电子科技大学
University of Electronic Science and Technology of China

Probabilistic Graphical Model
Chapter 7 & 8

Reporter: Xinzuo Wang

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Ch.7 Outline:

- ***Gaussian Network Models***
  - **Properties of Multivariate Gaussians**
    1. **Operations of Gaussians (i.e.** marginalization, conditioning**)**
    2. **Independencies in Gaussians**
  - **Gaussian Bayesian Networks**
  - **Gaussian Markov Random Fields**

Multivariate Gaussians :

- **Multivariate Gaussians**

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

, where $|\Sigma|$ is the determinant of Σ which should be ***positive definite.***

- **Information matrix and information form**

Let $\boxed{J = \Sigma^{-1}}$, thus

Information matrix

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T J(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= -\frac{1}{2}\left[\boldsymbol{x}^T J\boldsymbol{x} - 2\boldsymbol{x}^T J\boldsymbol{u} + \boldsymbol{\mu}^T J\boldsymbol{\mu}\right]. \end{aligned}$$

and

$$\boxed{p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^T J\boldsymbol{x} + \boxed{(J\boldsymbol{\mu})}^T\boldsymbol{x}\right]}.$$

Potential vector

Information form

Properties of Gaussians – operations :

数据挖掘实验室
Data Mining Lab

- **A little trick -- 'Completing the square'**

-- A Gaussian distribution is totally determined by its μ & Σ, i.e. the quadratic form

Let $\mathrm{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\quad \mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$

Consider :

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \underline{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathrm{const}} \tag{1}$$

$$\begin{aligned} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\ -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\ -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b). \end{aligned} \tag{2}$$

For example

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \qquad \begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \end{aligned}$$

Properties of Gaussians – operations :

- **Operation – conditioning & marginalization :**

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}.$$

Conditional distribution:

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b). \end{aligned}$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}).$$

Properties of Gaussians – Independencies:

数据挖掘实验室
Data Mining Lab

- The relationship between **variables**
  - Determined by ***covariance matrix*** .

$Theorem\ 7.3\ (without\ proof)$: $Let\ \boldsymbol{X} = X_1, \dots, X_n\ have\ a\ joint\ normal\ distribution\ N\ (\boldsymbol{\mu}; \Sigma).\ Then\ \boldsymbol{X_i}$ ***and*** $\boldsymbol{X_j}$ ***are independent if and only if*** $\boldsymbol{\Sigma_{ij} = 0}$.

Properties of Gaussians – Independencies:

数据挖掘实验室 Data Mining Lab

- The relationship between **Gaussians** and **graph structures**
  - Independence structure in the distribution is ***apparent*** in the ***information matrix***.

$Theorem\ 7.3\ (without\ proof)$: $Consider\ a\ Gaussian\ distribution\ p(X_1, \ldots, X_n) = N\,(\boldsymbol{\mu};\ \Sigma), and\ let\ J = \Sigma^{-1}\ be\ the\ information\ matrix.\ Then\ J_{i,j} = 0\ \ if\ and\ only\ if\ \ p \mid = (X_i \perp X_j \mid X - \{X_i, X_j\}).$

Indicating Pairwise Markov independencies

Information matrix $\Longleftrightarrow$ defines $\Longleftrightarrow$ A minimal I-map Markov network for p

Gaussian Bayesian Networks:

数据挖掘实验室
Data Mining Lab

- **Conditional -> Joint:**

Let Y be a linear Gaussian of its parents $X_1, \ldots, X_k$:

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}; \sigma^2\right).$$

Assume that $X_1, \ldots, X_k$ are jointly Gaussian with distribution $\mathcal{N}(\boldsymbol{\mu}; \Sigma)$. Then:

- *The distribution of Y is a normal distribution $p(Y) = \mathcal{N}\left(\mu_Y; \sigma_Y^2\right)$ where:*

$$\begin{aligned} \mu_Y &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu} \\ \sigma_Y^2 &= \sigma^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}. \end{aligned}$$

- *The joint distribution over $\{\boldsymbol{X}, Y\}$ is a normal distribution where:*

$$\mathit{Cov}[X_i; Y] = \sum_{j=1}^{k} \beta_j \Sigma_{i,j}.$$

- **Joint -> Conditional (the same as above)**

Gaussian Markov Random Fields:

数据挖掘实验室
Data Mining Lab

- **Gaussian distribution -> Pairwise Markov networks:**
  - Note that ***the Pairwise Markov independencies*** are indicated by the ***information matrix of a Gaussian distribution***.

  - ***Node potentials*** are derived from h and J_{ii};
  - ***Edge potentials*** are derived from the off-diagonal entries of the information matrix.

By breaking up the expression in the exponent into two types of terms:

$$-\frac{1}{2}J_{i,i}x_i^2 + \boldsymbol{h}_i x_i, \quad -\frac{1}{2}[J_{i,j}x_ix_j + J_{j,i}x_jx_i] = -J_{i,j}x_ix_j,$$

Gaussian Markov Random Fields:

数据挖掘实验室
Data Mining Lab

- **Pairwise Markov networks (Gaussian Markov networks) -> Gaussian distribution :**

Consider any pairwise Markov network ***with quadratic node and edge potentials***.

$$\epsilon_i(x_i) = d_0^i + d_1^i x_i + d_2^i x_i^2$$
$$\epsilon_{i,j}(x_i, x_j) = a_{00}^{i,j} + a_{01}^{i,j} x_i + a_{10}^{i,j} x_j + a_{11}^{i,j} x_i x_j + a_{02}^{i,j} x_i^2 + a_{20}^{i,j} x_j^2.$$

$$\Downarrow$$

$$p'(\boldsymbol{x}) = \exp(-\frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{x}) \longrightarrow J \text{ should be positive definte !}$$

There is no simple way to check whether the MRF is valid!
But we do have some simpler sufficient conditions (see p255, 256).

Ch.8 Outline:

- *The Exponential Family*
  - Exponential Families
  - Factored Exponential Families
    - Product Distributions
    - Bayesian Networks
  - Entropy and Relative Entropy
  - Projections
    - M-projection
    - I-projection

Exponential Family -- definition:

数据挖掘实验室
Data Mining Lab

Let $\mathcal{X}$ be ***a set of variables***. An ***Exponential Family*** P over $\mathcal{X}$ is specified by four components:

- *A* $\boxed{\text{sufficient statistics function } \tau}$ *from assignments to* $\mathcal{X}$ *to* $\mathcal{R}^K$.
- *A* parameter space *that is a convex set* $\Theta \subseteq \mathcal{R}^M$ *of* legal parameters.
- *A* natural parameter *function* t *from* $\mathcal{R}^M$ *to* R^K.
- *An auxiliary measure* A *over* $\mathcal{X}$.

Each vector of parameters $\boldsymbol{\theta} \in \Theta$ *specifies a distribution* $P_{\boldsymbol{\theta}}$ *in the family as*

$$\boxed{P_{\boldsymbol{\theta}}(\xi) = \frac{1}{Z(\boldsymbol{\theta})} A(\xi) \exp\{\langle \mathrm{t}(\boldsymbol{\theta}), \tau(\xi)\rangle\}}$$

where $\langle \mathrm{t}(\boldsymbol{\theta}), \tau(\xi)\rangle$ *is the inner product of the vectors* $\mathrm{t}(\boldsymbol{\theta})$ *and* $\tau(\xi)$*, and*

$$Z(\boldsymbol{\theta}) = \sum_{\xi} A(\xi) \exp\{\langle \mathrm{t}(\boldsymbol{\theta}), \tau(\xi)\rangle\}$$

Exponential Family – an example:

数据挖掘实验室
Data Mining Lab

Consider a Gaussian distribution over a single variable. Recall that

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}.$$

Define

$$\begin{aligned} \tau(x) &= \langle x, x^2 \rangle \\ \mathsf{t}(\mu, \sigma^2) &= \langle \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \rangle \\ Z(\mu, \sigma^2) &= \sqrt{2\pi}\sigma \exp\left\{ \frac{\mu^2}{2\sigma^2} \right\}. \end{aligned}$$

$$\boxed{P_{\boldsymbol{\theta}}(\xi) = \frac{1}{Z(\boldsymbol{\theta})} A(\xi) \exp\left\{ \langle \mathsf{t}(\boldsymbol{\theta}), \tau(\xi) \rangle \right\}}$$

We can easily verify that

$$P(x) = \frac{1}{Z(\mu, \sigma^2)} \exp\left\{ \langle \mathsf{t}(\theta), \tau(X) \rangle \right\}.$$

Read Linear Exponential Families in PGM Ch8.2.1 by yourself ☺

Factored Exponential Families:

数据挖掘实验室
Data Mining Lab

- **Exponential factor family**

An (unnormalized) exponential factor family Φ *is defined by* τ, t, A, *and* Θ *(as in the exponential family). A factor in this family is*

$$\phi_{\boldsymbol{\theta}}(\xi) = A(\xi) \exp \{ \langle \mathsf{t}(\boldsymbol{\theta}), \tau(\xi) \rangle \} .$$

■

- **Family composition**

Let $\Phi_1, \ldots, \Phi_k$ *be exponential factor families, where each* Φ_i *is specified by* τ_i, t_i, A_i, *and* Θ_i. *The* composition *of* $\Phi_1, \ldots, \Phi_k$ *is the family* $\Phi_1 \times \Phi_2 \times \cdots \times \Phi_k$ *parameterized by* $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \circ \boldsymbol{\theta}_2 \circ \cdots \circ \boldsymbol{\theta}_k \in \Theta_1 \times \Theta_2 \times \cdots \times \Theta_k$, *defined as*

$$P_{\boldsymbol{\theta}}(\xi) \propto \prod_i \phi_{\boldsymbol{\theta}_i}(\xi) = \left(\prod_i A_i(\xi) \right) \exp \left\{ \sum_i \langle \mathsf{t}_i(\boldsymbol{\theta}_i), \tau_i(\xi) \rangle \right\}$$

where $\phi_{\boldsymbol{\theta}_i}$ *is a factor in the* i*'th factor family.* ■

See examples in PGM Ch8.3.2 ☺

Entropy and Relative Entropy:

数据挖掘实验室
Data Mining Lab

- **Definition**

$$\mathbb{H}_P(X) = \mathbb{E}_P\left[\log \frac{1}{P(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)}$$

- *A measure of the amount of "stochasticity" or "noise" in the distribution;*
- *The number of bits needed, on average, to encode instances in the distribution.*

A low entropy	⇨	*Distribution mass is on a few instances.*
A high entropy	⇨	*A more uniform Distribution.*

Relative Entropy:

- **Definition**

  Consider a distribution Q and a distribution P_θ in an exponential family defined by τ and t. Then

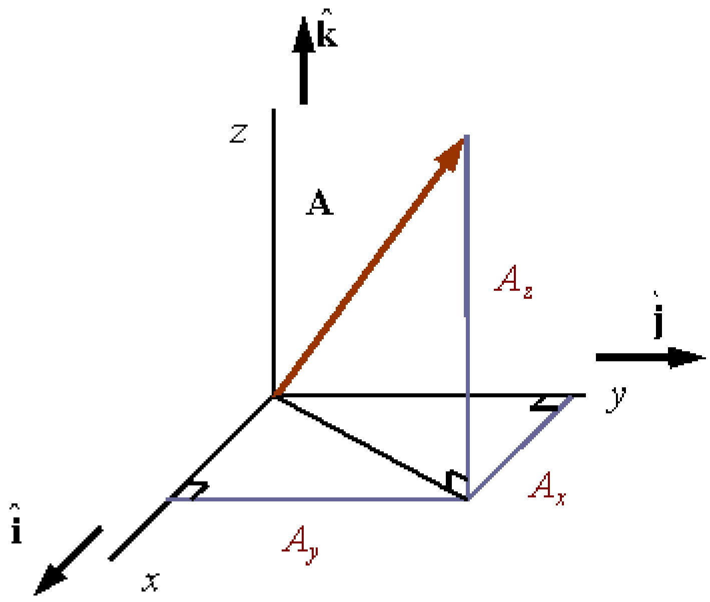
$$\mathbb{D}(Q \| P_{\boldsymbol{\theta}}) = -\boldsymbol{H}_Q(\mathcal{X}) - \langle \boldsymbol{E}_Q[\tau(\mathcal{X})], \mathrm{t}(\boldsymbol{\theta}) \rangle + \ln Z(\boldsymbol{\theta})$$

  - *A measure of* ***distance*** *between two distributions.*
  - *Relative entropy is* ***not symmetric*** *(i.e. $D(P||Q) \neq D(Q||P)$)*
  - *There are more elegant results if the two distributions are from the same distribution family (i.e. exponential family).*

Projections -- motivation :

- **Motivation**

Finding the distribution, within a given exponential family, that is ***closest*** to a given distribution ***in terms of relative entropy.***



An orthogonal projection of a vector in R^3 .

Finding the closest vector on a given subspace.

Projections -- definition:

- **Definition**

Let P be a distribution and let $\mathcal{Q}$ be a convex set of distributions.

- *The* I-projection *(information projection) of P onto $\mathcal{Q}$ is the distribution*

$$Q^I = \arg\min_{Q \in \mathcal{Q}} \boldsymbol{D}(Q \| P).$$

- *The* M-projection *(moment projection) of P onto $\mathcal{Q}$ is the distribution*

$$Q^M = \arg\min_{Q \in \mathcal{Q}} \boldsymbol{D}(P \| Q).$$

Projections -- comparison:

Weight (distribution of P) is known, find optimal Q. (consider mixed Gaussian.)

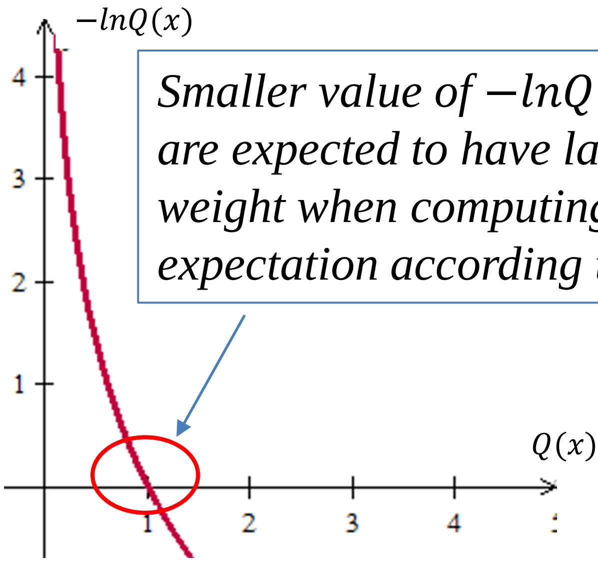
- **M-projection**

Finding Q that minimize $\mathbb{D}(P\|Q) = -\boldsymbol{H}_P(X) + \boldsymbol{E}_P[\underline{-\ln Q(X)}]$

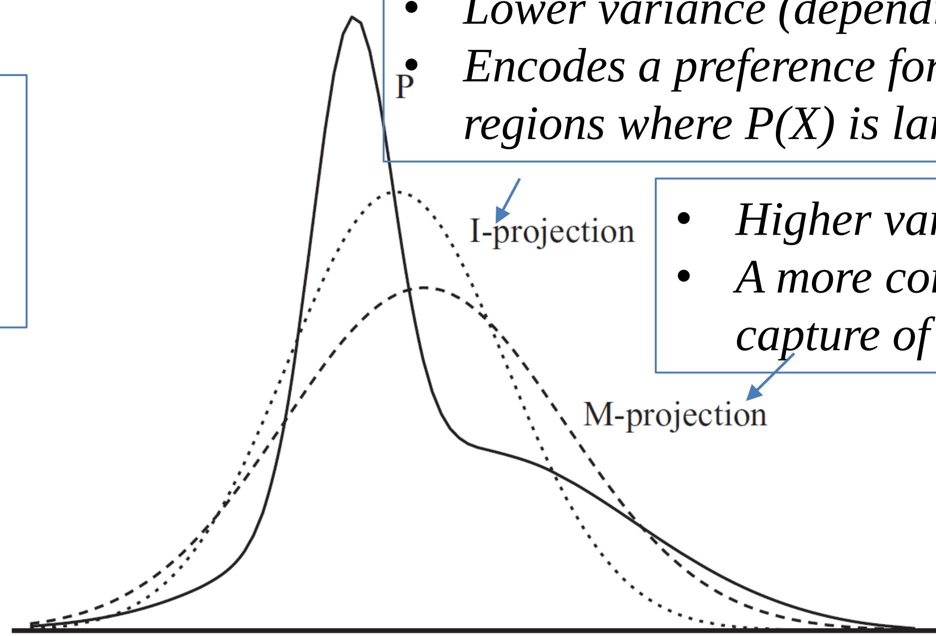
P is known, find optimal weight (distribution Q). (consider mixed Gaussian. Always give the largest weight to the x with largest P(x).)

- **I-projection**

Finding Q that minimize $\mathbb{D}(Q\|P) = \underline{-\boldsymbol{H}_Q(X)} + \boldsymbol{E}_Q[\underline{-\ln P(X)}]$



Smaller value of $-lnQ(x)$ are expected to have larger weight when computing the expectation according to P.



- *Lower variance (depending on H);*
- *Encodes a preference for to regions where P(X) is large.*

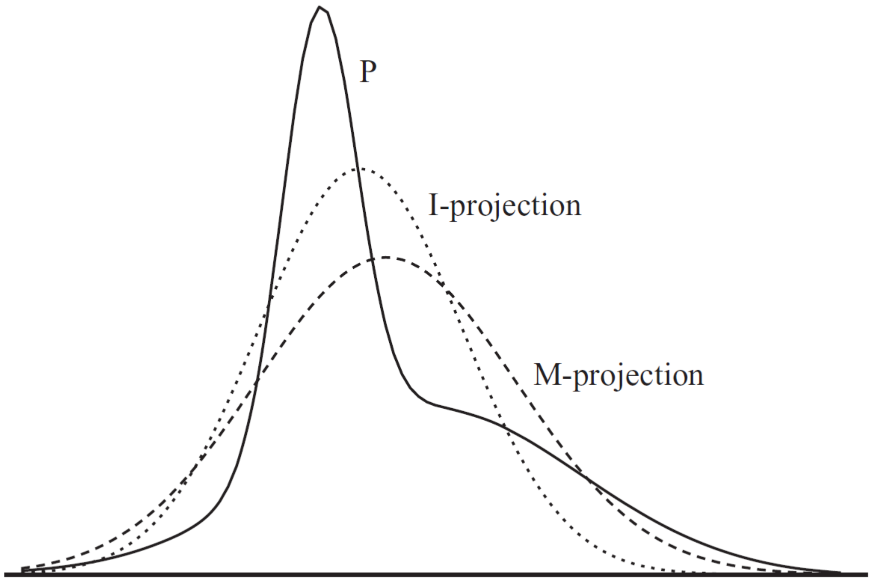
- *Higher variance;*
- *A more comprehensive capture of P .*

Projections -- comparison:

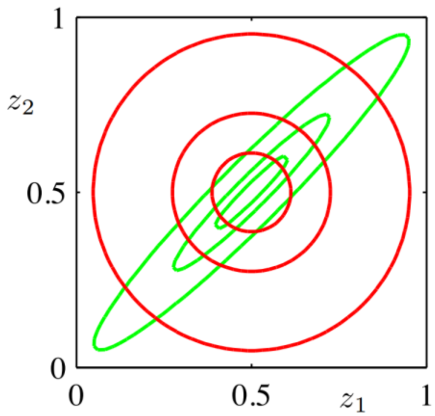
- **M-projection :** Although the M-projection attempts to match the main mass of *P*, its ***high variance*** is a compromise to ensure that it assigns reasonably high density to ***all regions*** that are in the support of *P*.

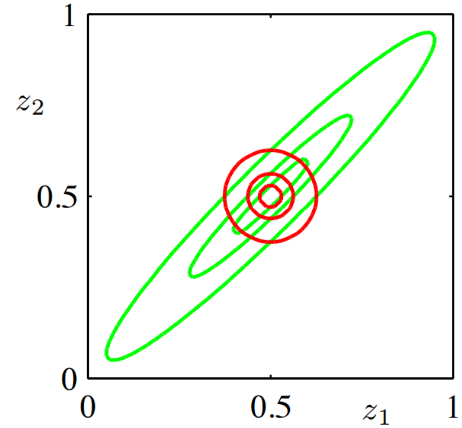
- **I-projection:** The first term brings ***a penalty on small variance***. The second term, i.e. Q[−ln P(X)], encodes ***a preference for assigning higher density to regions where P(X) is large*** and ***very low density to regions where P(X) is small.***

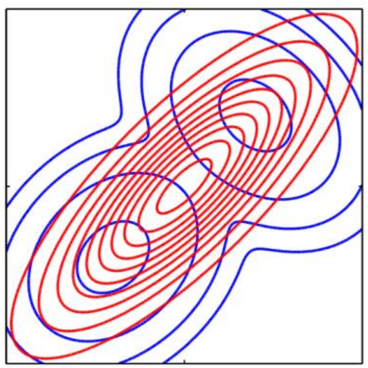
The M-projection attempts to ***give all assignments reasonably high probability***, whereas the I-projection attempts to ***focus on high-probability assignments*** in P while maintaining a reasonable entropy**.**



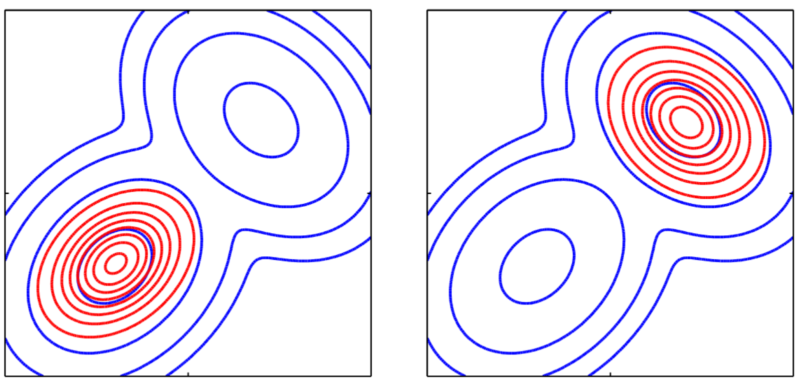
Projections -- comparison:







p=Blue, q=Red

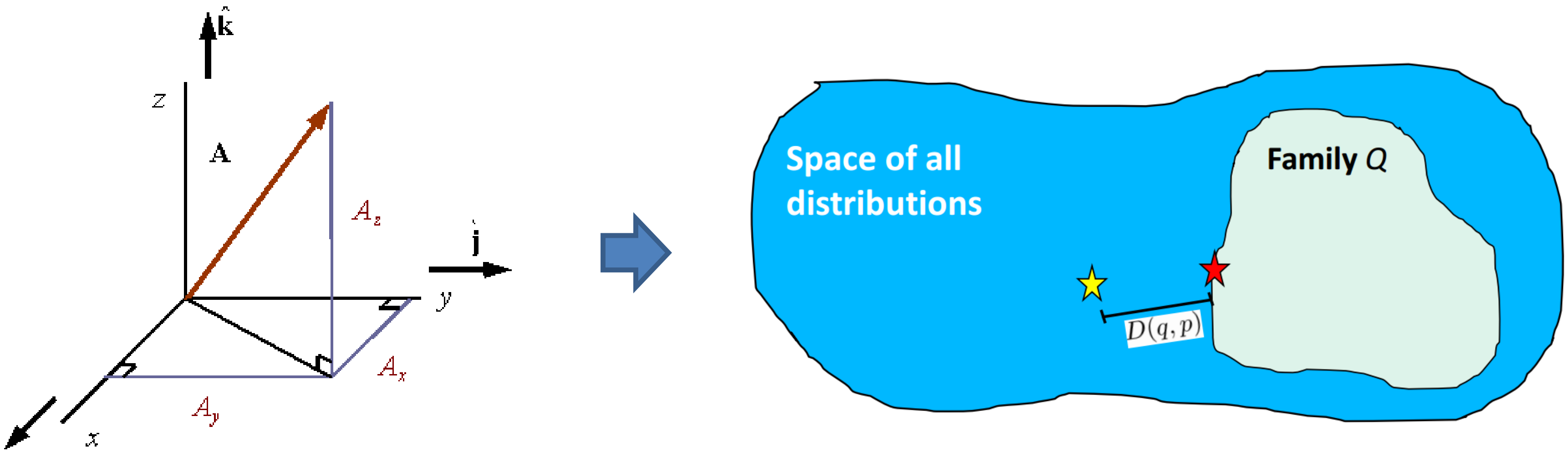


p=Blue, q=Red (two equivalently good solutions!)

M-projection
(maintains the mean)

I-projection
(fails to maintains the mean)

More About M—Projection (*moment matching*):



More About M -- Projection (*moment matching*):

Theorem 8.6

Let P be a distribution over $\mathcal{X}$, and let $\mathcal{Q}$ be an exponential family defined by the functions $\tau(\xi)$ and $\mathrm{t}(\boldsymbol{\theta})$. If there is a set of parameters $\boldsymbol{\theta}$ such that $\boldsymbol{E}_{Q_{\boldsymbol{\theta}}}[\tau(\mathcal{X})] = \boldsymbol{E}_P[\tau(\mathcal{X})]$, then the M-projection of P is $Q_{\boldsymbol{\theta}}$.

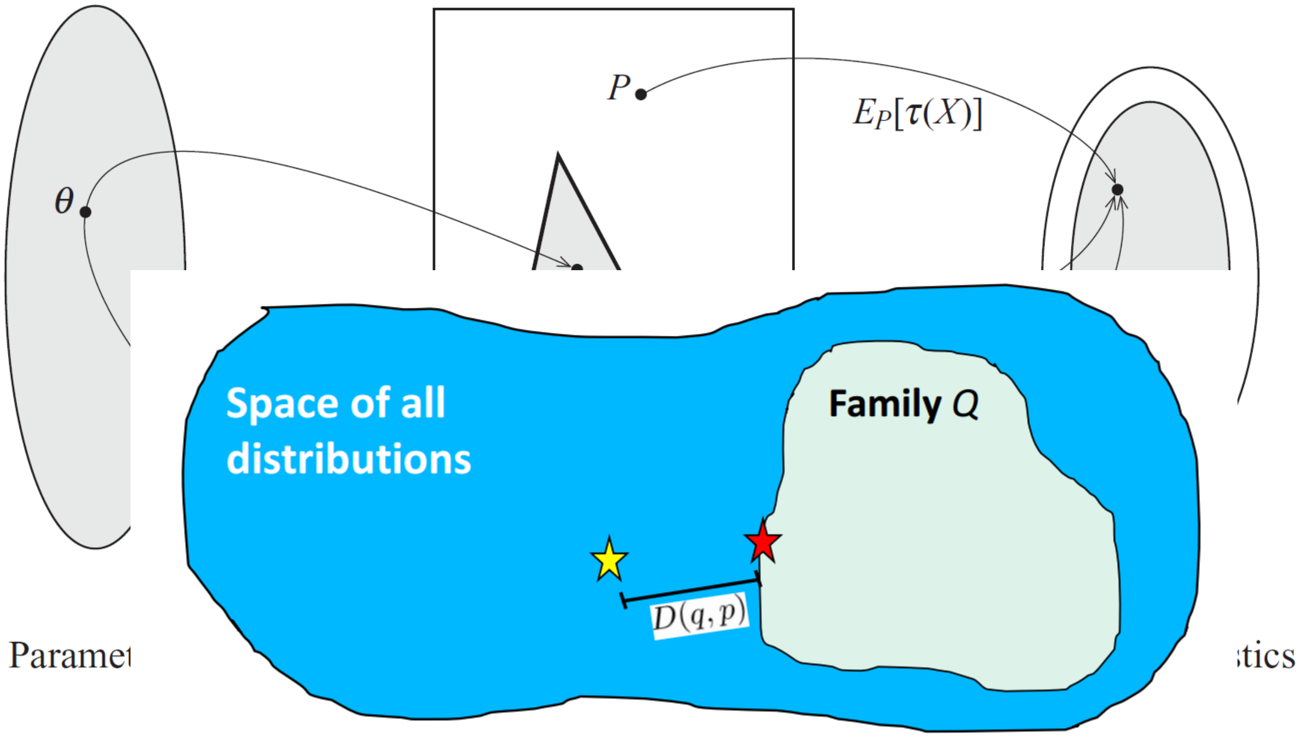
Proof Suppose that $\boldsymbol{E}_P[\tau(\mathcal{X})] = \boldsymbol{E}_{Q_{\boldsymbol{\theta}}}[\tau(\mathcal{X})]$, and let $\boldsymbol{\theta}'$ be some set of parameters. Then,

$$\boldsymbol{D}(P\|Q_{\boldsymbol{\theta}'}) - \boldsymbol{D}(P\|Q_{\boldsymbol{\theta}})$$

$$= \quad \boldsymbol{D}(Q_{\boldsymbol{\theta}}\|Q_{\boldsymbol{\theta}'}) \geq 0.$$

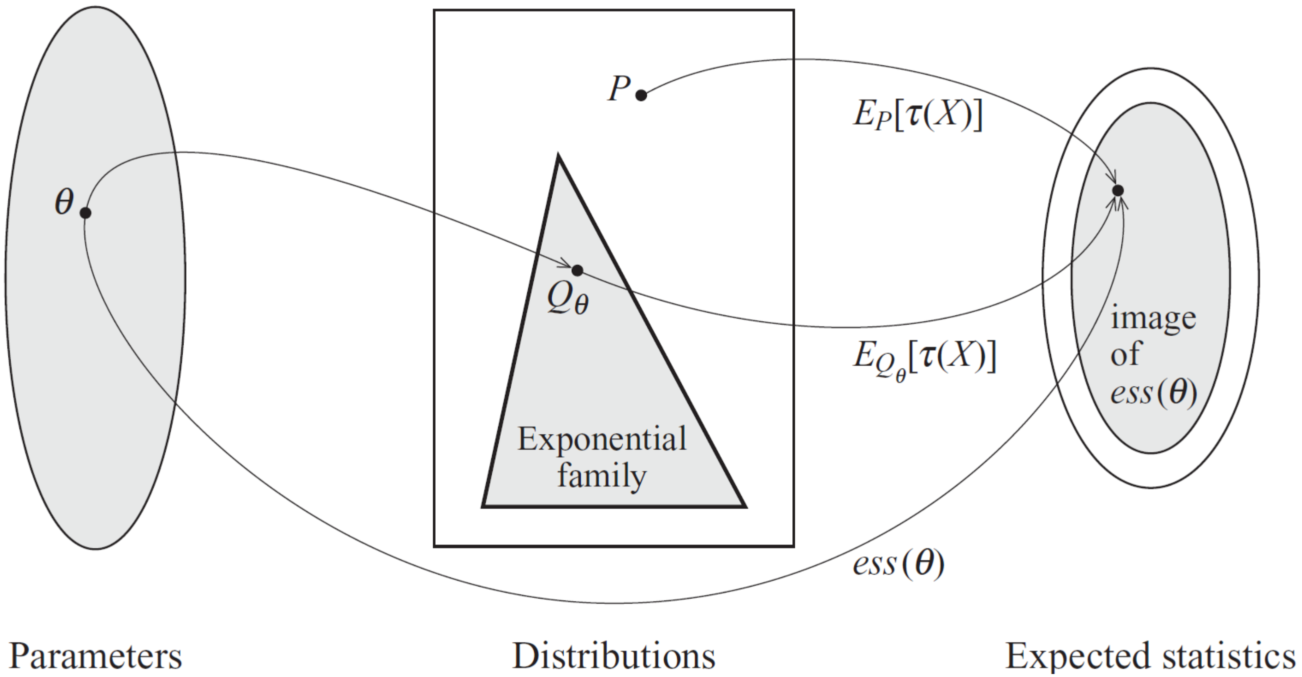
We conclude that the M-projection of P is $Q_{\boldsymbol{\theta}}$. ■

More About M -- Projection (*moment matching*):



- Each ***parameter*** corresponds to a ***distribution***, which in turn corresponds to ***a value of the expected statistics***.
- The ***function ess*** maps parameters directly to ***expected statistics***.
- If the expected statistics of ***P and Q$_\theta$ match***, then Q_θ is the ***M-projection*** of P.

More About M -- Projection (*moment matching*):



Let $\boldsymbol{s}$ *be a vector. If* $\boldsymbol{s} \in image(ess)$ *and ess is invertible, then* $M - project(\boldsymbol{s}) = ess^{-1}(\boldsymbol{s})$.

More About M -- Projection (*moment matching*):

数据挖掘实验室
Data Mining Lab

A gentle example:

What is the best Gaussian approximation (in the M-projection sense) to a non-Gaussian distribution over *X*?

Consider the exponential family of ***Gaussian*** *distributions. Recall that the sufficient statistics function for this family is* $\tau(x) = \langle x, x^2 \rangle$. *Given parameters* $\boldsymbol{\theta} = \langle \mu, \theta^2 \rangle$, *the expected value of* τ *is:*

$$ess(\langle \mu, \sigma^2 \rangle) = E_{Q_{\langle \mu, \sigma^2 \rangle}}[\tau(X)] = \langle \mu, \sigma^2 + \mu^2 \rangle.$$

- *For any distribution* P, $E_P[\tau(X)]$ *must be in the image of this function (see exercise 8.4).*
  → *for any choice of P, we can apply theorem 8.6.*

- *By* ***inverting*** *the ess function*

$$\text{M-project}(\langle s_1, s_2 \rangle) = ess^{-1}(\langle s_1, s_2 \rangle) = \langle s_1, s_2 - s_1^2 \rangle.$$

- *By substituting* s_1 *and* s_2 *with* $E_P[X]$ *and* $E_P[X^2]$, *Thus, the estimated parameters are the mean and variance of X according to P, as we would expect.*

Thanks

Q & A?