi \stackrel{\text{def}}{=} \Phi(\Phi^\dagger\Phi)^{-1}\Phi^\dagger = \mathsf{Proj}(\mathsf{Span}(\Phi))$$

Basis-free characterization: $\Pi_\Phi y^\star = y^\star$ and $\Pi_\Phi(y - y^\star) = 0$ (orthogonality).

• Prediction $\widehat{Y}$ of a random **scalar** Y by a linear combination of p zero-mean random variables: $\widehat{Y} = \sum_{i=1}^p \alpha_i \xi_i$. The best ($\min \mathbf{E}(\widehat{Y} - Y)^2$) prediction is

$$\widehat{Y} = X^\dagger \mathbf{E}(XX^\dagger)^{-1}\mathbf{E}(XY) \qquad X = [\xi_1, \ldots, \xi_p]^\dagger$$

Actually, we are working in a vector space of scalar random variables with scalar product $\langle \xi_i, \xi_j \rangle = \mathbf{E}(\xi_i \xi_j)$ and (squared) norm $\|\xi\|^2 = \langle \xi, \xi \rangle = \mathbf{E}\xi^2$.

Orthogonal decomposition as $Y = \widehat{Y} + (Y - \widehat{Y})$ since $\mathbf{E}\left((\widehat{Y} - Y)\widehat{Y}\right) = 0$.
It entails $\mathbf{E}Y^2 = \mathbf{E}\widehat{Y}^2 + \mathbf{E}(Y - \widehat{Y})^2$.

Spoiler: tangent planes can be constructed like that.

The Frechet-Darmois-Cramér-Rao bound a.k.a. the CRB...

Decompose the error $e = \widehat{\theta} - \theta$ made by an estimator $\widehat{\theta} = \mathcal{T}(X)$ as

$$\widehat{\theta} - \theta = e = e_{/\!/} + e_{\perp}$$

where $e_{/\!/}$ is the projection of e onto the linear span of $\varphi(X;\theta)$ and $e_{\perp}$ is the orthogonal (uncorrelated) part. The projected error $e_{/\!/}$ is:

$$e_{/\!/} = \mathbf{E}_\theta\left(e\ \varphi^\dagger\right)\mathbf{E}_\theta\left(\varphi\varphi^\dagger\right)^{-1}\varphi \qquad \text{with } \varphi = \varphi(X;\theta)$$

If $\mathcal{T}(x)$ is unbiased, $\mathbf{E}_\theta\left(e\ \varphi^\dagger\right)$ must be the identity matrix:

$$\mathbf{E}_\theta\left(e\ \varphi^\dagger\right) = \mathbf{E}_\theta\left((\widehat{\theta}-\theta)\varphi^\dagger\right) = \frac{\partial}{\partial\theta}\mathbf{E}_\theta\widehat{\theta} - \theta\,\mathbf{E}_\theta\varphi^\dagger = I - 0.$$

Hence, its parallel error is $e_{/\!/} = F_\theta^{-1}\varphi(X;\theta)$. It does *not* depend on $\mathcal{T}$!!

$$\widehat{\theta} = \theta + F_\theta^{-1}\varphi(X;\theta) + e_{\perp}$$

Hence the Cramér-Rao bound: covariance of the unavoidable error:

$$\mathbf{Cov}_\theta(\widehat{\theta}) = \mathbf{Cov}(e_{/\!/}) + \mathbf{Cov}(e_{\perp}) \geq \mathbf{Cov}(e_{/\!/}) = \mathbf{Cov}_\theta\left(F_\theta^{-1}\varphi\right) = F_\theta^{-1}$$

An *efficient estimator* is unique: it must behave as $\mathcal{T}(X) = \theta + F_\theta^{-1}\varphi(X;\theta)$.

Summary

We have a statistically well founded Riemmanian metric.

Further, it is **Kullback-compatible:** For infinitesimal $d\theta$,

$$K\left[\theta \,|\, \theta + d\theta\right] = K\left[\theta + d\theta \,|\, \theta\right] = \frac{1}{2} d\theta^{\dagger} F(\theta) d\theta$$

So, we are happy.

Tangent planes

Representing the tangent plane with random variables

- Tangent plane at point θ_* = span of the scores there.

$$\mathcal{T}(\theta_*) = \Big\{\alpha^\dagger S(x;\theta_*) \;=\; \sum_i \alpha_i S_i(x;\theta_*) \;=\; \sum_i \alpha_i \frac{\partial \log p(X;\theta_*)}{\partial \theta_i}\Big\}$$

So the tangent plane is a vector space of zero-mean scalar random variables with the norm controlled by the Fisher matrix

$$\|\alpha\|^2 = \mathbf{E}(\alpha^\dagger S(x;\theta_*))^2 = \alpha^\dagger F(\theta_*)\alpha$$

- More generally, for $P(x)$ a prob. dist. and $d(X)$ a zero-mean variable $\mathbf{E}_P d(X) = 0$, one can define a 'translation' $P \to Q : P\vec{Q} = \vec{d}$ by $\vec{d} = d(X)$ by:

$$Q(x) = P(x)\, e^{d(x) - \psi(\vec{d})} \qquad \psi(\vec{d}) = \log \mathbf{E}_P e^{d(X)} \approx \frac{1}{2}\mathbf{E}_P d^2(X) = \frac{1}{2}\|\vec{d}\|_P^2.$$

Note that $K\left[P \,|\, P + \vec{d}\right] = K\left[P \,|\, Q\right] = \mathbf{E}_P\Big(\log P/Q\Big) = \psi(\vec{d})$.

If $P(x) = p(x;\theta)$ and $d(x) = \alpha^\dagger S(x;\theta)$, the translated distribution stays in the model approximately. It stays in it **exactly** for an exponential family.

The exponential family tangent to a smooth model

- 2nd-order expansion around point θ_* of the log-density in a smooth model:

$$\log p(x;\theta) \approx \log p(x;\theta_*) + (\theta-\theta_*)^\dagger \frac{\partial \log p(x;\theta_*)}{\partial \theta} + \frac{1}{2}(\theta-\theta_*)^\dagger \frac{\partial^2 \log p(x;\theta_*)}{\partial \theta^2}(\theta-\theta_*)$$

- Exponential family based at θ_* with sufficient statistic $\varphi(x;\theta_*) = \frac{\partial \log p(x;\theta_*)}{\partial \theta}$:

$$p(x;\theta) = p(x;\theta_*) e^{(\theta-\theta_*)^\dagger \varphi(x;\theta_*) - \psi(\theta-\theta_*)}.$$

The potential ψ satisfies $\psi(0) = 0$ and $\frac{\partial \psi(0)}{\partial \theta} = \mathbf{E}_* \varphi(X;\theta_*) = 0$ and

$$\frac{\partial^2 \psi(0)}{\partial \theta^2} = \mathbf{Cov}_*(\varphi(X,\theta_*)) = F(\theta_*)$$

so, at second order,

$$\psi(\theta-\theta_*) \approx \frac{1}{2}(\theta-\theta_*)^\dagger F(\theta_*)(\theta-\theta_*).$$

Hence the two models are almost identical at second-order since

$$-\frac{\partial^2 \log p(x;\theta_*)}{\partial \theta^2} \approx -\mathbf{E}_* \frac{\partial^2 \log p(x;\theta_*)}{\partial \theta^2} = \mathbf{E}_* \frac{\partial \log p(x;\theta_*)}{\partial \theta} \frac{\partial \log p(x;\theta_*)}{\partial \theta}^\dagger = F(\theta_*)$$

and they become equivalent if many independent samples are available.

Smooth models as curved exponential families

For a smooth model, define $\ell(x;\theta) = \log p(x;\theta)$, $\theta \in \mathbb{R}^n$, pick a θ_* and define the higher order scores:

$$\ell_i(x) = \frac{\partial \ell(x;\theta_*)}{\partial \theta_i}, \quad \ell_{ij}(x) = \frac{\partial^2 \ell(x;\theta_*)}{\partial \theta_i \partial \theta_j}, \quad \ell_{ijk}(x) = \frac{\partial^3 \ell(x;\theta_*)}{\partial \theta_i \partial \theta_j \partial \theta_k},$$

In general $d = n + n(n+1)/2 + n(n+1)(n+2)/6$ such independent scores.

Then, build with them a d-dimensional exponential family $p_3(x;T)$, $T \in \mathbb{R}^d$

$$p_3(x;T) = p(x;\theta_*)\ e^{T^\dagger S(x) - \psi(T)} \quad \text{with} \quad S(x) \stackrel{\text{def}}{=} [\ell_i(x)\ \ell_{ij}(x)\ \ell_{ijk}(x)]^\dagger_{1\le i\le j\le k\le n}.$$

The third order expansion of the log-density $\log p(x;\theta_* + t)$ reads

$$\log p(x;\theta_* + t) \approx \log p(x;\theta_*) + \sum_i t_i\, \ell_i(x) + \sum_{ij} t_i t_j\, \ell_{ij}(x) + \sum_{ijk} t_i t_j t_k\, \ell_{ijk}(x) + \ldots$$

so the log-density $\log p(x;\theta)$ is approximated at third-order around θ_* by

$$\log p(x;\theta_* + t) \approx \log p_3(x;T(t)) \quad T(t) \stackrel{\text{def}}{=} [t_i\ t_{ij}\ t_{ijk}]^\dagger_{1\le i\le j\le k\le n}.$$

We have approximated a smooth model by a **curved exponential family**.
It is curved because the $\mathbb{R}^n \mapsto \mathbb{R}^d$ mapping $t \to T(t)$ is not linear.

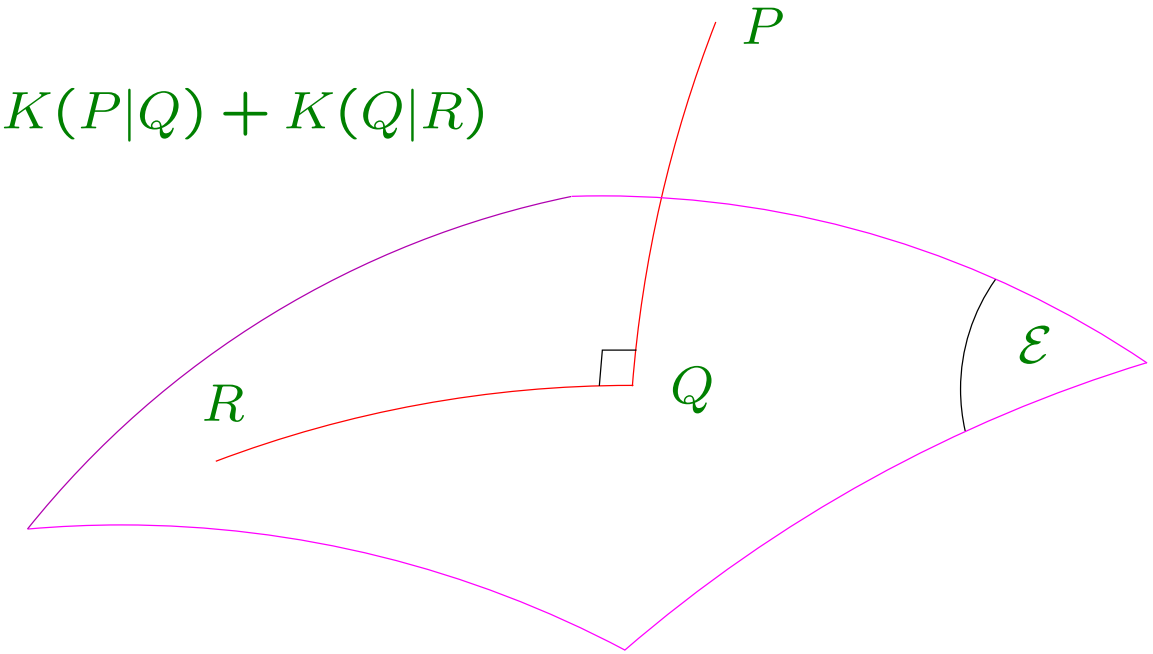
Summary

- Tangent plane tp $p(x;\theta)$ at θ_* as a vector space ...
- of zero-mean random variables ...
- spanned by the scores at θ_*.
- Exponential model tangent to a smooth family
- 'Osculating' higher-dimensional exponential models
  to embed (approximately) a smooth manifold as a curved submodel.

The Pythagorean theorem of information geometry.

The Pythagorean theorem of information geometry

$$K(P|R) = K(P|Q) + K(Q|R)$$



- $\mathcal{E}$: an exponential family of distributions.
- Q: the minimizer in $\mathcal{E}$ of $K(P\|\cdot)$. It is the m-projection of P onto $\mathcal{E}$.
- Then, for any other distribution R of $\mathcal{E}$

$$K(P\|R) = K(P\|Q) + K(Q\|R).$$

MLE and foliation `foliation here, maybe with maxent`

Kullback divergence in an exponential family

Recall the generic exponential family:

$$p(x;\theta) = g(x)\, e^{\theta^\dagger S(x) - \psi(\theta)} \qquad \psi(\theta) = \log \int g(x)\, e^{\theta^\dagger S(x)}\, dx$$

$$\frac{\partial \psi(\theta)}{\partial \theta} = \mathbf{E}_\theta S(x) \stackrel{\text{def}}{=} \eta(\theta) \qquad \frac{\partial^2 \psi(\theta)}{\partial \theta^2} = \mathbf{Cov}_\theta S(x)$$

Since $\psi(\theta)$ is convex, there is a 1-to-1 mapping $\theta \longleftrightarrow \eta = \frac{\partial \psi(\theta)}{\partial \theta}$. Any point is uniquely labeled with $\eta = \eta(\theta) = \mathbf{E}_\theta S(x)$, the conjugate variable.

The convex conjugate of $\psi(\theta)$ is $\varphi(\eta)$ which is minus the intercept of the tangent plane with slope η, so $\psi(\theta) = -\varphi(\eta) + \theta^\dagger \eta$ or

$$\varphi(\eta) = \theta^\dagger \eta - \psi(\theta) \quad \text{for } \theta \text{ s.t. } \frac{\partial \psi(\theta)}{\partial \theta} = \eta$$

Then $\psi(\theta) \geq -\varphi(\eta) + \eta^\dagger \theta$ for all θ, η with equality if $\frac{\partial \psi(\theta)}{\partial \theta} = \eta$. Then

$$K\,[1\,|\,2] = \varphi(\eta_1) + \psi(\theta_2) - \eta_1^\dagger \theta_2 \qquad \text{Kullback as Bregman}$$

The dual potential functions can also be expressed as

$$\varphi(\eta) = K\,[p(x;\eta)\,|\,g(x)] \qquad \psi(\theta) = K\,[g(x)\,|\,p(x;\theta)]$$

Euclidean vs Information geometry

- Do we have, *non locally in distribution space* something like this?

$$\|\vec{AC}\|^2 = \|\vec{AB} + \vec{BC}\|^2 = \|\vec{AB}\|^2 + \|\vec{BC}\|^2 - 2\vec{BA} \cdot \vec{BC}$$

- Take any three distributions P, Q, R, and check:

$$K\left[P \,|\, R\right] - K\left[P \,|\, Q\right] - K\left[Q \,|\, R\right] = \int P \log \frac{P}{R} - \int P \log \frac{P}{Q} - \int Q \log \frac{Q}{R} = \int (P-Q) \log \frac{Q}{R}$$

- Making sense of the difference:

$$\int (P-Q) \log \frac{Q}{R} = -\int Q \left(\frac{P}{Q} - 1\right) \left(\log \frac{R}{Q} + K\left[Q \,|\, R\right]\right)$$

- Everything is *beautiful* with mixture and exponential vectors:

$$K\left[P \,|\, R\right] = K\left[P \,|\, Q\right] + K\left[Q \,|\, R\right] \; - \; \langle \, \vec{QP}^{\,m}, \vec{QR}^{\,e} \, \rangle_Q$$

$$\vec{QP}^{\,m} = \frac{P(X)}{Q(X)} - 1 \qquad \vec{QR}^{\,e} = \log \frac{R(X)}{Q(X)} + K\left[Q \,|\, R\right]$$

and $\langle \, \vec{QP}^{\,m}, \vec{QR}^{\,e} \, \rangle_Q = 0$ for the Pythagorean theorem.

A dual mixture/exponential structure is hidden in the KLD.

Back to Hardy-Weinberg

Parallel transport and connections

Connecting tangent planes, the regular way

Let $V_i(\theta)$ be a basis for the tangent plane at $\theta = [\theta^1, \ldots, \theta^n]$.

The scalar product between two vectors at θ:

$$\langle X|Y\rangle = \langle \sum_i x^i V_i | \sum_j y^j V_j \rangle = \sum_{ij} x^i y^j \langle V_i | V_j \rangle \stackrel{\text{def}}{=} \sum_{ij} x^i y^j \; g_{ij}(\theta)$$

Variation of a vector field $X(\theta) = \sum_i x^i(\theta) V_i(\theta)$:

$$\delta X(\theta) = \sum_i \delta[x^i V_i] = \sum_i \sum_j \Big[\frac{\partial x^i}{\partial \theta^j} V_i + x^i \frac{\partial V_i}{\partial \theta^j}\Big] \delta\theta^j$$

First order variation of the basis controlled by n^3 Christoffel symbols Γ^k_{ij}:

$$\frac{\partial V_i}{\partial \theta^j} = \sum_k \Gamma^k_{ij} V_k$$

The metric g_{ij} and the Christoffel symbols Γ^k_{ij} fully determine the geometry.

You/the math/the physics/the stats decide what to pick for $g_{ij}(\theta)$ and $\Gamma^k_{ij}(\theta)$...

...but a unique connection can be derived from the metric.

Statistical connections

Connections can be determined from statistical principles. Define:

$$\ell_i = \ell_i(x;\theta) = \frac{\partial \log p(x;\theta)}{\partial \theta_i} \qquad \ell_{ij} = \ell_{ij}(x;\theta) = \frac{\partial^2 \log p(x;\theta)}{\partial \theta_i \partial \theta_j}$$

The metric (Levi-Civita) connection for the Fisher metric $g_{ij}(\theta) = \mathbf{E}_\theta \ell_i \ell_j$ is

$$\Gamma^0_{ij,k}(\theta) \;=\; \mathbf{E}_\theta\Big((\ell_{ij} + \frac{1}{2}\ell_i\ell_j)\; \ell_k\Big)$$

but we do no want it. Amari introduced a pair of dual connections:*

$$\Gamma^{(e)}_{ij,k}(\theta) = \mathbf{E}_\theta\Big(\ell_{ij}\ell_k\Big) \qquad \Gamma^{(m)}_{ij,k}(\theta) = \mathbf{E}_\theta\Big((\ell_{ij} + \ell_i\ell_j)\ell_k\Big)$$

$\Gamma^{(e)}_{ij,k} = 0$ in an exponential model. $\Gamma^{(m)}_{ij,k} = 0$ in a mixture model.

The scalar product of two vectors is preserved during parallel transport along any curve if one is e-transported while the other is m-transported.

But we are not going to do that.

*Actually, he introduced a 1D family of such pairs of dual connections.

Some parallel transport

Dual parallel transport preserves the scalar product.

Let a and b be two vectors of the tangent plane at P,
represented by two random random variables with zero mean under P.

$$\mathbf{E}_P a(X) = \mathbf{E}_P b(X) = 0 \text{ with scalar product } \langle a|b\rangle_P = \mathbf{E}_P\left(a(X)b(X)\right)$$

The vectors are parallelly transported to Q.
Vector a undergoes an e-parallel transport and becomes a^e.
Vector b undergoes an m-parallel transport and becomes b^m.

How does parallel transport work here ?

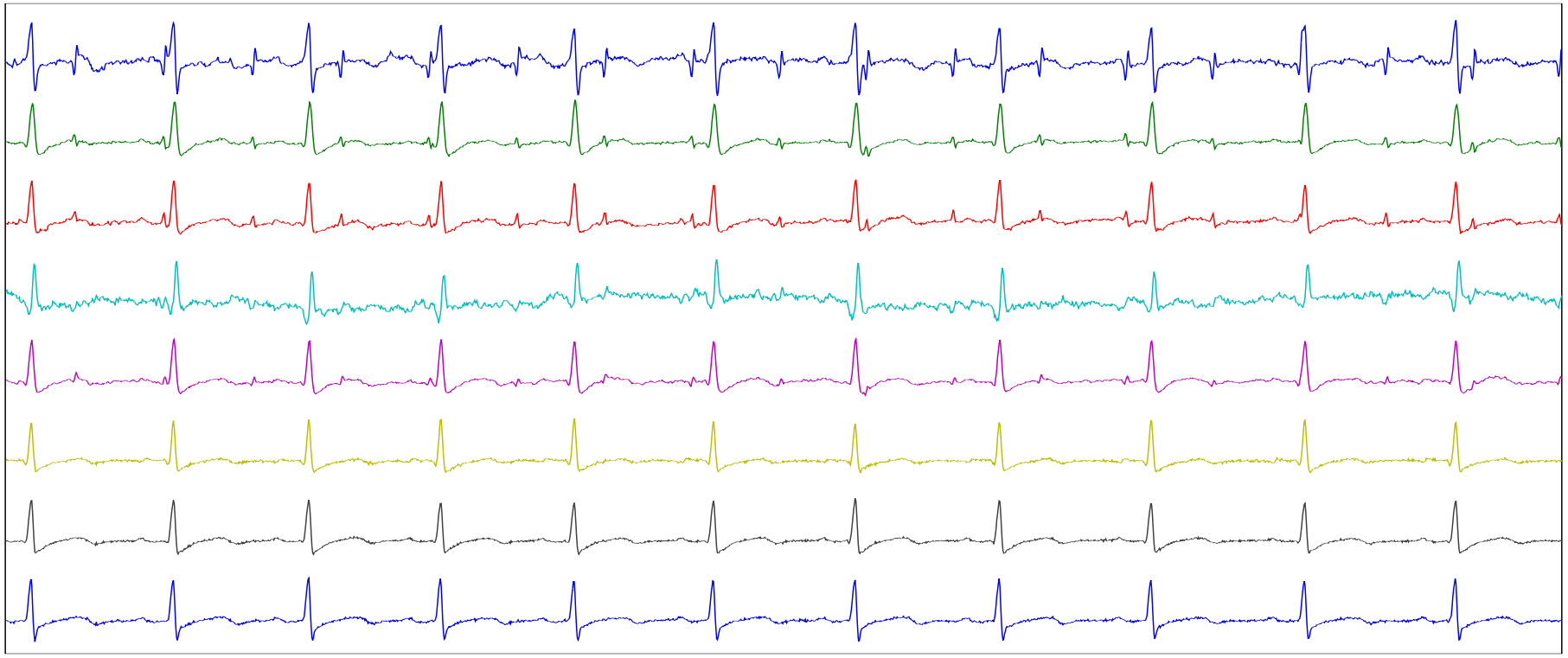
$$a^e(X) = a(X) - \mathbf{E}_Q\ a(X) \qquad b^m(X) = b(X)\ \frac{p(X)}{q(X)}$$

Check that $a^e(X)$ and $b^m(X)$ have zero-mean under Q and that

$$\langle a|b\rangle_P = \langle a^e|b^m\rangle_Q$$

Information geometry of ICA

Multi-channel time series: a biomedical example

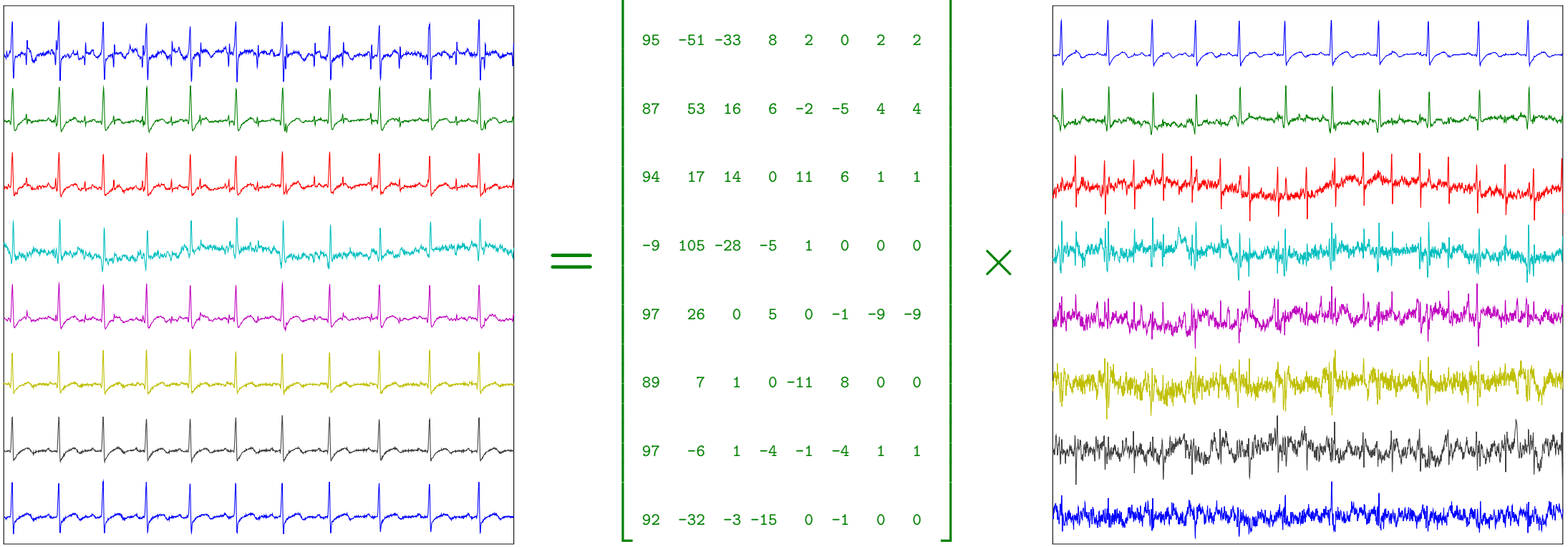


8 ECG electrodes located on the thorax and the abdomen of a pregant woman.

Looking for linear decompositions: Data = Mixing matrix $\times$ Sources.

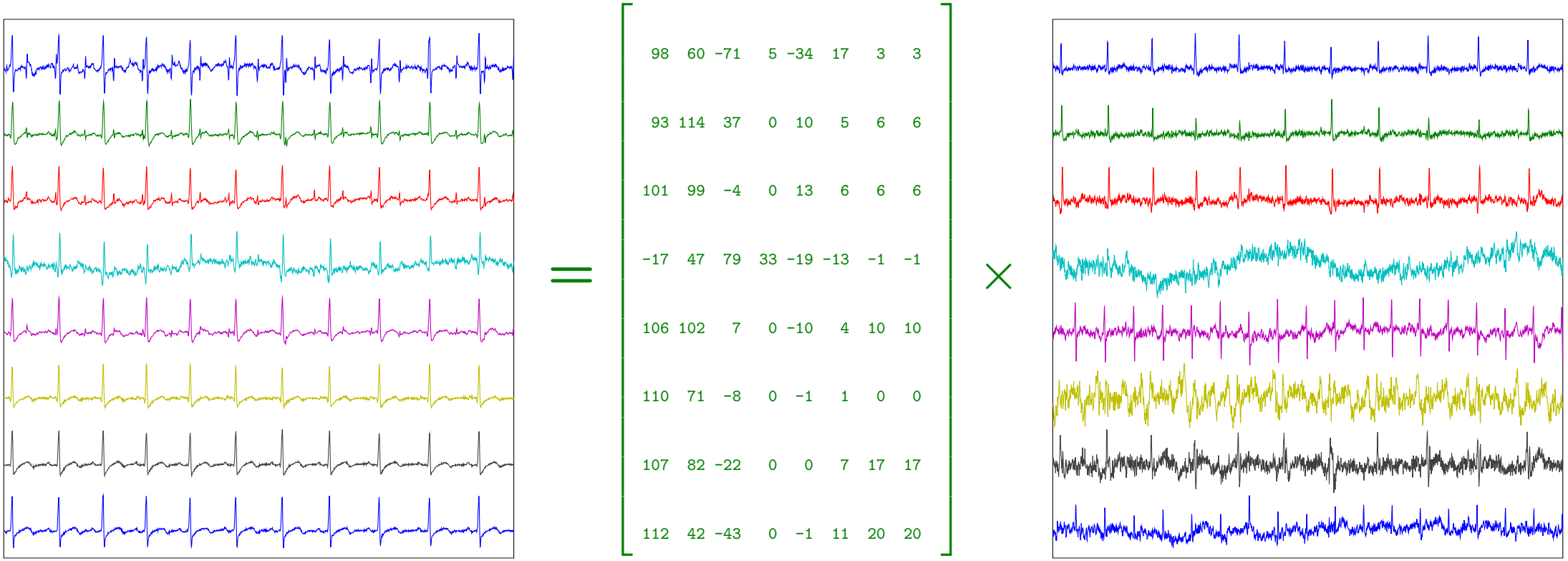
Data Credits: Daisy database

Principal component analysis



- Orthogonal mixture, uncorrelated components $\frac{1}{T}\sum_t y_i(t)y_j(t) = 0$ for $i \neq j$
- Decorrelation is weak (always posible), orthogonality is implausible.

Independent component analysis

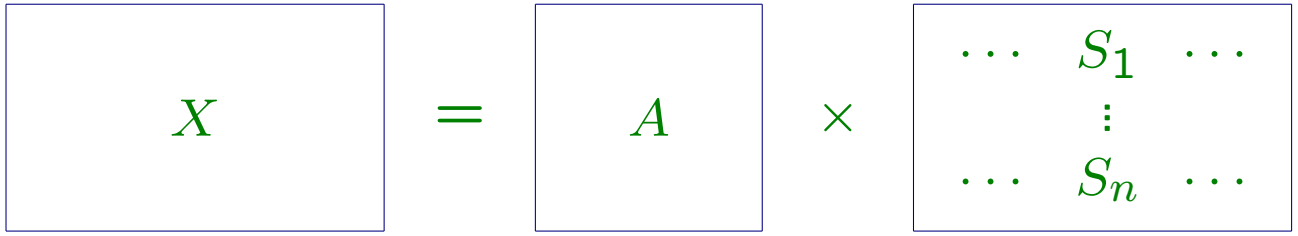


- Linear decomposition into "the most independent sources"
- Blind: only independence is at work but it must go beyond decorrelation.
- Independence is statistically very strong but often physically plausible.
- Weak assumptions $\longrightarrow$ wide applicability

The basic ICA model

- An $n \times T$ data set $X = \{x_i(t) \mid 1 \leq i \leq n,\ 1 \leq t \leq T\}$ ...

  ... modelled as $X = AS$ with an $n \times T$ source matrix of *independent* rows.



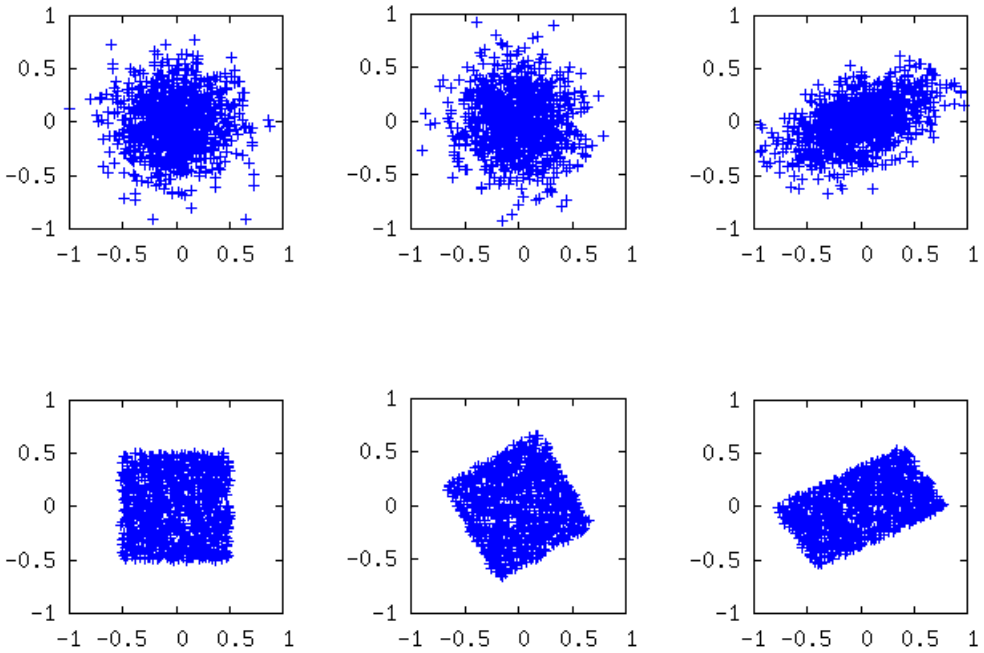
Assuming $X = AS$, is the assumption of statistical independence sufficient to recover the sources *blindly* and *efficiently* ?

Blindness: why not? If mixing always destroys independence, then recovering independence should unmix, *shouldn't it*?

The crux of the matter

Blind methods good statistical descriptors.

Mixing Gaussian (top row) or uniform(bottom) random variables
with the identity (left), a rotation (center), a generic transform (right):



Rotation is invisible unless the data **and** the model are non Gaussian.

Mere decorrelation, $\frac{1}{T}\sum_t y_i(t)y_j(t) = 0$, does not cut it.

How to do it ?

- Ignore time/space structure but use non Gaussianity . . .
  - Minimize dependence between recovered sources . . .
  - . . . as measured by mutual information, *why ? hard !*
  - . . . or approximated using high order cumulants, *clumsy ?*
  - or find the most non Gaussian sources *why ? how ?*
  - or find sources uncorrelated through non linear functions:
    e.g $\frac{1}{T}\sum_t \phi(y_i(t))\psi(y_j(t)) = 0$ for $i \neq j$. Which functions ϕ, ψ ?
- . . . or assume Gaussianity but use temporal/spatial structure, such as
  - . . . correlations, or spectral diversity,
  - . . . or non stationarity / inhomogeneities

But how to make sense of all that ? How to do it properly ? *Statistics!*

Likelihood and Kullback matching

The distribution of $X = AS$ is specified by $\theta = (A, P_{S_1}, \ldots, P_{S_n})$.

	$-\log P_\theta(X)$	- log-likelihood
$=$	$-E_X \log P_\theta(X) + \text{stoch.}$	The average likelihood landscape
$\stackrel{c}{=}$	$K\left[P_X \,|\, P_\theta\right] + H(X)$	$H(X) = \text{entropy} = \text{cst}$
$\stackrel{c}{=}$	$K\left[P_X \,|\, P_{AS}\right]$	The model is $X = AS$.
$=$	$K\left[P_{A^{-1}X} \,|\, P_S\right]$	Invariance under invertible transforms.
$=$	$K\left[P_Y \,|\, P_S\right]$	No X, no A, only $Y \stackrel{\text{def}}{=} A^{-1}X$.

The likeliest A is the one making P_Y the closest to $P_S = \prod_i P_{S_i}$. Nice !

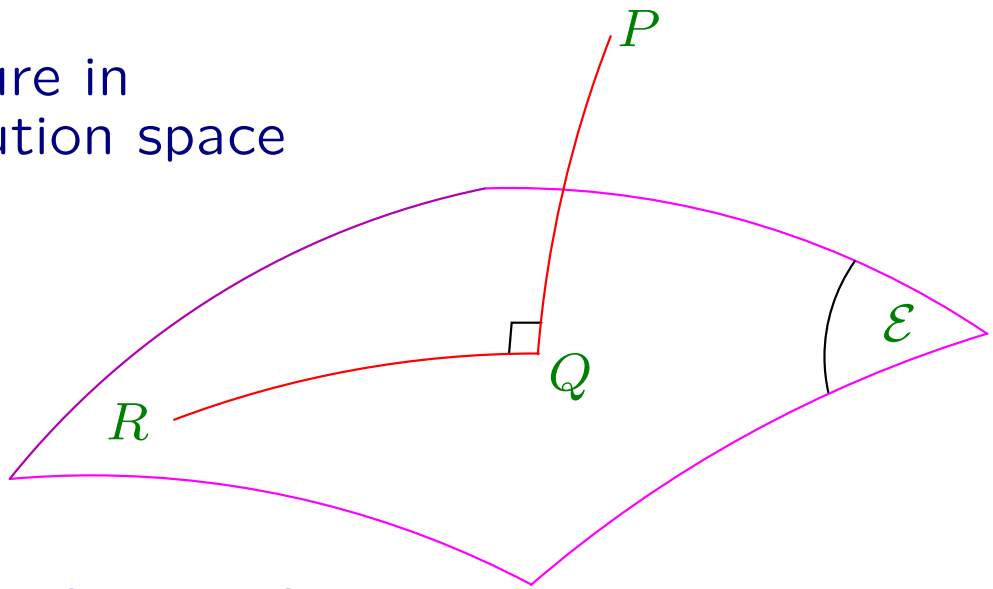
$$P_X \longrightarrow \boxed{A^{-1}} \longrightarrow P_Y \stackrel{\mathcal{D}}{\approx} P_S = \prod_i P_{S_i}$$

Data X and mixing A enter only through $Y = A^{-1}X$: equivariance.

The Pythagorean theorem of information geometry

The Kullback divergence may not be a distance,
it still has its own private Pythagoras theorem.

A picture in
distribution space



- $\mathcal{E}$: an *exponential* family of distributions.
- Q: the minimizer in $\mathcal{E}$ of $K[P\,|\,\cdot]$. It is a projection of P onto $\mathcal{E}$.
- Then, for any other distribution R of $\mathcal{E}$

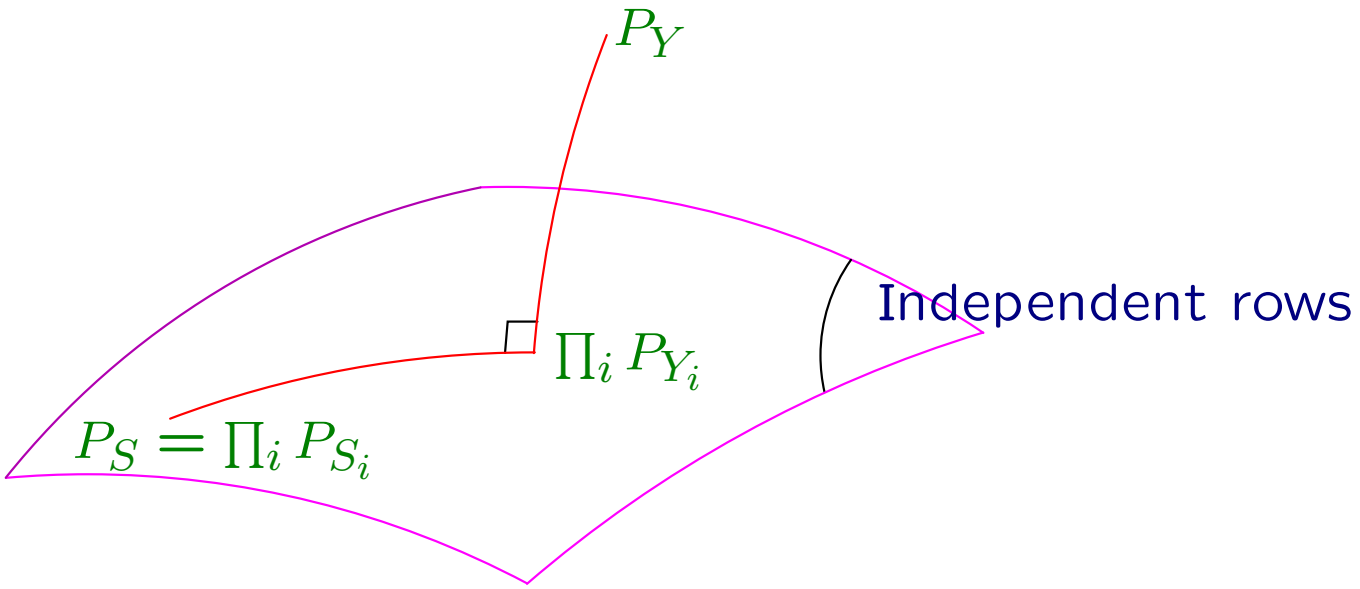
$$K[P\,|\,R] = K[P\,|\,Q] + K[Q\,|\,R]$$

The Kullback divergence behaves here as a squared Euclidean distance.

Likelihood and mutual information

- Maximizing the likelihood $p(X|A)$ for $X = AS$ and $S \sim P_S$ is equivalent to minimizing the (empirical) Kullback mismatch $K\left[P_Y \,|\, P_S\right]$ with $Y = A^{-1}X$.

- The target source distribution P_S is usually unknown except for $P_S = \prod_i P_{S_i}$.



$$\begin{aligned} K\left[P_Y \,|\, P_S\right] &= K\left[P_Y \,|\, \textstyle\prod_i P_{Y_i}\right] + K\left[\textstyle\prod_i P_{Y_i} \,|\, \prod_i P_{S_i}\right] \\ &= I(Y) + \textstyle\sum_i K\left[P_{Y_i} \,|\, P_{S_i}\right] \\ &= \text{dependence} + \text{sum of marginal mismatches} \end{aligned}$$

The likelihood suggests to measure dependence
by the mutual information $I(Y) = K\left[P_Y \,|\, \prod_i P_{Y_i}\right]$.

Form Kullback matching to mutual information

- Maximizing the ICA likelihood is equivalent to minimizing $K\left[P_Y \,|\, P_S\right]$.

- The target source distribution P_S is usually unknown except for $P_S = \prod_i P_{S_i}$.

$$K\left[P_Y \,|\, P_S\right] = I(Y) + \sum_i K\left[P_{Y_i} \,|\, P_{S_i}\right]$$

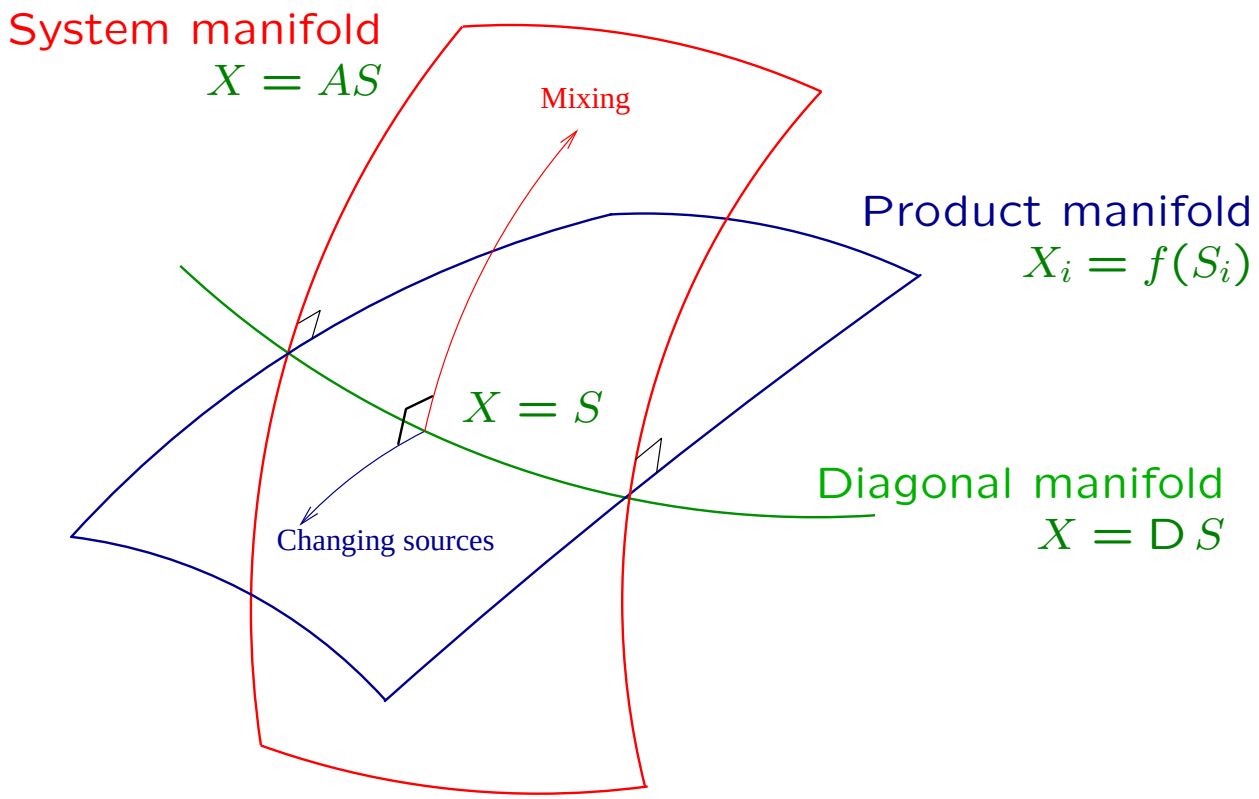
Optimizing over nuisance parameters P_{S_i} kills each $K\left[P_{Y_i} \,|\, P_{S_i}\right]$ and leaves us with *(in)dependence*:

$$I(Y) \stackrel{\text{def}}{=} K\left[P_Y \,|\, \prod_i P_{Y_i}\right] \qquad \text{a.k.a. } \textit{mutual information}$$

$\rightarrow$ Maximum likelihood leads to maximum independence. . .
$\rightarrow$ . . . and provides a definition for it.

A key property:
mixing matrix and source distributions are orthogonal parameters.
$\rightarrow$ Black board time.

Some statistical manifolds of ICA



- $X = S$: the distribution of a vector S with independent entries. Move from there.
- The *system manifold*: the family of distributions of $X = AS$ for all invertible matrices A.
- The *product manifold*: distort the marginals of S, retaining independence.
- The *diagonal manifold*: change only the scales of the entries of S.

$\rightarrow$ ICA is doable: the system manifold intersects the product manifold along the diagonal manifold *only*. And even easy by nuisance orthogonality.
But breaks for Gaussian S...

Classic ICA: mining non Gaussianity

Classic ICA ignores any 'time' structure and thus **must** rely on non Gaussianity. It assumes (explicitly or not) i.i.d. sequences:

$$\begin{aligned} P(Y) &= P(Y(1), Y(2), \ldots, Y(T)) \\ &= \prod_t P_t(Y(t)) \quad \text{Independently and } \ldots \\ &= \prod_t p(Y(t)) \quad \ldots \text{ identically distributed.} \end{aligned}$$

where $p()$ is the n-dimensional pdf common to all $Y(t)$, $1 \leq t \leq T$.

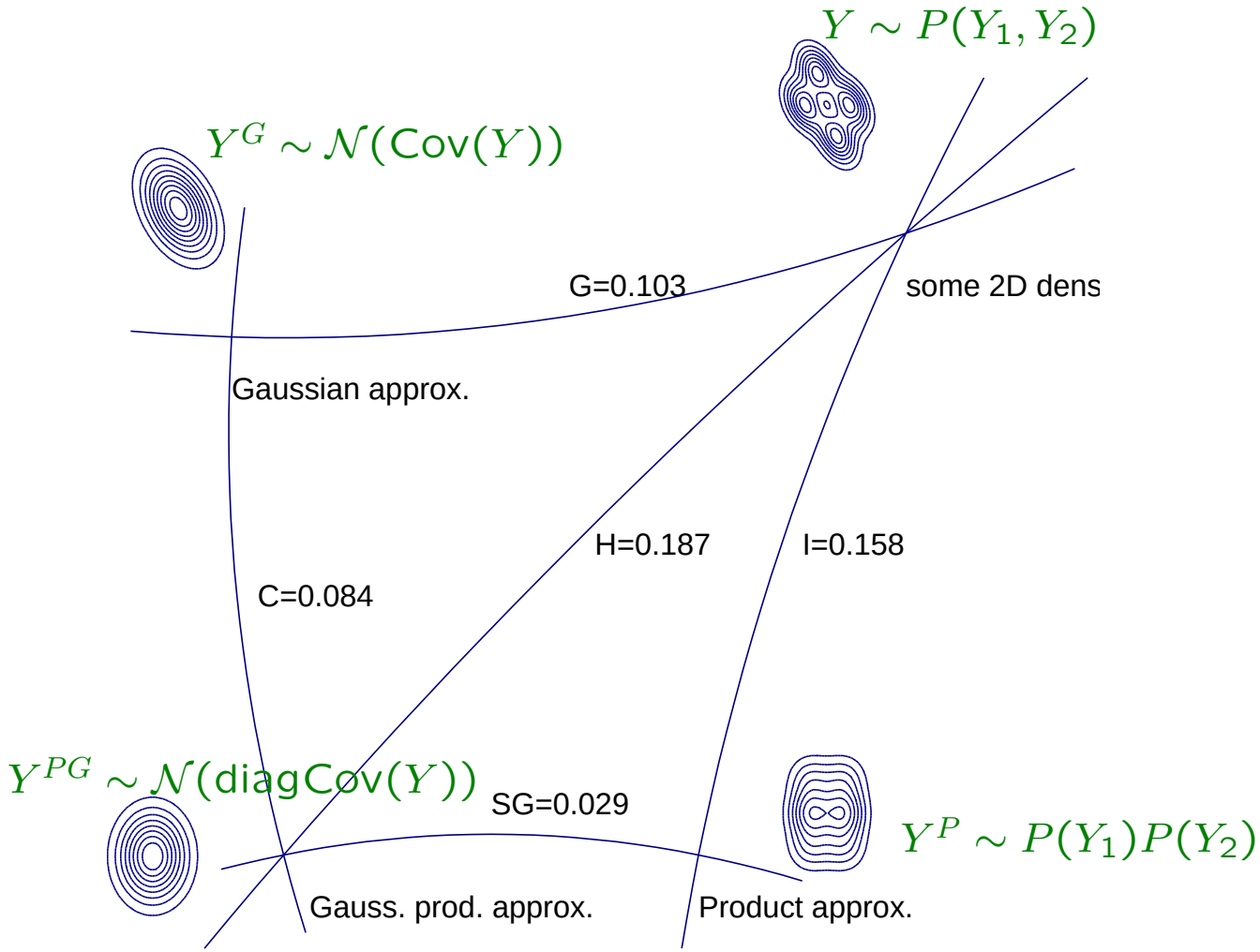
Then, divergence between T-long time series distributions reduces to

$$K[P \,|\, Q] = T \times K[p \,|\, q]$$

depending only on the n-dimensional distributions p and q.

Note: for readibility, we drop the T factor in the following.

Geometry of non Gaussian dependence

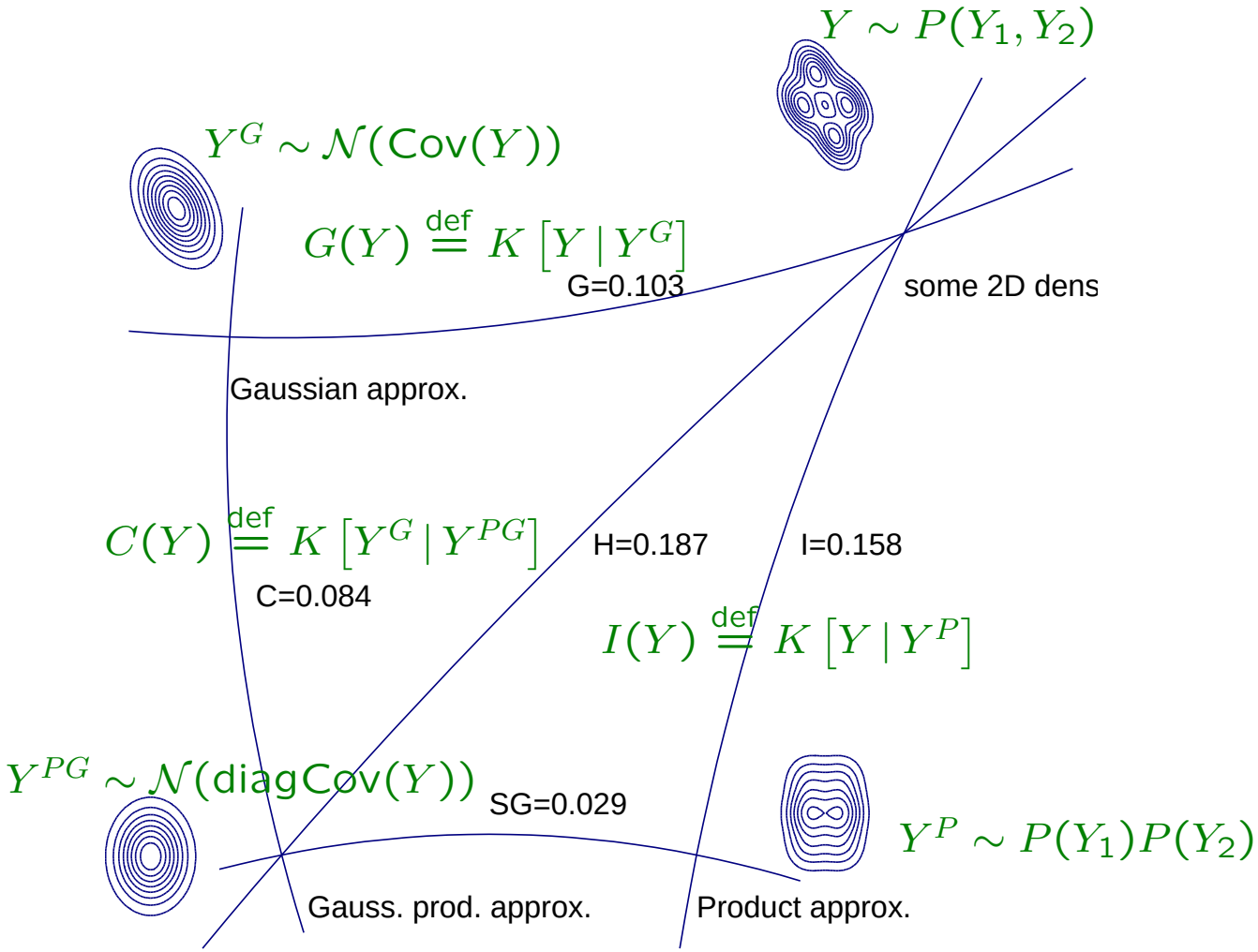


Two familiar statistical approximations can be seen as projections:

Oh, let's assume it's Gaussian...

Ho, let's assume they're independent...

Geometry of non Gaussian dependence



Orthogonal projections $\rightarrow$ two right triangles $\rightarrow$ two Pythagorean theorems

Dependence and non Gaussianity

- *Non Gaussianity*. Define the non Gaussianity $G(Y)$ of Y as

$$G(Y) = K\left[P_Y \,|\, \mathcal{N}(R_Y)\right]$$

i.e. how much the best Gaussian approx. fails to mimic the distrib. of Y.

- The *correlation* $C(Y)$ of Y

$$C(Y) = K\left[\mathcal{N}(R_Y) \,|\, \mathcal{N}(\mathrm{diag} R_Y)\right]$$

i.e. how much the covariance matrix R_Y of Y fails to be diagonal.

- All these are (geometrically) connected by

$$I(Y) + \sum_i G(Y_i) = C(Y) + G(Y)$$

- Under *linear* transforms, $G(Y)$ is constant. The mutual information then is

$$I(Y) = C(Y) - \sum_i G(Y_i) + \mathrm{cst}$$

$\rightarrow$ Under linear transforms, making the entries of Y as independent as possible ($\min I(Y)$) is *identical* to making (as much as possible) Y uncorrelated and *each of its entries* non Gaussian.

Non Gaussianity (cont.)

Repeat: Under linear transforms, making the entries of Y as independent as possible ($\min I(Y)$) is *identical* to making (as much as possible) Y uncorrelated and *each of its entries* non Gaussian.

- Note 1: The relation $I(Y) = C(Y) - \sum_i G(Y_i) + G(Y)$ also reads

$$\text{Complicated} = \text{Simple - Simples} + \text{Complicated constant}$$

- Note 2: Some people/algos *enforce* $C(Y) = 0$ (pre-whitening) but even that leaves us with with the not-so-easy-after-all measure of marginal non-Gaussianity $G(Y_i)$. So, how do we do that?

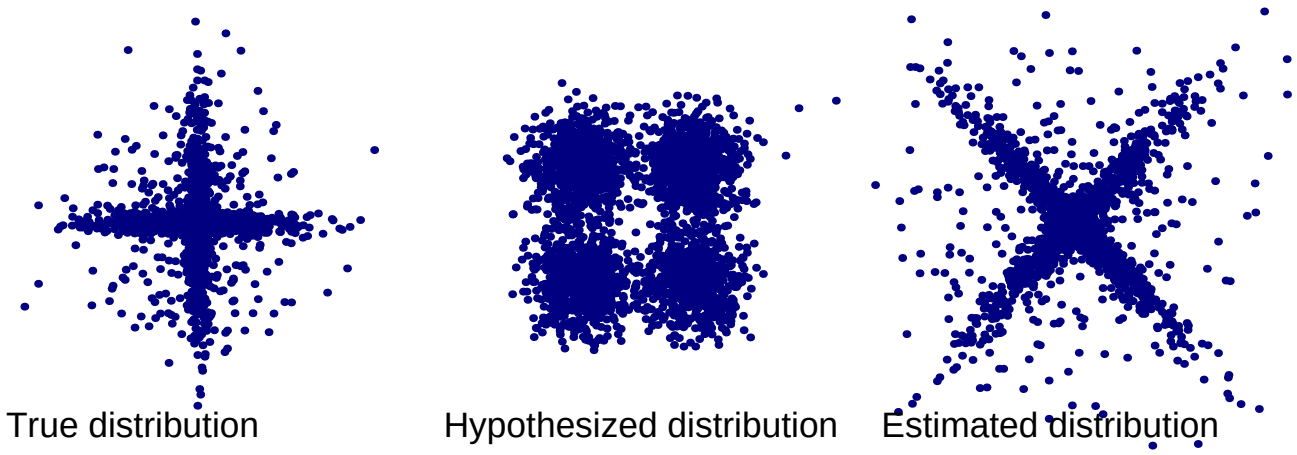
- Note 3: Connection between non Gaussianity and sparsity.

Non Gaussianity, sparsity and bad luck

To avoid estimating (explicitly or implicitly) the source distributions, one may use fixed targets $P_{S_i} = Q_i$ for each component and just minimize $K\left[P_Y \mid \prod_i Q_i\right]$.

If the pdf's of the sources are known, one should do just that, indeed. If the pdf's are *not* known but are expected to be "sparse", then one could use *some* sparse guess Q_i in place of the true but unknown distribution,

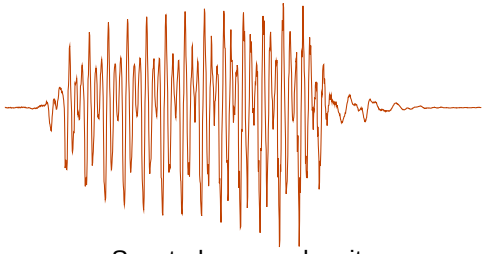
This works (provably) provided the mismatch between P_{S_i} and Q_i is not too large. It does *not* work if a sparse model is used and the true sources are "anti-sparse" or vice-versa.



It's not only wrong; it's maximally wrong!

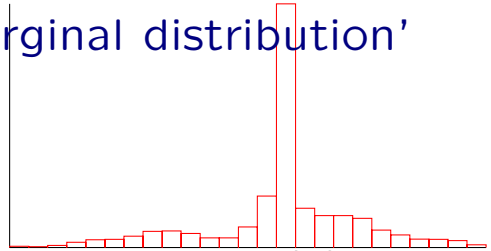
Other geometries: three points of view on a time series.

A random (?!) sequence

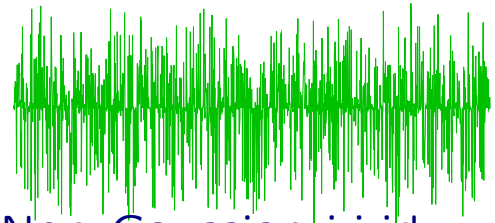


Marginal probability density

Its 'marginal distribution'



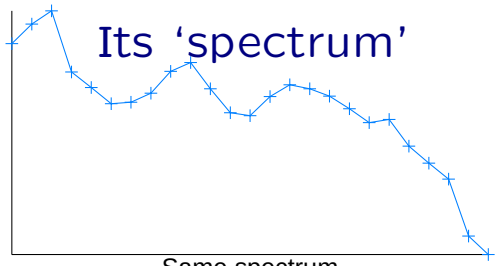
Same marginal



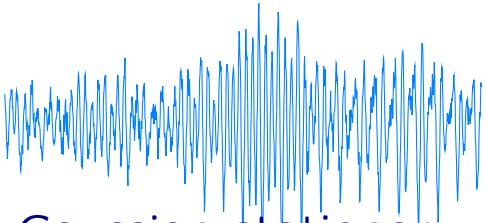
Non Gaussian i.i.id

Spectral energy density

Its 'spectrum'



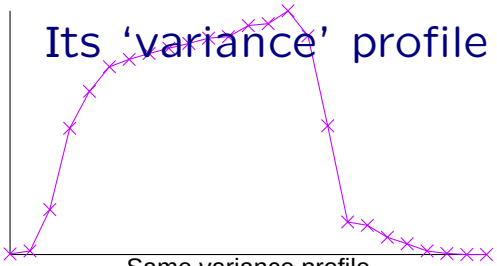
Same spectrum



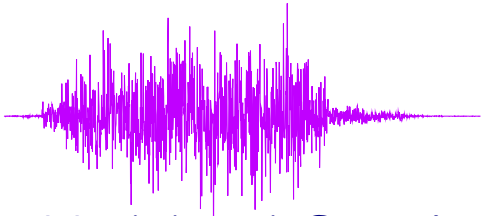
Gaussian stationary

Temporal energy density

Its 'variance' profile



Same variance profile



Modulated Gaussian i.i.d.

- *All models are wrong, but some are useful* — George Box

Other instances of mutual information by projecting the data onto simple Gaussian models for time series:
stationary but colored, white but non non stationary.

Conclusion

Some nice conlusion