A Bayesian Probability Calculus for Density Matrices

Manfred K. Warmuth and Dima Kuzmin

University of California - Santa Cruz

Web: Google.com "manfred"

22nd Conference on Uncertainty in Artificial Intelligence
Last updated August 4, 2006
Outline

1. Matrices

2. Conventional and Generalized Probability Distributions

3. Conventional and generalized Bayes rule

4. Bounds, derivation, calculus
Density Matrices?

- Symmetric: $A^T = A$
- Positive definite: $u^T A u \geq 0 \quad \forall u$
- Trace one: sum of diagonal elements is one

Here: generalizations of finite probability distributions
Ellipses

- We illustrate symmetric matrices as ellipses - affine transformations of the unit ball:

\[ \text{Ellipse} = \{ Au : \|u\|_2 = 1 \} \]

- Dotted lines connect \( u \) on unit ball and \( Au \)
Ellipses

- We illustrate symmetric matrices as ellipses - affine transformations of the unit ball:

Ellipse = \{Au : \|u\|_2 = 1\}

- Dotted lines connect \(u\) on unit ball and \(Au\)

- For symmetric matrices, the eigenvectors form the axes of the ellipse and eigenvalues their lengths
Matrices

Eigendecomposition

- $Ax = \alpha x$, $x$ is an eigenvector, $\alpha$ is an eigenvalue
- Symmetric matrices always have an eigendecomposition:

$$A = \begin{pmatrix} \mathbf{A} & \mathbf{\alpha} \end{pmatrix}$$

orthogonal mat. diagonal mat. of eigenvectors

of eigenvectors real eigenvalues

$$A^\top = A \left( \sum_i \alpha_i \mathbf{e}_i \mathbf{e}_i^\top \right) A^\top$$

$$= \sum \alpha_i (\mathbf{Ae}_i)(\mathbf{e}_i^\top \mathbf{A}^\top) = \sum_i \alpha_i \mathbf{a}_i \mathbf{a}_i^\top$$

eigenvalues dyads

- Density matrices again
  - Positive definite: $\alpha_i \geq 0$
  - Trace one: $\sum_i \alpha_i = 1$
  - $n$ eigenvalues form a probability vector
Density Matrices as mixtures of dyads

- Dyads are degenerate ellipses:
- Many mixtures lead to same density matrix

$$0.2 + 0.3 + 0.5 = 0.29 + 0.71$$

- Decomposition into $n$ dyads that correspond to eigenvectors
View the symmetric positive definite matrix \( A \) as a covariance matrix of some random cost vector \( c \in \mathbb{R}^n \), i.e.

\[
A = E \left( (c - E(c)(c - E(c)))^\top \right)
\]

The variance along any vector \( u \) is

\[
\nabla (c^\top u) = E \left( \left( c^\top u - E(c^\top u) \right)^2 \right)
= E \left( \left( (c^\top - E(c^\top)) u \right)^2 \right)
= u^\top Au
\]

\[
u^\top Au = \text{tr}(u^\top Au) = \text{tr}(Au u^\top)
\]
Plotting Variance

Curve of the ellipse is plot of vector $Au$, where $u$ is unit vector. The outer figure eight is direction $u$ times the variance $u^\top Au$. For an eigenvector, this variance equals the eigenvalue and touches the ellipse.
Conventional Probability Theory

- **Space** is set $A$ of $n$ **elementary events** / points
  \[ \{a_1, a_2, a_3, a_4, a_5\} \]

- **Event** is subset
  \[ (0, 1, 1, 0, 1) \]

- **Distribution** is probability vector
  \[ (.1, .2, .3, .1, .3) \]

- Probability of event $S$: $P(S) = \sum_i I(a_i \in S)P(a_i)$

- **Random variable**: $f: A \rightarrow \mathbb{R}$
  
  **Expectation**: $E(f) = \sum_i f(a_i)P(a_i)$
Generalized Probabilites over $\mathbb{R}^n$

- **Elementary event** is dyad $uu^\top$ where $u$ unit vector
  - One-dimensional projection matrix onto $u$
  - Degenerate ellipse:

- **Event** is symmetric matrix $P$ with $\{0, 1\}$ eigenvalues

$$P = U \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} U^\top,$$
  where $U$ orthogonal

- Projection matrix onto arbitrary subspace of $\mathbb{R}^n$: $P^2 = P$

- **Distribution** is density matrix $A$:
  symmetric positive definite matrix of trace one

$$A = U \begin{pmatrix} .1 & 0 & 0 & 0 & 0 \\ 0 & .2 & 0 & 0 & 0 \\ 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & 0 & .1 & 0 \\ 0 & 0 & 0 & 0 & .3 \end{pmatrix} U^\top,$$
  Eigenvalues form probability vector
Density Matrices Continued

- Density matrix $A$ assigns generalized probability $\text{tr}(A uu^\top)$ to dyad $uu^\top$.

Sum of probabilities over an orthonormal basis $u_i$ is 1:

$$\sum_i \text{tr}(A u_i u_i^\top) = \text{tr}(A \sum_i u_i u_i^\top) = \text{tr}(A) = 1$$

- Uniform density matrix: $\frac{1}{n}I$

$$\text{tr}\left(\frac{1}{n}I uu^T\right) = \frac{1}{n} \text{tr}(uu^T) = \frac{1}{n}$$

- All dyads have generalized probability $\frac{1}{n}$.
  Probability of $n$ orthogonal dyads sum to 1.
Conventional and Generalized Probability Distributions

Probability of Events

- Probability of event $P$ is

$$\text{tr}(AP) = \text{tr}\left(\sum \alpha_i a_i a_i^\top P\right) = \sum \alpha_i a_i^\top P a_i$$

- Random variable is symmetric matrix $S$ of expectation

$$\text{tr}(AS) = \sum \sigma_i s_i^\top A s_i$$

- Trace is quantum measurement
  for mixture state $A$ and instrument $S$
Gleason’s Theorem

**Definition**

Scalar function $\mu(u)$ from unit vectors $u$ in $\mathbb{R}^n$ to $\mathbb{R}$ is called *generalized probability measure* if:

- $\forall u$, $0 \leq \mu(u) \leq 1$
- If $u_1, \ldots, u_n$ form an orthonormal basis for $\mathbb{R}^n$, then $\sum \mu(u_i) = 1$

**Theorem**

Let $n \geq 3$. Then any generalized probability measure $\mu$ on $\mathbb{R}^n$ has the form:

$$\mu(u) = \text{tr}(A uu^\top)$$

for a uniquely defined density matrix $A$.
Outline

1. Matrices
2. Conventional and Generalized Probability Distributions
3. Conventional and generalized Bayes rule
4. Bounds, derivation, calculus
Conventional Setup

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

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

expected likelihood
Conventional and generalized Bayes rule

Conventional 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
Conventional and generalized Bayes rule

Bayes Rule for Density Matrices

\[ D(M|y) = \frac{\exp \left( \log D(M) + \log D(y|M) \right)}{\text{tr} \left( \text{above matrix} \right)} \]

- 20 updates with same data likelihood matrix \( D(y|M) \)
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation
The product of two symmetric positive matrices can be neither symmetric nor positive definite.

\[
D(M | y) = \exp \left( \log \left( \frac{D(M)}{\text{sym.pos.def}} \right) + \log \left( \frac{D(y | M)}{\text{sym.pos.def}} \right) \right) \frac{\text{tr}(\cdots)}{\text{sym.}}
\]
Conventional and generalized Bayes rule

Conventional Rule Special Case

- If $D(M)$ and $D(y|M)$ have the same eigensystem, then generalized Bayes rule specializes the conventional case

$$
\begin{pmatrix}
\vdots & 0 \\
P(M_i | y) & \\
0 & \ddots
\end{pmatrix} = 
\begin{pmatrix}
\vdots & 0 \\
P(M_i) & \\
0 & \ddots
\end{pmatrix}
\begin{pmatrix}
\vdots & P(y|M_i) & 0 \\
0 & \ddots & \\
\text{tr}(\cdots) & & 
\end{pmatrix}
$$
Diagonal matrices “don’t see” Hadamard matrices

\[
H = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\]

\[
HH^\top = \begin{pmatrix}
n & 0 & 0 & 0 \\
0 & n & 0 & 0 \\
0 & 0 & n & 0 \\
0 & 0 & 0 & n
\end{pmatrix}
\]

- \( h \) is any of the \( n \) columns
- \( u = \frac{1}{\sqrt{n}} h \) is unit vector, \( uu^\top \) is \( n \times n \) matrix with \( \frac{1}{n} \) in diagonal
- For all diagonal density matrices \( A \)

\[
\text{tr}(A uu^\top) = \sum_i A_{i,i} \frac{1}{n} = \frac{1}{n}
\]

- Density matrix \( uu^\top \) gives likelihood one to \( uu^\top \)

\[
\text{tr}(uu^\top uu^\top) = ||u||_2^2 = 1
\]
Visualization of Hadamard example

\[ H = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \mathbf{u} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \]

\[ \mathbf{u}\mathbf{u}^\top = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} = \mathbf{U} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{U}^\top, \]

where \( \mathbf{U} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \).

- Any diagonal matrix
  \[
  \text{tr} \left( \begin{pmatrix} \alpha & 0 \\ 0 & 1-\alpha \end{pmatrix} \mathbf{u}\mathbf{u}^\top \right) = \frac{1}{2},
  \]

  largest eigenvalue is not “visible” in basis \( \mathbf{I} \)

- If prior \( \mathbf{D}(\mathbf{M}) \) uniform and data likelihood matrix \( \mathbf{D}(\mathbf{y}|\mathbf{M}) \) is off-diagonal matrix \( \mathbf{u}\mathbf{u}^\top \), then posterior \( \mathbf{D}(\mathbf{M}|\mathbf{y}) \) equals \( \mathbf{u}\mathbf{u}^\top \) and

  \[
  \text{tr} \left( \mathbf{D}(\mathbf{M}|\mathbf{y})\mathbf{D}(\mathbf{y}|\mathbf{M}) \right) = 1
  \]
Intersection Properties

Conventional Bayes:

\[
\begin{array}{ccc}
P(M_i) & P(M_i|y) & P(y|M_i) \\
0 & 0 & 0 \\
a & 0 & 0 \\
0 & b & 0 \\
a & b & \frac{ab}{P(y)} \\
\end{array}
\]

- Computes intersection of two sets
Avoiding Logs of Zeros

Replace

\[ \exp(\log S + \log T) \] by

\[ S \circ T := \lim_{n \to \infty} (S^{1/n}T^{1/n})^n \]

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

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

\[ D(M|y) = \frac{D(M) \circ D(y|M)}{\text{tr}(D(M) \circ D(y|M))} \]
Plain matrix product is non-commutative and can violate symmetry and positive definiteness. \( \odot \) does not have these drawbacks.
Behaviour of the Limit for $\circ$

- “Ears” indicating negative definiteness are smaller for $(S^{1/2}T^{1/2})^2$ compared to $ST$
- Non-commuting part shrinks as well

\[ S \circ T = T \circ S \]
Properties

1. Commutative, associative, identity matrix as neutral elmt, preserves symmetry and positive definiteness
2. \( S \odot T = ST \) iff \( S \) and \( T \) commute
3. \( \text{range}(S \odot T) = \text{range}(S) \cap \text{range}(T) \)
4. \( \text{tr}(S \odot T) \leq \text{tr}(ST) \) with equality when \( S \) and \( T \) commute
5. For any unit direction \( u \in \text{range}(S) \),
   \[ uu^\top \odot S = e^{u^\top (\log + S) u} uu^\top \]
6. \( \det(S \odot T) = \det(S) \det(T) \), as for the regular matrix product
7. Typically \( S \odot (T + U) \neq S \odot T + S \odot U \)
Setups

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

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

(expected likelihood)

Generalized:

$$D(y) = \text{tr}(D(M) \odot D(y|M)) \leq \text{tr}(D(M)D(y|M))$$

(variance)

$$= \sum \delta_i d_i^\top D(y|M)d_i$$

(expected variance)

Only decouples when $D(M)$ and $D(y|M)$ have same eigensystem
Outline

1. Matrices

2. Conventional and Generalized Probability Distributions

3. Conventional and generalized Bayes rule

4. Bounds, derivation, calculus
Bounds into MAP

- Conventional:
  \[- \log P(y) = - \log \sum_i P(y|M_i)P(M_i) \leq \min_i (\log P(y|M_i) - \log P(M_i)) \]

- Generalized:
  \[- \log m^T S m \leq -m^T \log(S) m, \text{ for any unit vector } m \text{ and symmetric positive definite matrix } S \]
Derivation of Updates

\[ \inf_w \Delta(w, w_0) + \eta \text{ Loss}(w) \]

- 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
Conventional Bayes Rule

- Mixture parameter $\gamma_i$
- Prior $P(M_i)$

$$\inf_{\gamma_i \geq 0, \sum_i \gamma_i = 1} \sum_i \gamma_i \log \frac{\gamma_i}{P(M_i)} - \eta \sum_i \gamma_i \log P(y|M_i)$$

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

Lagrangian:

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

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

Setting partials zero:

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

Enforcing sum constraint:

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

$\eta = 1$: Conventional Bayes rule
Conventional Bayes Again

\[ \inf_{\gamma_i \geq 0, \sum \gamma_i = 1} \sum_i \gamma_i \log \gamma_i - \sum_i \gamma_i \log P(M_i) - \eta \sum_i \gamma_i \log P(y|M_i) \]

- Prior and data treated the same when \( \eta = 1 \)
- Commutativity
Bayes Rule for Density Matrices

- Parameter is density matrix $\mathbf{G}$
- Prior is density matrix $\mathbf{D}(\mathbf{M})$

$$
\inf_{\mathbf{G} \text{ dens. mat.}} \mathbf{tr}(\mathbf{G}(\log \mathbf{G} - \log \mathbf{D}(\mathbf{M}))) - \eta \mathbf{tr}(\mathbf{G} \log \mathbf{D}(\mathbf{y}|\mathbf{M}))
$$

Quantum rel. entr.
Fancier mixture loss

- $\eta = 1$: Generalized Bayes Rule
  - Soft maximum eigenvalue calculation
- $\eta = \infty$: minimized when $\mathbf{G}$ is dyad $\mathbf{uu}^T$ and $\mathbf{u}$ is the eigenvector belonging to a minimum eigenvalue of $-\log \mathbf{D}(\mathbf{y}|\mathbf{M})$

- Special case of Matrix Exponentiated Gradient update
Generalized Bayes Rule Again

\[ \inf_{\mathbf{G}} \text{tr} (\mathbf{G} \log \mathbf{G}) - \text{tr} (\mathbf{G} \log \mathbf{D}(\mathbf{M})) - \eta \text{tr} (\mathbf{G} \log \mathbf{D}(y|M)) \]

- Von Neumann Entropy is just entropy of eigenvalues
- Prior and data treated the same when \( \eta = 1 \)
- Commutativity
Where does data likelihood matrix $D(y|M)$ come from?

From a joint distribution on space $(Y, M)$
Joint Distributions

Conventional joints:

- Two sets of elementary events - $A$ and $B$
- Joint space $A \times B$
- Elementary events are pairs $(a_i, b_j)$
- Joint distribution is a probability vector over pairs

Generalized joints:

- Two real vector spaces: $A$ and $B$ of dimension $n_A$ and $n_B$
- Joint space: tensor product $A \otimes B$ - real space of dimension $n_An_B$
- Elementary events are dyads of joint space
- Joint distribution is a density matrix over joint space
Joint Probability?

Given joint density matrix $D(A, B)$
a dyad $aa^T$ from space $A$
a dyad $bb^T$ from space $B$
What’s the joint probability of $aa^T$ and $bb^T$?

- $D(a, b) =$?
- Recall $D(a) = \text{tr}(D(A) aa^T)$.
- Thus $D(a, b) = \text{tr}(D(A, B))$?

Conventional: look up probability of jointly specified event $(a_i, b_j)$ in joint table

What is a jointly specified dyad?
**Kronecker Product**

*Kronecker* product of $n \times m$ matrix $A$ and $p \times q$ matrix $B$ is a $np \times mq$ matrix $A \otimes B$ which in block form is given as:

$$A \otimes B = \begin{pmatrix}
    a_{11}B & a_{12}B & \ldots & a_{1m}B \\
    a_{21}B & a_{22}B & \ldots & a_{2m}B \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n1}B & a_{n2}B & \ldots & a_{nm}B
\end{pmatrix}$$

Properties:

- $(A \otimes B)(C \otimes D) = AC \otimes BD$
- $\text{tr}(A \otimes B) = \text{tr}(A)\text{tr}(B)$
- If $D(A)$ and $D(B)$ are density matrices, then so is $D(A) \otimes D(B)$
  - $(a \otimes b)(a \otimes b)^T = aa^T \otimes bb^T$
    - is a dyad of space $(A, B)$
Joint Probability

Use $aa^T \otimes bb^T$ as \textit{jointly specified dyad}

Joint probability: $D(a, b) = \text{tr}(D(A, B)(aa^T \otimes bb^T))$

\textbf{Not every dyad on the joint space can be written as $aa^T \otimes bb^T$!!!}

This issue in quantum physics is known as \textit{entanglement}
More!

- Conditionals
  - Marginalization
  - Theorem of total probability

- Need additional Kronecker product properties
  - Partial trace, etc

- Goes beyond the scope of this talk

- Many subtle quantum physics issues show up in the calculus
Sample Calculus Rules

- \( D(A) = \text{tr}_B(\mathbf{D}(A, B)) \)
- \( D(A, b) = \text{tr}_B(\mathbf{D}(A, B)(I_A \otimes bb^\top)) \)
  
  **Marginalization**

- \( D(A|B) = D(A, B) \odot (I_A \otimes \mathbf{D}(B))^{-1} \)
  
  **Conditional in terms of the joint**
  
  Introduced by Cerf and Adami

- \( D(A) = \text{tr}_B(\mathbf{D}(A|B) \odot (I_A \otimes \mathbf{D}(B))) \)
  
  **Theorem of total probability**

- \( D(M|y) = \frac{D(M) \odot D(y|M)}{\text{tr}(D(M) \odot D(y|M))} \)
  
  **Our Bayes rule**

- \( D(b|A) = D(b)D(A|b) \odot (D(A|B) \odot (I_A \otimes \mathbf{D}(B)))^{-1} \)
  
  **Another Bayes rule**
Summary

- We maintain uncertainty about direction of maximum variance with a density matrix
- Update generalizes conventional Bayes’s rule
- Motivate the update based on a maxent principle
- Probability calculus that retains conventional probabilities as a special case
Outlook

- Calculus for other matrix classes
- On-line update for PCA :-) 
- Other applications
- Connections to quantum computation