A Bayes Rule for Density Matrices

Manfred K. Warmuth

UCSC

Poster NIPS 05
Revised January 11, 2006

Thanks to Dima Kuzmin

Other Titles?

A “Quantum” Bayes Rule

or

A Non-commutative Bayes Rule

Matrix Priors

Prior probability of model M_i is $P(M_i)$
Denote prior as diagonal matrix:

$$\boldsymbol{I} \quad \begin{pmatrix} \ddots & & 0 \\ & P(M_i) & \\ 0 & & \ddots \end{pmatrix} \qquad \boldsymbol{I}^T \quad = \quad \sum_i P(M_i)\, \boldsymbol{e}_i \boldsymbol{e}_i^T$$

$$\boldsymbol{D} \quad \begin{pmatrix} \ddots & & 0 \\ & \delta_i & \\ 0 & & \ddots \end{pmatrix} \qquad \boldsymbol{D}^T \quad = \quad \sum_i \delta_i \quad \underbrace{\boldsymbol{d}_i \boldsymbol{d}_i^T}_{\text{dyad}},$$

In above generalization of the prior,
the eigenvalues (δ_i) are probability vector
and $\boldsymbol{D}$ is orthogonal system of eigenvectors

Prior now a Density Matrix:

- Symmetric positive definite matrix of trace one

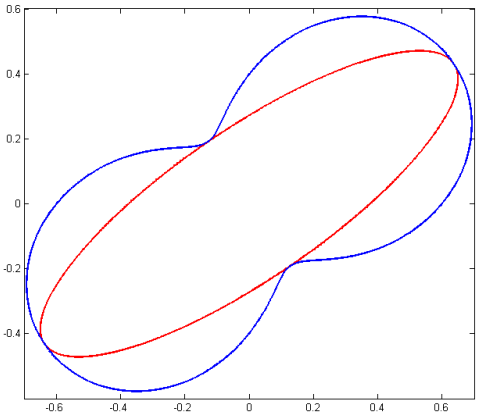
Covariance Matrices

- View the symmetric positive definite matrix $\mathcal{S}$ as a covariance matrix of some random cost vector $\boldsymbol{c} \in \mathbb{R}^n$, i.e.

$$\mathcal{S} = \mathbb{E}\left((\boldsymbol{c} - \mathbb{E}(\boldsymbol{c})(\boldsymbol{c} - \mathbb{E}(\boldsymbol{c}))^\top\right)$$

- The variance along any vector $\boldsymbol{u}$ has the form

$$\mathbb{V}(\boldsymbol{c}^\top \boldsymbol{u}) = \mathbb{E}(\left(\boldsymbol{c}^\top \boldsymbol{u} - \mathbb{E}(\boldsymbol{c}^\top \boldsymbol{u})\right)^2) = \mathbb{E}(\left((\boldsymbol{c}^\top - \mathbb{E}(\boldsymbol{c}^\top))\, \boldsymbol{u}\right)^2) = \boldsymbol{u}^\top \mathcal{S} \boldsymbol{u}$$



An ellipse $\mathcal{S}$ in $\mathbb{R}^2$: The eigenvectors are the directions of the axes and the eigenvalues their lengths. Ellipses are weighted combinations of the one-dimensional degenerate ellipses (dyads) corresponding to the axes.

The curve of the ellipse is a plot of direction $\boldsymbol{u}$ times $||\mathcal{S}\boldsymbol{u}||_2$ and the outer figure eight is direction $\boldsymbol{u}$ times the variance $\boldsymbol{u}^\top \mathcal{S} \boldsymbol{u}$. At the eigenvectors, this variance equals the eigenvalues and touches the ellipse.

Two More Interpretations of $u^T \mathcal{S} u$

- $\boldsymbol{u}^\top \mathcal{S} \boldsymbol{u}$ is the squared length of $\boldsymbol{u}$ w.r.t. the basis $\sqrt{\mathcal{S}}$:

$$\boldsymbol{u}^\top \mathcal{S} \boldsymbol{u} = \boldsymbol{u}^T \sqrt{\mathcal{S}} \sqrt{\mathcal{S}} \boldsymbol{u} = ||\sqrt{\mathcal{S}} \boldsymbol{u}||_2^2$$

- Quantum physics interpretation:

  Since the squared length of $\boldsymbol{u}$ w.r.t. any orthogonal basis $\boldsymbol{S}$ is one, any such basis turns the unit vector into an n-dimensional probability vector $((\boldsymbol{u}^\top \boldsymbol{s}_i)^2)$
  Now $\boldsymbol{u}^\top \mathcal{S} \boldsymbol{u}$ is the expected eigenvalue w.r.t. this probability vector:

$$\begin{aligned} \boldsymbol{u}^\top \mathcal{S} \boldsymbol{u} &= \boldsymbol{u}^T \left(\sum_i \sigma_i \boldsymbol{s}_i \boldsymbol{s}_i^T \right) \boldsymbol{u} \\ &= \sum_i (\boldsymbol{u}^\top \boldsymbol{s}_i)^2 \, \sigma_i \end{aligned}$$

Classical Setup

- Model M_i is chosen with prior probability $P(M_i)$
- Datum y is generated with probability $P(y|M_i)$

$$P(y) = \sum_i \underbrace{P(M_i)P(y|M_i)}_{\text{expected likelihood}}$$

Quantum Setup

- If prior is $\mathfrak{D} = \sum_{i=1}^{n} \delta_i \; \boldsymbol{d}_i \boldsymbol{d}_i^T$, then dyad $\boldsymbol{d}_i \boldsymbol{d}_i^T$ is chosen with probability δ_i
- A random variable $\boldsymbol{c}^T \boldsymbol{d}_i$ is observed, where $\boldsymbol{c}$ has covariance matrix $\mathfrak{D}(y|.)$
- Variance along direction $\boldsymbol{d}_i$

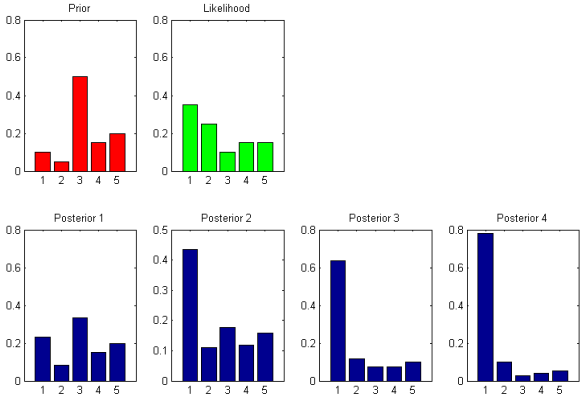
$$\mathbb{V}(\boldsymbol{c}^T \boldsymbol{d}_i) = \boldsymbol{d}_i^T \mathfrak{D}(y|.) \boldsymbol{d}_i$$

- $$\underbrace{\mathrm{tr}(\mathfrak{D}(.)\mathfrak{D}(y|.))}_{\substack{\text{quantum} \\ \text{measurement}}} = \mathrm{tr}(\sum_i \delta_i \; \boldsymbol{d}_i \boldsymbol{d}_i{}^{\top} \mathfrak{D}(y|.)) = \underbrace{\sum_i \delta_i \; \boldsymbol{d}_i{}^{\top} \mathfrak{D}(y|.) \boldsymbol{d}_i}_{\text{expected variance}}$$

- When $\mathfrak{D}(.)$ and $\mathfrak{D}(y|.)$ diagonal, then in classical case

Classical Bayes Rule

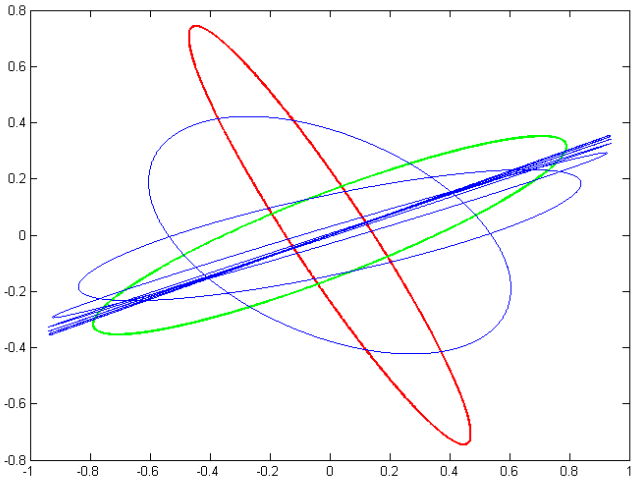
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



- 4 updates with the same data likelihood
- Update maintains uncertainty information about maximum likelihood
- Soft max

Bayes Rule for Density Matrices

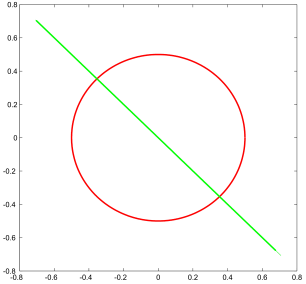
$$\mathcal{D}(.|y) = \frac{\exp\left(\log \mathcal{D}(.) + \log \mathcal{D}(y|.)\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 7 updates with same data density matrix
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

Using of Off-diagonal Elements

$$\mathcal{D}(.) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \quad \mathcal{D}(y|.) = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \boldsymbol{U}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\boldsymbol{U}^\top,$$



where $\boldsymbol{U} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$

- $\forall$ diagonal $\mathcal{D}(.)$, $\mathrm{tr}(\mathcal{D}(.)\,\mathcal{D}(y|.)) = \frac{1}{2}$, largest eigenvalue is not "visible" in basis $\boldsymbol{I}$
- But

$$\mathrm{tr}\left(\underbrace{\boldsymbol{U}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\boldsymbol{U}^\top}_{\mathcal{D}(.|y) \text{ of new rule}} \quad \mathcal{D}(y|.)\right) = 1$$

Non-commutative Bayes Rule?

- The product of two symmetric positive matrices can be neither symmetric nor positive definite

$$\mathfrak{D}(.|y) = \frac{\overbrace{\exp\Big(\log \overbrace{\mathfrak{D}(.)}^{\text{sym.pos.def}} + \log \overbrace{\mathfrak{D}(y|.)}^{\text{sym.pos.def}}\Big)}^{\text{sym.pos.def}}}{\operatorname{tr}(\cdots)}$$

Classical Rule as Special Case

- If $\mathfrak{D}(.)$ and $\mathfrak{D}(y|.)$ have the same eigensystem, then quantum Bayes rule specializes the the classical case

$$\begin{pmatrix} \ddots & & 0 \\ & P(M_i|y) & \\ 0 & & \ddots \end{pmatrix} =$$

$$\frac{\begin{pmatrix} \ddots & & 0 \\ & P(M_i) & \\ 0 & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & 0 \\ & P(y|M_i) & \\ 0 & & \ddots \end{pmatrix}}{\mathrm{tr}(\ \cdots\ )}$$

Classical Intersection Properties

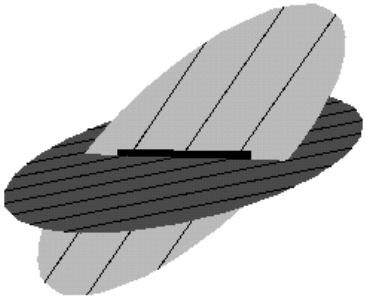
$P(M_i)$	$P(M_i\|y)$	$P(y\|M_i)$
0	0	0
a	0	0
0	b	0
a	b	ab

- Computes intersection of two sets

Generalized Case

$$\mathcal{S} \odot \mathcal{T} := \exp(\log \mathcal{S} + \log \mathcal{T}) = \lim_{n \to \infty} (\mathcal{S}^{1/n} \mathcal{T}^{1/n})^n$$

- Lie-Trotter Formula
- Limit alwavs exists and well behaved



- “Product” lies in intersection of both spans
- In example, product is degenerate ellipse of dimension one (bold line)
- New rule:

$$\mathcal{D}(.|y) = \frac{\mathcal{D}(.) \odot \mathcal{D}(y|.)}{\mathrm{tr}(\mathcal{D}(.) \odot \mathcal{D}(y|.))}$$

Log Loss

- For classical case

$$-\log P(y) = -\log \sum_i P(M_i) P(y|M_i)$$

- In matrix case similar inequality

$$-\log \mathrm{tr}(\mathfrak{D}(.)\mathfrak{D}(y|.)) \overset{\text{G.Th.}}{\leq} -\log \underbrace{\mathrm{tr}\left(\mathbf{exp}\left(\mathbf{log}\,\mathfrak{D}(.) + \mathbf{log}\,\mathfrak{D}(y|.)\right)\right)}_{\text{normaliz.factor}}$$

$$= -\log \mathrm{tr}(\mathfrak{D}(.) \odot \mathfrak{D}(y|.))$$

- Golden-Thompson inequality:

  $\forall$ symmetric $\mathcal{S}$ and $\mathcal{T}$: $\mathrm{tr}(\mathbf{exp}(\mathcal{S}+\mathcal{T})) \leq \mathrm{tr}(\mathbf{exp}\,\mathcal{S}\,\mathbf{exp}\,\mathcal{T})$

- Telescoping properties of log loss still hold as inequalities
- Worst case loss bound generalize

Derivation of Updates [KW97]

$$\inf_{\boldsymbol{w}} \quad \underbrace{\Delta(\boldsymbol{w}, \boldsymbol{w}_0)}_{\text{divergence to } \boldsymbol{w}_0} \quad + \quad \underbrace{\eta}_{\text{learning rate}} \quad \text{Loss}(\boldsymbol{w})$$

- Can derive large variety of updates by varying divergence, loss function and learning rate
- Examples: Gradient descent update, exponentiated gradient update, Ada-Boost ($\eta \to \infty$)
- Here we will derive Bayes rule with this framework

Classical Bayes Rule

- Mixture parameter γ_i
- Prior $P(M_i)$

$$\inf_{\gamma_i \geq 0,\, \sum_i \gamma_i = 1} \underbrace{\sum_i \gamma_i \log \frac{\gamma_i}{P(M_i)}}_{\text{rel.entropy}} - \eta \underbrace{\sum_i \gamma_i \log P(y|M_i)}_{\text{expected loss}}$$

- $\eta = 1$: Bayes Rule
  - Soft max
- $\eta = \infty$: maximum likelihood
- Special case of Exponentiated Gradient update

Minimization of γ

Lagrangian:

$$L(\boldsymbol{\gamma}) = \sum_i \gamma_i \log \frac{\gamma_i}{P(M_i)} - \eta \sum_i \gamma_i \log P(y|M_i) + \lambda((\sum_i \gamma_i) - 1)$$

$$\frac{\partial L(\boldsymbol{\gamma})}{\partial \gamma_i} = \log \frac{\gamma_i}{P(M_i)} + 1 - \eta \log P(y|M_i) + \lambda$$

Setting partials zero:

$$\gamma_i^* = P(M_i) \exp(\lambda - 1 + \eta \log P(y|M_i))$$

Enforcing sum constraint:

$$\gamma_i^* = \frac{P(M_i) P(y|M_i)^\eta}{\sum_j P(M_j) P(y|M_j)^\eta}$$

$\eta = 1$: Classical Bayes rule

Bayes Rule for Density Matrices

- Parameter is density matrix $\mathcal{G}$
- Prior is density matrix $\mathcal{D}$

$$\inf_{\mathcal{G}\ \text{dens.matr.}} \underbrace{\mathrm{tr}(\mathcal{G}(\log \mathcal{G} - \log \mathcal{D}(.))}_{\text{Quantum rel. entr.}} - \eta \underbrace{\mathrm{tr}(\mathcal{G} \log \mathcal{D}(y|.))}_{\text{Fancier mixture loss}}$$

- $\eta = 1$: Quantum Bayes Rule
  - Soft maximum eigenvalue calculation
- $\eta = \infty$: minimized when $\mathcal{G}$ is dyad $\boldsymbol{u}\boldsymbol{u}^T$ and $\boldsymbol{u}$ is the eigenvector belonging to a minimum eigenvalue of

$$-\log \mathcal{D}(y|.)$$

- Special case of Matrix Exponentiated Gradient update

Future

- Find conditional probability calculus for density matrices
- On-line update for PCA
- Applications
- Bayes rule for arbitrary matrices