A Bayesian Probability Calculus for Density Matrices

Manfred K. Warmuth and Dima Kuzmin

University of California - Santa Cruz

Web: Google.com "manfred"

The Technion, Haifa, Israel, Nov 4, 2013

Outline

1. Multiplicative updates - the genesis of this research
2. Conventional and Generalized Probability Distributions
3. Conventional and generalized Bayes rule
4. Bounds, derivation, calculus

Outline

1. Multiplicative updates - the genesis of this research
2. Conventional and Generalized Probability Distributions
3. Conventional and generalized Bayes rule
4. Bounds, derivation, calculus

Multiplicative updates

- A set of experts i learns something online
- Maintain one weight w_i per expert i
- Multiplicative update

$$w_i := \frac{w_i \; e^{-\eta \; L_i}}{\text{normalization}}$$

- Bayes rule is special case
  when $\eta = 1, \quad L_i := -\log(P(y|M_i)), \quad$ and $\quad w_i = P(M_i)$

$$P(M_i|y) := \frac{P(M_i) \;\; P(y|M_i)}{\text{normalization}}$$

- Motivated by using a relative entropy regularization
- Weight vector is uncertainty vector over experts / models

Matrix version of multiplicative updates

- Maintain density matrix $\mathbf{W}$ as a parameter
- Multiplicative update

$$\mathbf{W} := \frac{\mathbf{exp}(\mathbf{log}\,\mathbf{W} - \eta\,\mathbf{L})}{\text{normalization}}$$

- Motivated by using a quantum relative entropy regularization
- Density matrix is uncertainty over rank 1 subspaces
- Capping the eigenvalues at $\frac{1}{k}$: uncertainty over rank k subspaces

Matrix version of multiplicative updates

- Maintain density matrix $\mathbf{W}$ as a parameter
- Multiplicative update

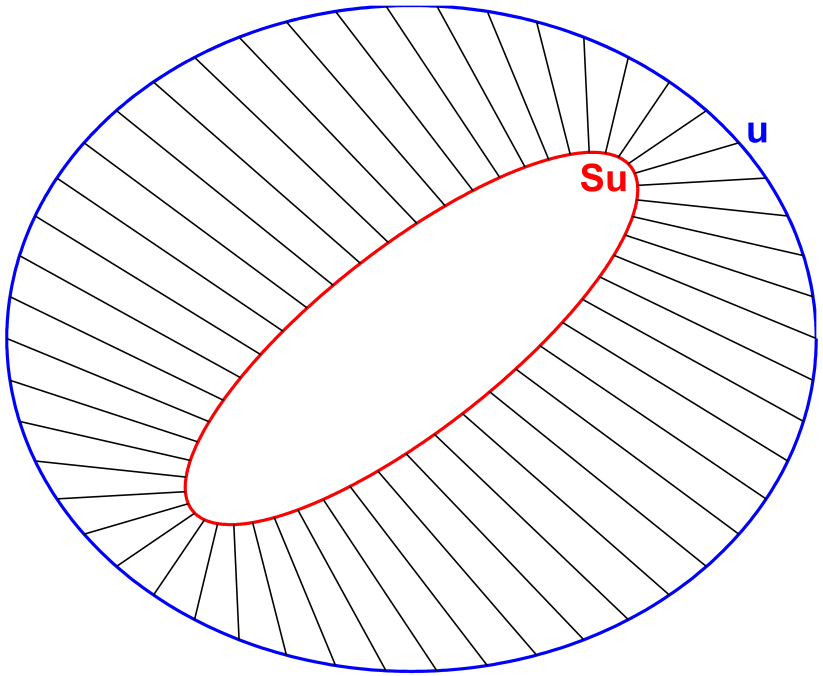
$$\mathbf{W} := \frac{\mathbf{exp}(\mathbf{log}\,\mathbf{W} - \eta\,\mathbf{L})}{\text{normalization}}$$

- Motivated by using a quantum relative entropy regularization
- Density matrix is uncertainty over rank 1 subspaces
- Capping the eigenvalues at $\frac{1}{k}$: uncertainty over rank k subspaces

- Today: What corresponds to Bayes ???

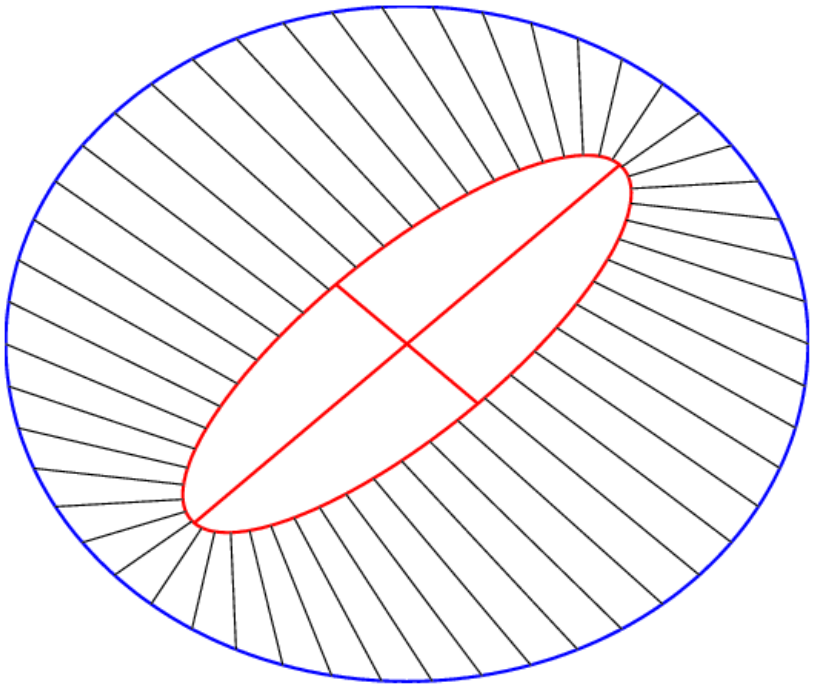
Visualisations: ellipses

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



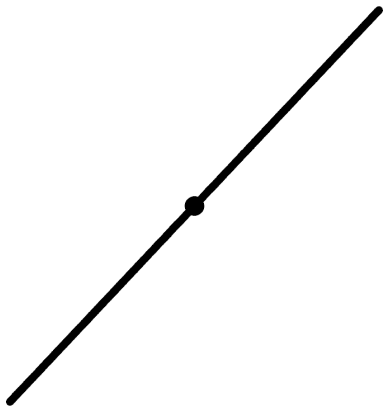
- Ellipse $= \{\mathbf{Su} : \|\mathbf{u}\|_2 = 1\}$
- Dotted lines connect point **u** on unit ball with point **Su** on ellipse

Ellipses cont.



- For symmetric matrices, the eigenvectors form the axes of the ellipse and eigenvalues their lengths
- $\mathbf{Su} = \sigma\mathbf{u}$, $\mathbf{u}$ is an eigenvector, σ is an eigenvalue

Dyads



- One eigenvalue one
- All others zero

From vectors to matrices

Real vectors $\mathbf{x} = \sum_i \underbrace{x_i}_{\text{real coefficients}} \mathbf{e}_i$

Symmetric matrices $\mathbf{S} = \sum_i \underbrace{\sigma_i}_{\text{real coefficients}} \underbrace{\mathbf{s}_i \mathbf{s}_i^\top}_{\text{dyads}}$

i.e. linear combinations of unit vectors / dyads

From vectors to matrices

Real vectors $\mathbf{x} = \sum_i \underbrace{x_i}_{\text{real coefficients}} \mathbf{e}_i$

Symmetric matrices $\mathbf{S} = \sum_i \underbrace{\sigma_i}_{\text{real coefficients}} \underbrace{\mathbf{s}_i \mathbf{s}_i^\top}_{\text{dyads}}$

i.e. linear combinations of unit vectors / dyads

Probability vectors $\boldsymbol{\omega} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \mathbf{e}_i$

Density matrices $\mathbf{W} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{w}_i \mathbf{w}_i^\top}_{\text{pure density matrices}}$

i.e. mixtures of unit vectors / dyads

From vectors to matrices

Real vectors $\mathbf{x} = \sum_i \underbrace{x_i}_{\text{real coefficients}} \mathbf{e}_i$

Symmetric matrices $\mathbf{S} = \sum_i \underbrace{\sigma_i}_{\text{real coefficients}} \underbrace{\mathbf{s}_i \mathbf{s}_i^\top}_{\text{dyads}}$

i.e. linear combinations of unit vectors / dyads

Probability vectors $\boldsymbol{\omega} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \mathbf{e}_i$

Density matrices $\mathbf{W} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{w}_i \mathbf{w}_i^\top}_{\text{pure density matrices}}$

i.e. mixtures of unit vectors / dyads

Matrices are generalized distributions

Outline

1. Multiplicative updates - the genesis of this research
2. Conventional and Generalized Probability Distributions
3. Conventional and generalized Bayes rule
4. Bounds, derivation, calculus

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

$$\{\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_5\} = (0,\ 1,\ 1,\ 0,\ 1)^\top$$

- Distribution is probability vector

$$(.1, .2, .3, .1, .3)^\top$$

- Probability of event $0 \cdot .1 + 1 \cdot .2 + 1 \cdot .3 + 0 \cdot .1 + 1 \cdot .3 = .8$

Conventional case in matrix notation

- The n elementary events are matrices with a single one on diagonal

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{e}_4\mathbf{e}_4^\top$$

- Event is diagonal matrix $\mathbf{P}$ with $\{0, 1\}$ entries

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \mathbf{e}_2\mathbf{e}_2^\top + \mathbf{e}_3\mathbf{e}_3^\top + \mathbf{e}_5\mathbf{e}_5^\top$$

- Distribution is diagonal matrix $\mathbf{W}$
  with probability distribution along the diagonal

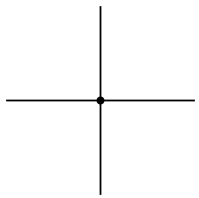
$$\mathbf{W} = \begin{pmatrix} .1 & 0 & 0 & 0 & 0 \\ 0 & .2 & 0 & 0 & 0 \\ 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & 0 & .1 & 0 \\ 0 & 0 & 0 & 0 & .3 \end{pmatrix}$$

- Probability of event $\mathbf{P}$ wrt distribution $\mathbf{W}$ is $\mathrm{tr}(\mathbf{WP}) = \mathbf{W} \bullet \mathbf{P}$

Elementary events = states in quantum physics

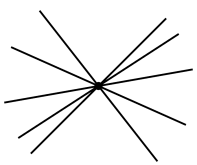
- Conventional: finitely many states

$$\{\mathbf{e}_i\mathbf{e}_i^\top : 1 \le i \le n\}$$



- Generalized: continuously many states $\mathbf{u}\mathbf{u}^\top$ called dyads

$$\{\mathbf{u}\mathbf{u}^\top : \mathbf{u} \in \mathbb{R}^n,\ ||\mathbf{u}||_2 = 1\}$$



  - Degenerate ellipses
  - $\mathbf{u}\mathbf{u}^\top$ one-dimensional projection matrix onto direction $\mathbf{u}$

- Prefer to use the dyads $\mathbf{u}\mathbf{u}^\top$ as our states instead of the unit vectors $\mathbf{u}$

Generalized probabilites over $\mathbb{R}^n$

- Elementary events are the dyads $\mathbf{u}\mathbf{u}^\top$
- Event is symmetric matrix $\mathbf{P}$ with $\{0, 1\}$ eigenvalues

$$\mathbf{P} = \mathcal{U} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \mathcal{U}^\top = \mathbf{u}_2\mathbf{u}_2^\top + \mathbf{u}_3\mathbf{u}_3^\top + \mathbf{u}_5\mathbf{u}_5^\top$$

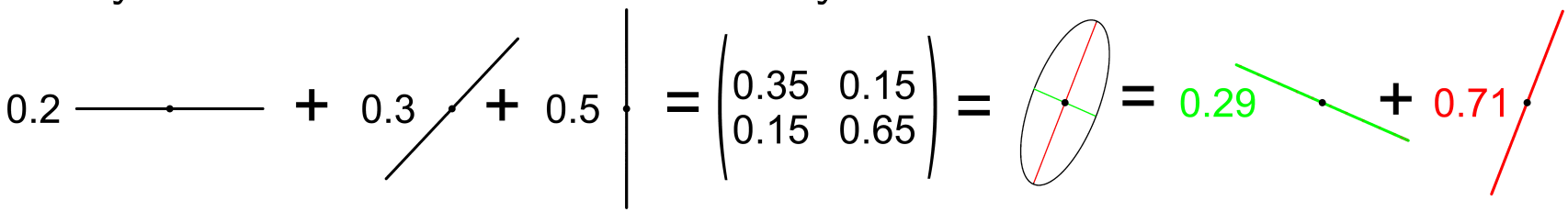
  - $\mathcal{U}$ orthogonal eigensystem, u_i columns/eigenvectors
  - $\mathbf{P}$ projection matrix onto arbitrary subspace of $\mathbb{R}^n$: $\mathbf{P}^2 = \mathbf{P}$
- Distribution is density matrix $\mathbf{W}$:

$$\mathbf{W} = \mathcal{U} \begin{pmatrix} .1 & 0 & 0 & 0 & 0 \\ 0 & .2 & 0 & 0 & 0 \\ 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & 0 & .1 & 0 \\ 0 & 0 & 0 & 0 & .3 \end{pmatrix} \mathcal{U}^\top = \underbrace{\sum_i \omega_i \, \mathbf{u}_i\mathbf{u}_i^\top}_{\text{mixture of dyads}}$$

  Eigenvalues ω_i form probability vector
- Event probabilities become traces (later)

Density matrices = mixtures of dyads

- Many mixtures lead to same density matrix



- There always exists a decomposition into n dyads that correspond to eigenvectors

Alternate defnitions of density matrices?

- Symmetric: $\mathbf{W}^\top = \mathbf{W}$
- Positive definite: $\mathbf{u}^\top \mathbf{W}\mathbf{u} \geq 0 \quad \forall \mathbf{u}$
- Trace one: sum of diagonal elements is one

Variance

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

$$\mathbf{C} = \mathbb{E}\left((\mathbf{c} - \mathbb{E}(\mathbf{c}))(\mathbf{c} - \mathbb{E}(\mathbf{c}))^\top\right)$$

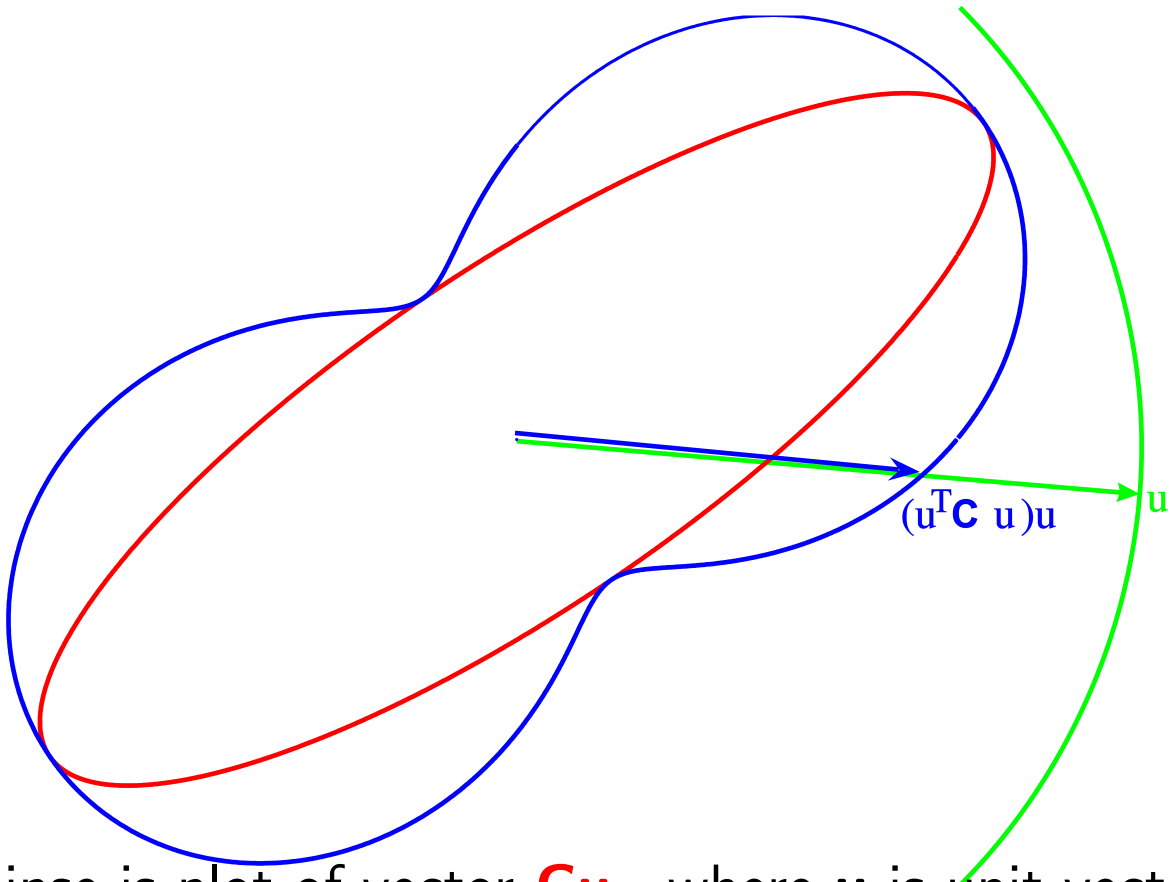
- The variance along any vector $\mathbf{u}$ is

$$\begin{aligned}
\mathbb{V}(\mathbf{c}^\top \mathbf{u}) &= \mathbb{E}(\left(\mathbf{c}^\top \mathbf{u} - \mathbb{E}(\mathbf{c}^\top \mathbf{u})\right)^2) \\
&= \mathbb{E}(\left((\mathbf{c}^\top - \mathbb{E}(\mathbf{c}^\top))\,\mathbf{u}\right)^2) \\
&= \mathbf{u}^\top \underbrace{\mathbb{E}\left((\mathbf{c} - \mathbb{E}(\mathbf{c}))(\mathbf{c} - \mathbb{E}(\mathbf{c}))^\top\right)}_{\mathbf{C}} \mathbf{u}
\end{aligned}$$

- Variance as trace

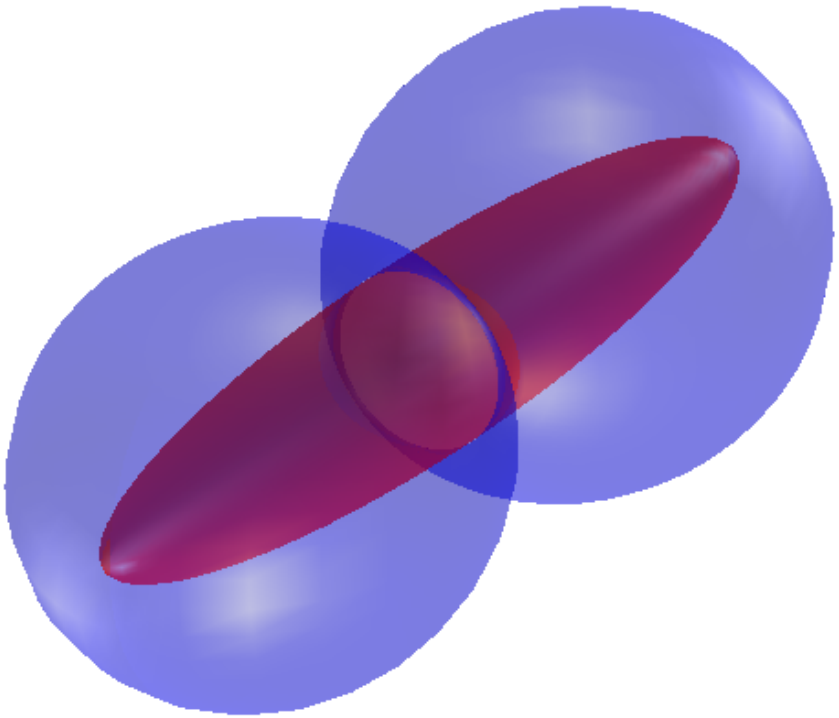
$$\mathbf{u}^\top \mathbf{C} \mathbf{u} = \operatorname{tr}(\mathbf{u}^\top \mathbf{C} \mathbf{u}) = \operatorname{tr}(\mathbf{C}\,\mathbf{u}\mathbf{u}^\top) = \mathbf{C} \bullet \mathbf{u}\mathbf{u}^\top \geq 0$$

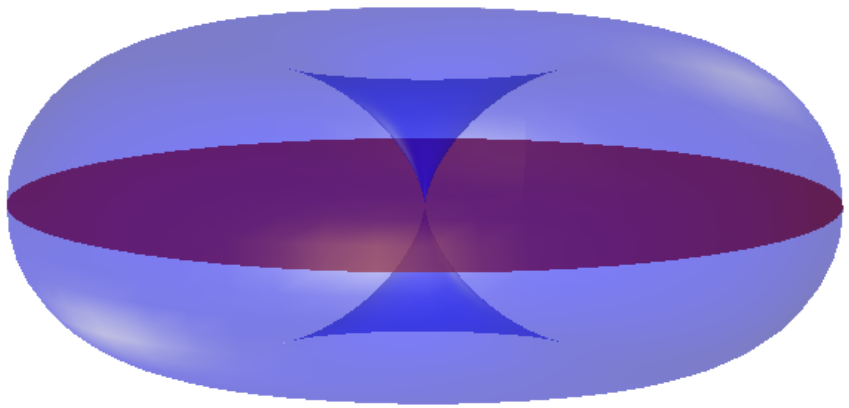
Plotting the variance



Curve of the ellipse is plot of vector **Cu** , where **u** is unit vector
The outer figure eight is direction **u** times the variance $\mathbf{u}^\top \mathbf{Cu}$
For an eigenvector, this variance equals the eigenvalue
and touches the ellipse

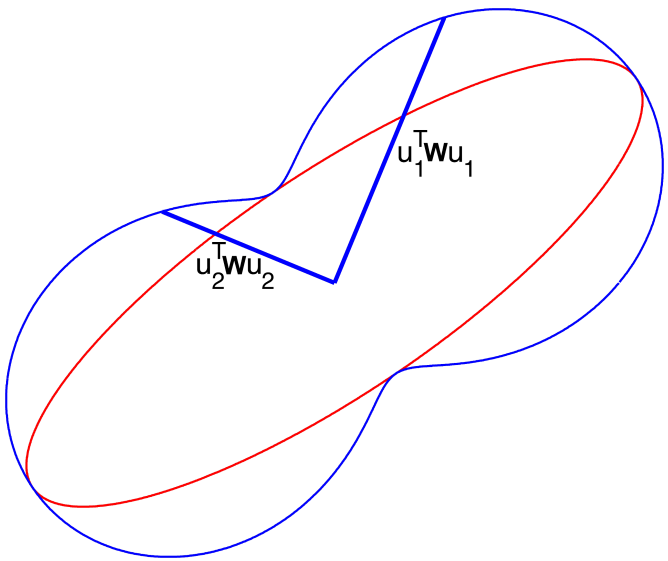
3 dimensional variance plots

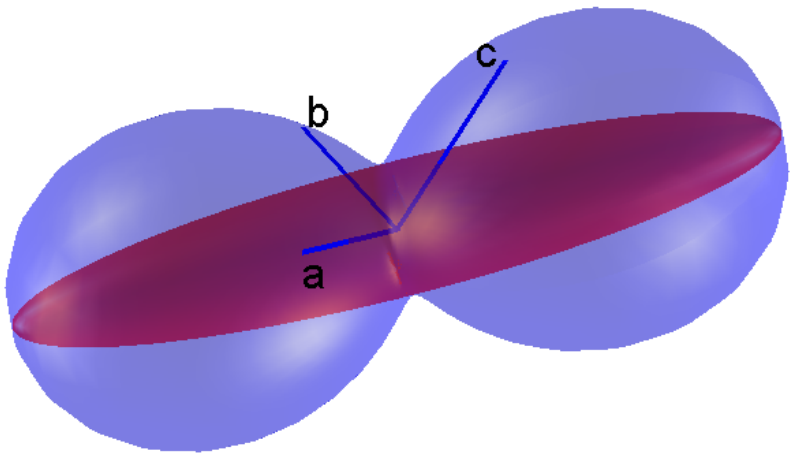




Assignment of generalized probabilities

- Density matrix $\mathbf{W}$ assigns generalized probability $\mathbf{u}^\top \mathbf{W} \mathbf{u} = \mathrm{tr}(\mathbf{W}\,\mathbf{u}\mathbf{u}^\top)$ to dyad $\mathbf{u}\mathbf{u}^\top$
- Sum of probabilities over an orthornormal basis $\mathbf{u}_i$ is 1





For any two orthogonal directions:
$\mathbf{u}_1^\top \mathbf{W} \mathbf{u}_1 + \mathbf{u}_2^\top \mathbf{W} \mathbf{u}_2 = 1$

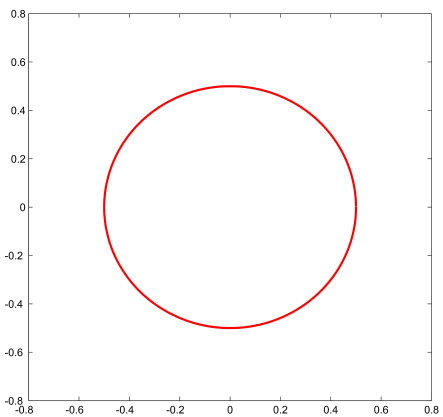
$a + b + c = 1$

Sum of probabilities over an orthornormal basis

- $$\sum_i \mathrm{tr}(\mathbf{W}\,\mathbf{u}_i\mathbf{u}_i^\top) = \mathrm{tr}(\mathbf{W}\sum_i \mathbf{u}_i\mathbf{u}_i^\top) = \mathrm{tr}(\mathbf{W}\underbrace{\mathcal{U}\mathcal{U}^\top}_{\mathbf{I}}) = \mathrm{tr}(\mathbf{W}) = 1$$

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

$$\mathrm{tr}(\frac{1}{n}\mathbf{I}\ \ \mathbf{u}\mathbf{u}^\top) = \frac{1}{n}\,\mathrm{tr}(\mathbf{u}\mathbf{u}^\top) = \frac{1}{n}$$



- All dyads have generalized probability $\frac{1}{n}$
- Generalized probability of n orthogonal dyads sums to 1

- Classical: for any probability vector $\boldsymbol{\omega}$

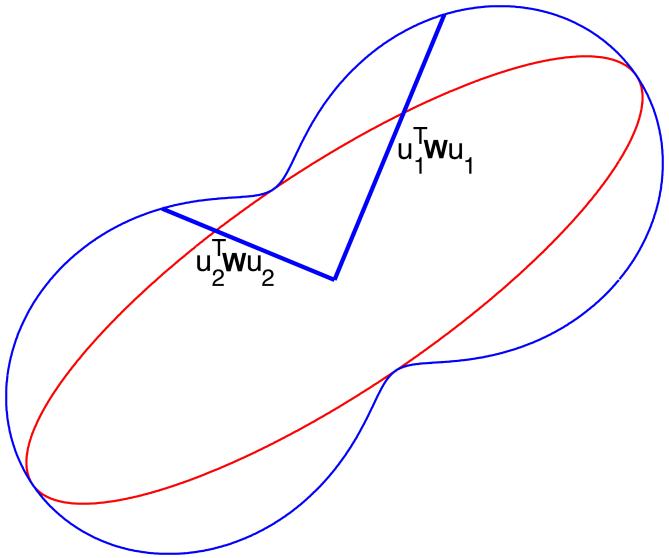
$$\sum_i \mathrm{tr}(\mathrm{diag}(\boldsymbol{\omega})\,\mathbf{e}_i\mathbf{e}_i^\top) = \sum_i \omega_i = 1$$

Total variance along orthogonal set of directions

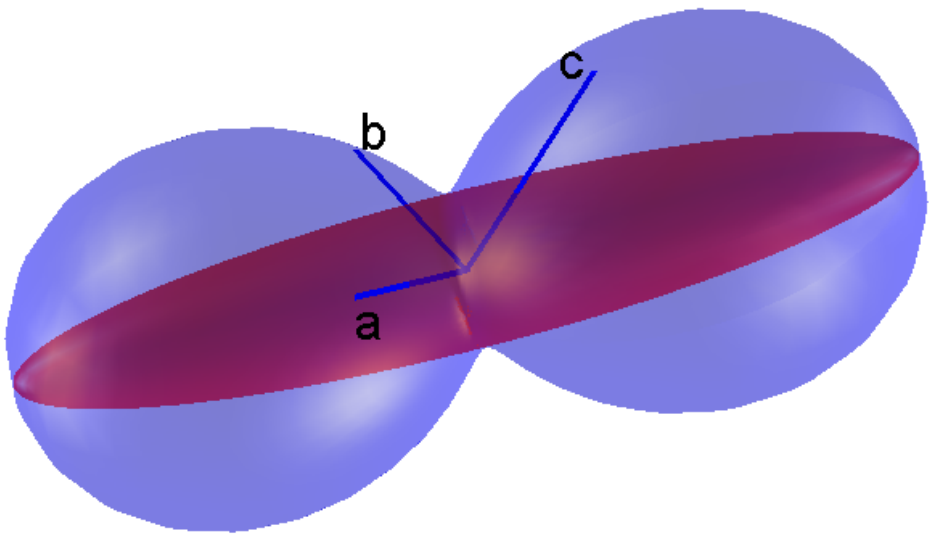
A density matrix

For any two orthogonal directions

$$\mathbf{u}_1^\top \mathbf{A}\mathbf{u}_1 + \mathbf{u}_2^\top \mathbf{A}\mathbf{u}_2 = 1$$



$$a + b + c = 1$$



Gleason's Theorem

Definition

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

- $\forall \mathbf{u}$, $0 \leq \mu(\mathbf{u}) \leq 1$
- If $\mathbf{u}_1, \ldots, \mathbf{u}_n$ form an orthonormal basis for $\mathbb{R}^n$, then $\sum \mu(\mathbf{u}_i) = 1$

Theorem

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

$$\mu(\mathbf{u}) = \mathrm{tr}(\mathbf{W}\,\mathbf{u}\mathbf{u}^\top)$$

for a uniquely defined density matrix $\mathbf{W}$

Key generalization

- Disjoint events now correspond to orthogonal events
- When events have same eigensystem, then

$$\mathrm{tr}\left(\underbrace{\boldsymbol{u}\begin{pmatrix}1&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&1\end{pmatrix}\boldsymbol{u}^\top\ \boldsymbol{u}\begin{pmatrix}0&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix}\boldsymbol{u}^\top}_{\text{orthogonal}}\right) = 0 \quad \text{iff} \quad \underbrace{\begin{pmatrix}1\\0\\0\\0\\1\end{pmatrix}\cdot\begin{pmatrix}0\\1\\1\\0\\0\end{pmatrix}}_{\text{disjoint}} = 0$$

Probability of events

- Generalized probability of event $\mathbf{P}$ is

$$\mathrm{tr}(\mathbf{WP}) = \mathrm{tr}\Big(\sum_i \omega_i \, \mathbf{w}_i \mathbf{w}_i^\top \mathbf{P}\Big) = \underbrace{\sum_i \overbrace{\omega_i}^{\text{probability}} \; \overbrace{\mathbf{w}_i^\top \mathbf{P} \mathbf{w}_i}^{\text{variance}}}_{\text{expected variance}}$$

- Random variable is symmetric matrix $\mathbf{S}$
  Expectation of $\mathbf{S}$ is

$$\mathrm{tr}(\mathbf{WS}) = \mathrm{tr}\Big(\mathbf{W} \sum_i \sigma_i \mathbf{s}_i \mathbf{s}_i^\top\Big) = \underbrace{\sum_i \overbrace{\sigma_i}^{\text{outcome}} \; \overbrace{\mathbf{s}_i^\top \mathbf{W} \, \mathbf{s}_i}^{\text{probability}}}_{\text{expected outcome}}$$

Quantum measurement

Quantum measurement for mixture state $\mathbf{W} = \sum_i \omega_i \mathbf{w}_i \mathbf{w}_i^\top$
and observable $\mathbf{S} = \sum_i \sigma_i \mathbf{s}_i \mathbf{s}_i^\top$:

- After measurement, state collapses into one of $\{\mathbf{s}_1\mathbf{s}_1^\top, \ldots, \mathbf{s}_n\mathbf{s}_n^\top\}$
- Successor state is $\mathbf{s}_i\mathbf{s}_i^\top$ with probability $\mathbf{s}_i{}^\top \mathbf{W}\, \mathbf{s}_i$
- The expected state is again a density matrix

$$\mathbf{W} \longrightarrow \sum_i \mathbf{s}_i\mathbf{s}_i^\top \mathbf{W}\, \mathbf{s}_i\mathbf{s}_i^\top = \underbrace{\sum_i \overbrace{\mathbf{s}_i^\top \mathbf{W}\, \mathbf{s}_i}^{\text{probability}}\ \mathbf{s}_i\mathbf{s}_i^\top}_{\text{expected state}}$$

- Successor state $\mathbf{s}_i\mathbf{s}_i^\top$ associated with outcome σ_i
- Expected outcome $\mathrm{tr}(\mathbf{WS}) = \sum_i \sigma_i\ \mathbf{s}_i^\top \mathbf{W}\, \mathbf{s}_i$
- We use the trace computations but our density matrices are updated differently

Outline

1. Multiplicative updates - the genesis of this research
2. Conventional and Generalized Probability Distributions
3. Conventional and generalized Bayes rule
4. Bounds, derivation, calculus

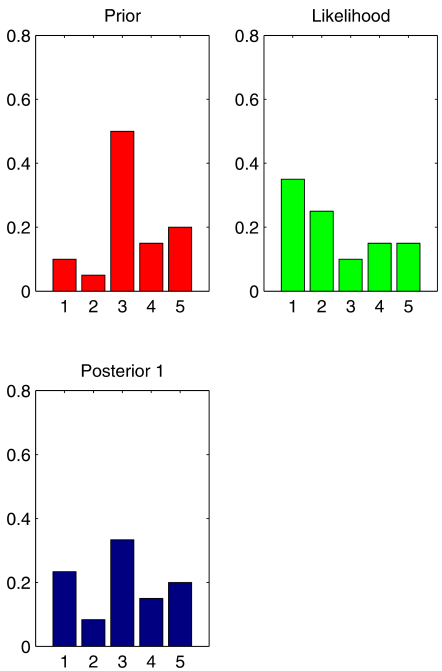
Conventional setup

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

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

Conventional Bayes Rule

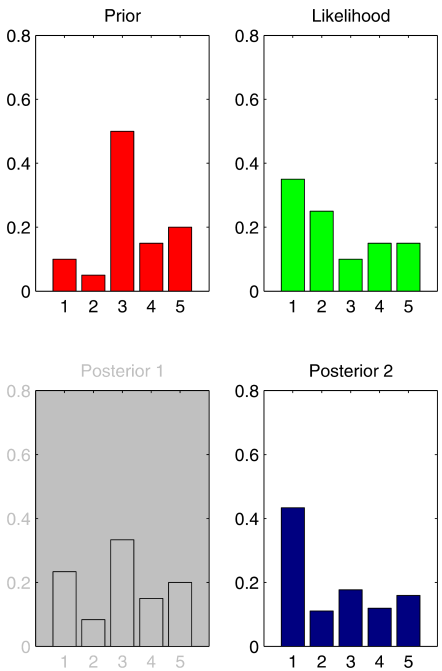
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



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

Conventional Bayes Rule

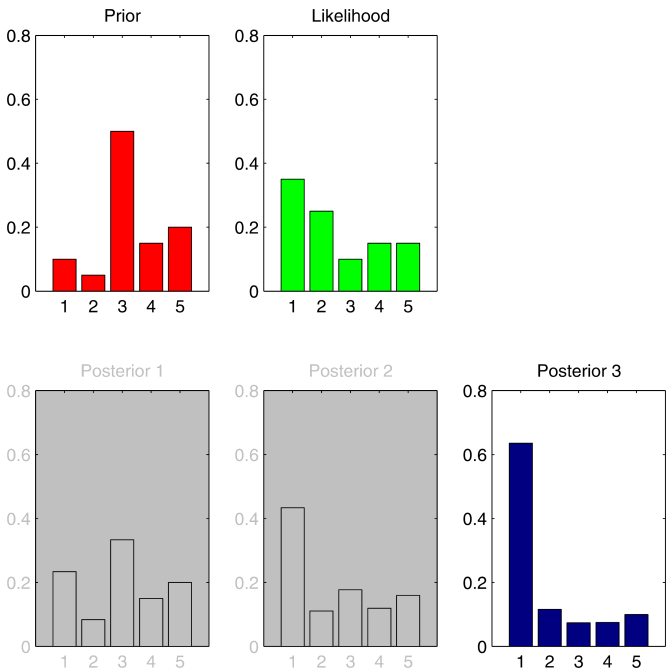
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



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

Conventional Bayes Rule

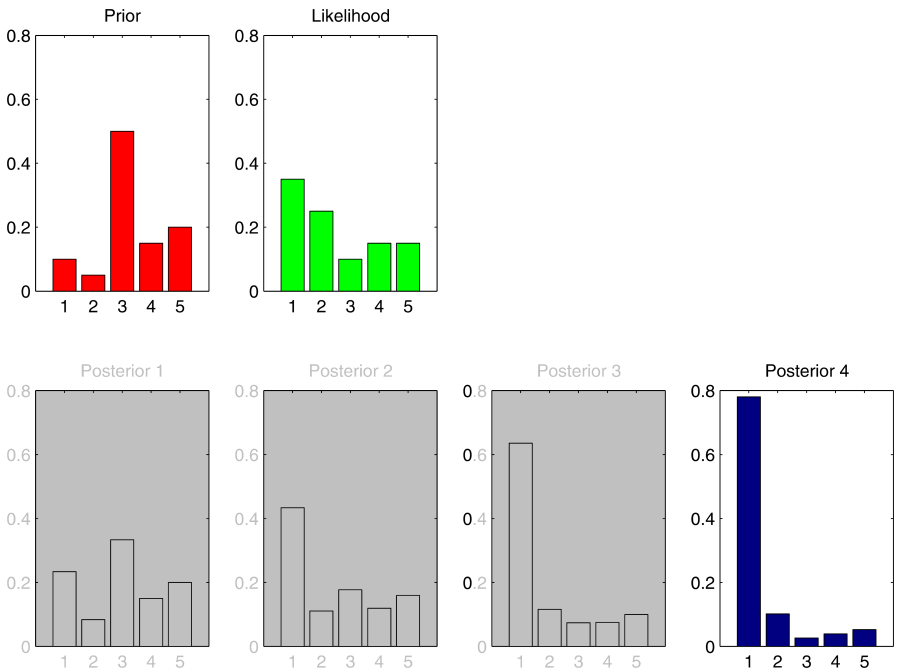
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



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

Conventional Bayes Rule

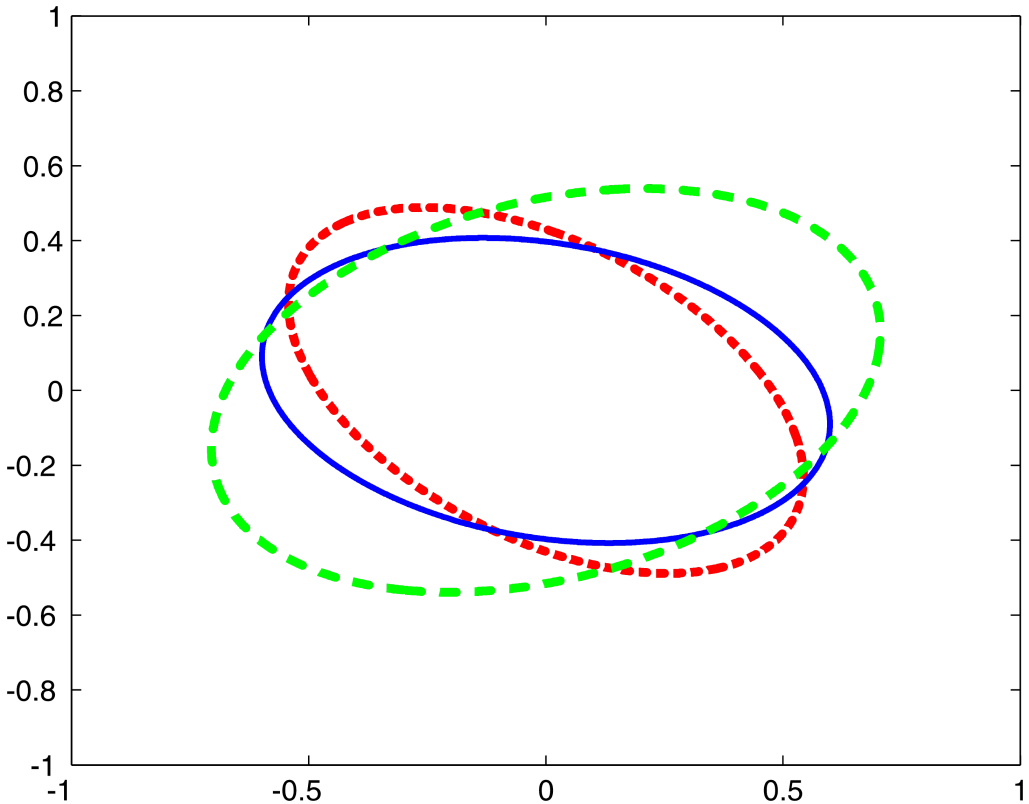
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



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

Bayes Rule for density matrices

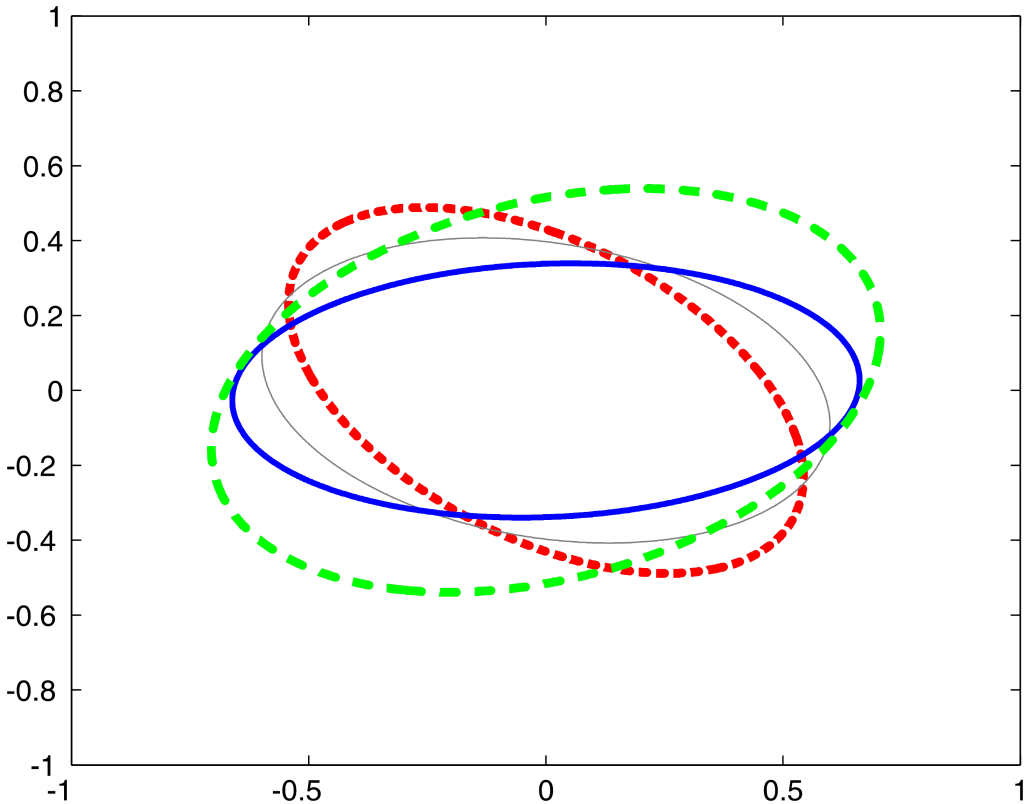
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 1 update with data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

Bayes Rule for density matrices

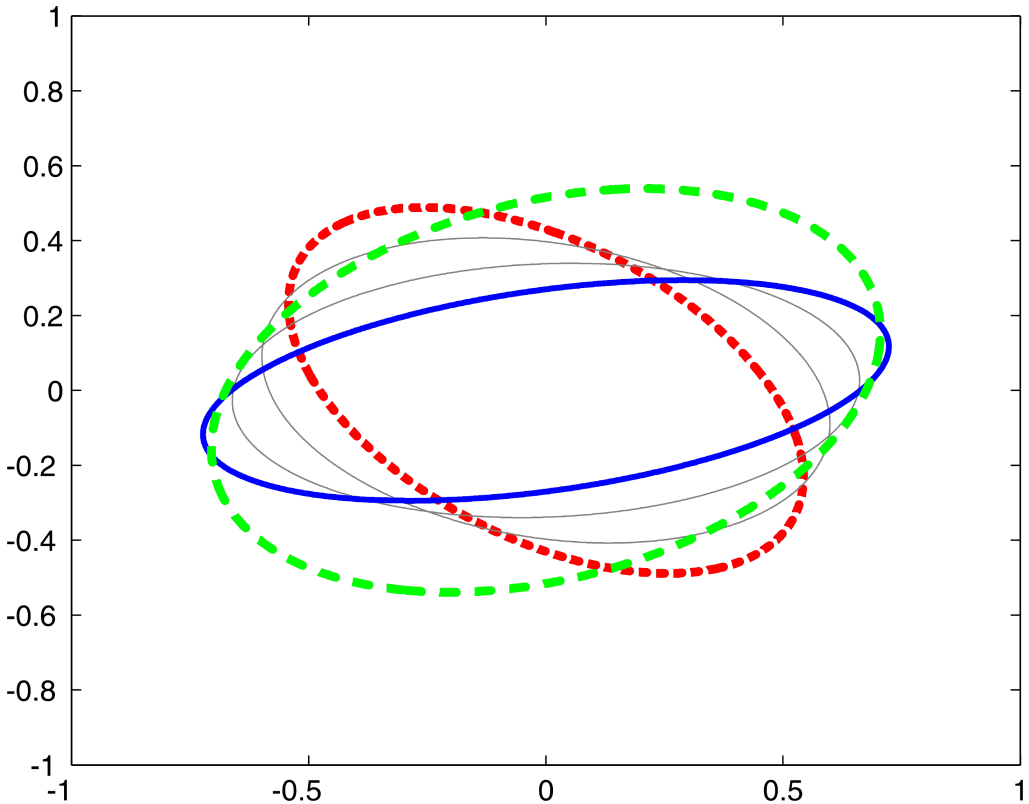
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 2 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

Bayes Rule for density matrices

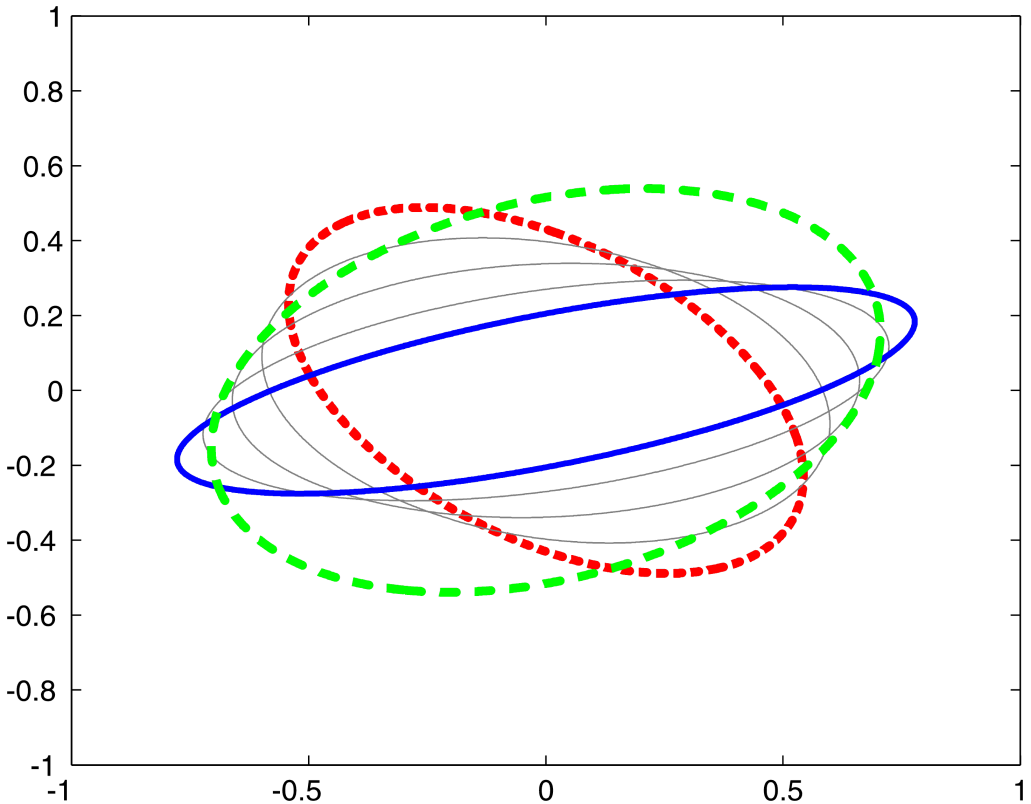
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 3 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

Bayes Rule for density matrices

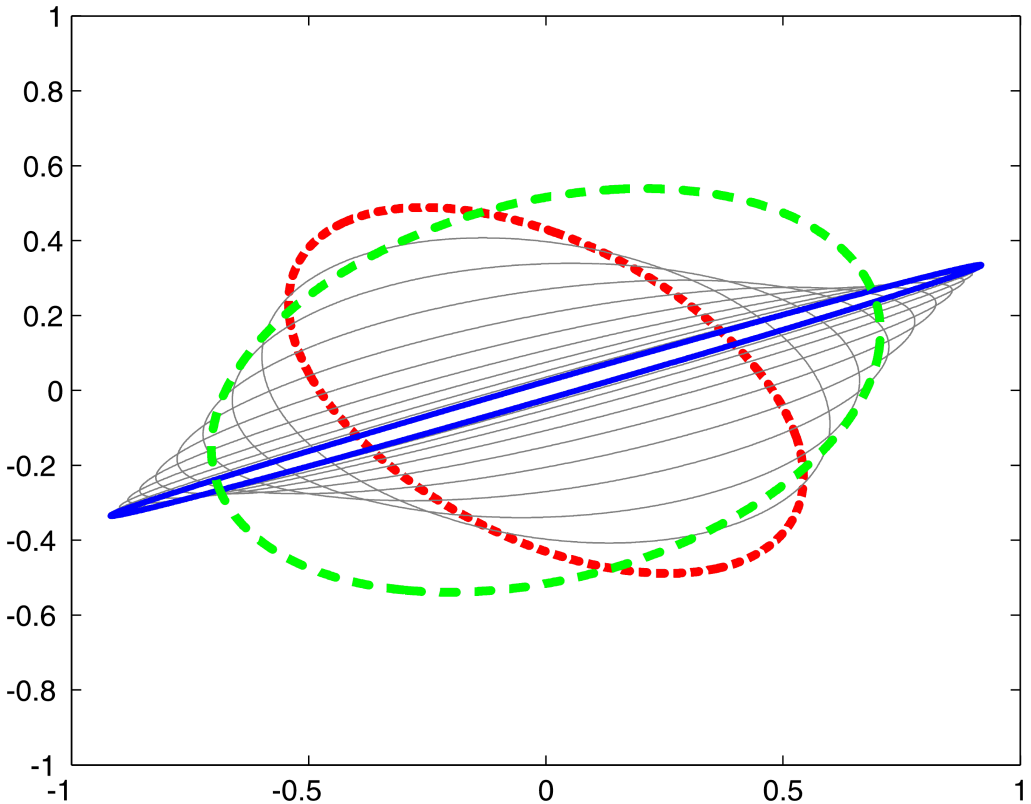
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 4 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

Bayes Rule for density matrices

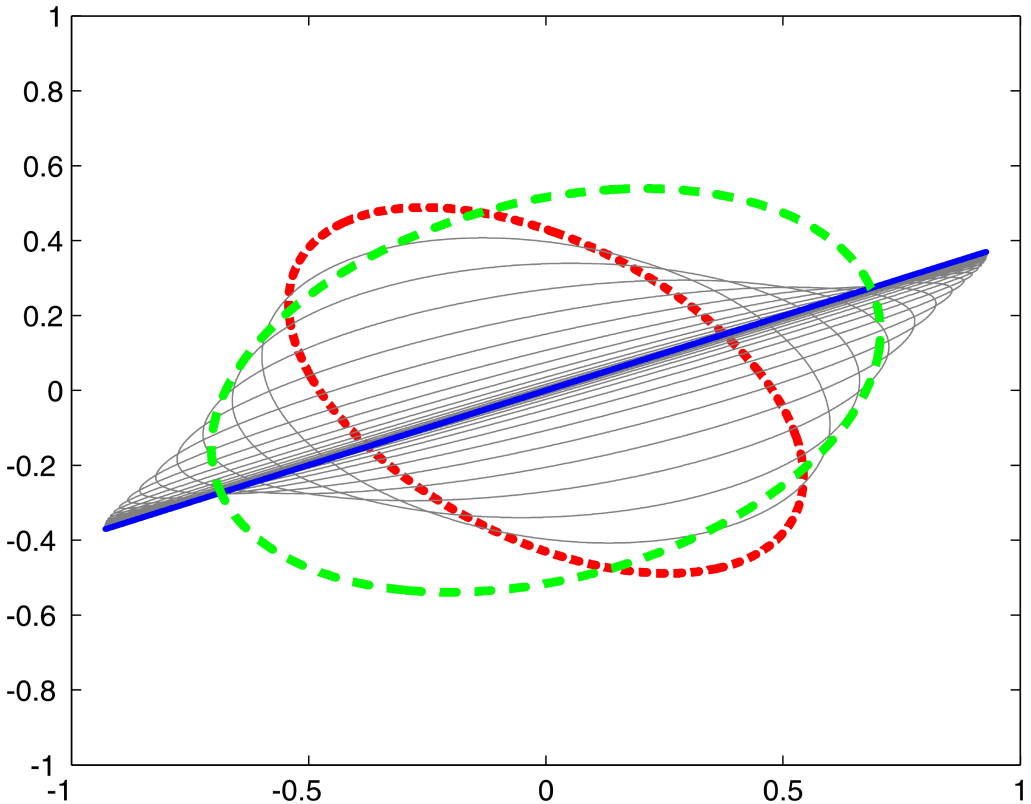
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 10 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

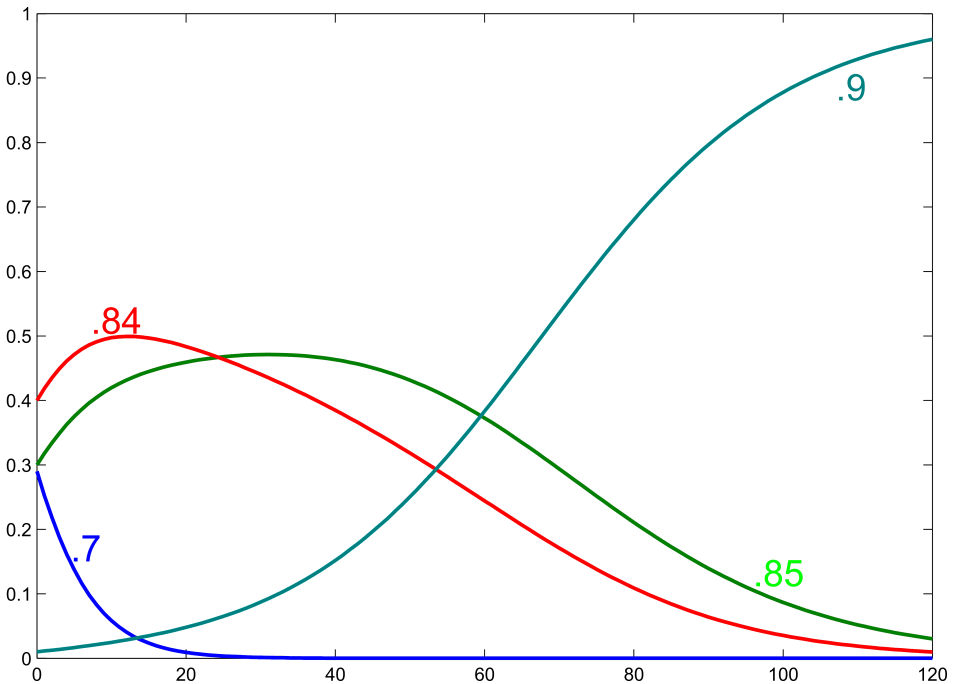
Bayes Rule for density matrices

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



- 20 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- Soft max eigenvalue calculation

Many iterations of conventional Bayes w. same data likelihood



Plot of posterior probability as a function of the iteration number

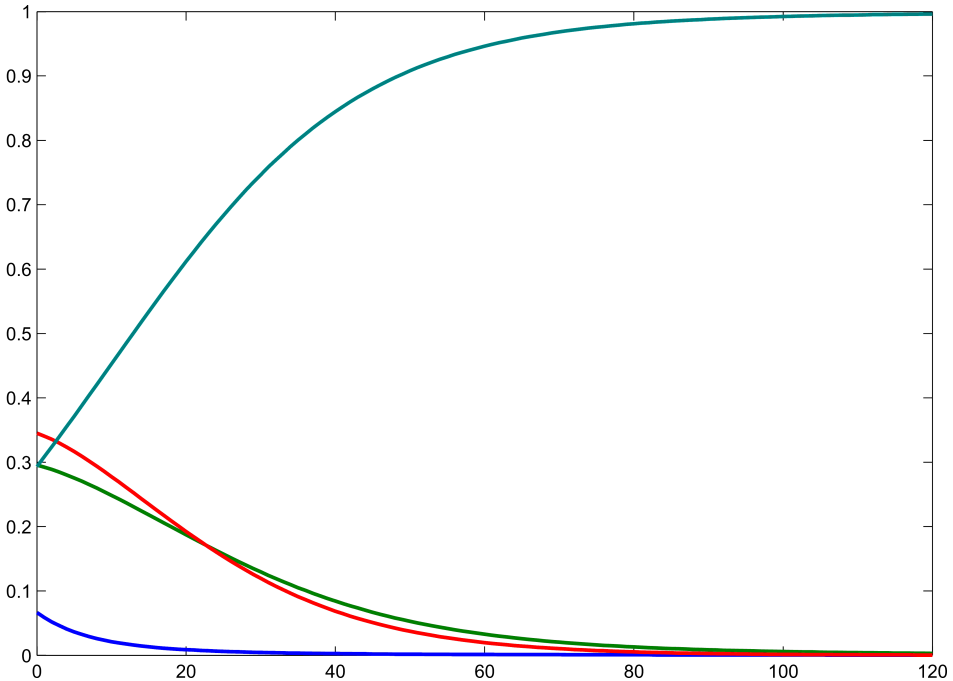
$$(P(M_i)) = (.29, .4, .3, .01) \quad (P(y|M_i)) = (.7, .84, .85, .9)$$

Initially .85 overtakes .84

Eventually .9 overtakes both

Largest likelihood: sigmoid smallest likelihood: reverse sigmoid

Many iterations of generalized Bayes Rule



Prior $\mathbf{D}(\mathbb{M})$ is diagonalized prior $(P(M_i))$ of previous plot

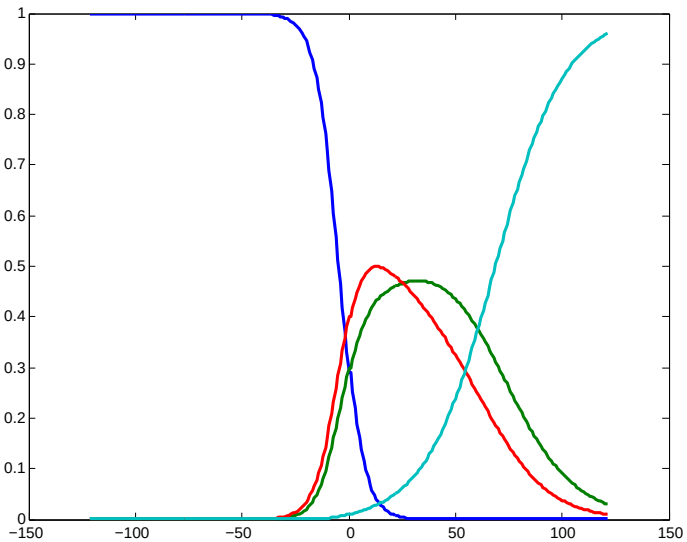
Data likelihood $\mathbf{D}(y|\mathbb{M}) = \mathbf{U}\,\mathrm{diag}((P(y|M_i))\mathbf{U}^T$,

where the eigensystem $\mathbf{U}$ is a random rotation matrix

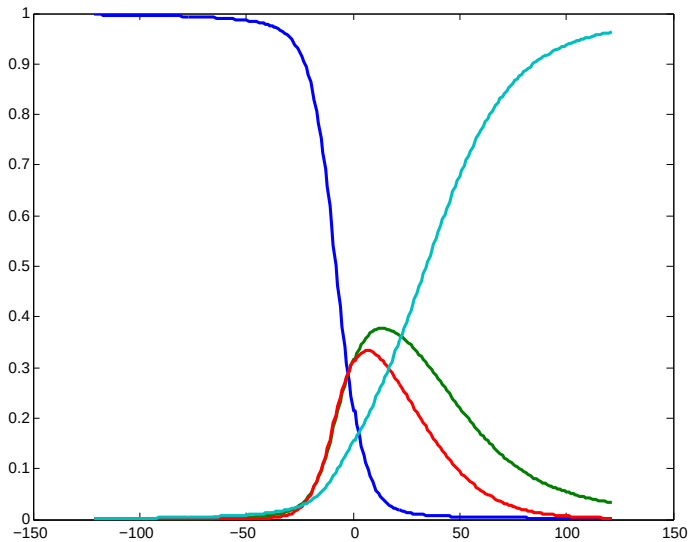
Plot of projections of the posterior onto the four eigendirections $\mathbf{u}_i$

Forward and backward

diagonal



with rotation



Conventional rule as special case

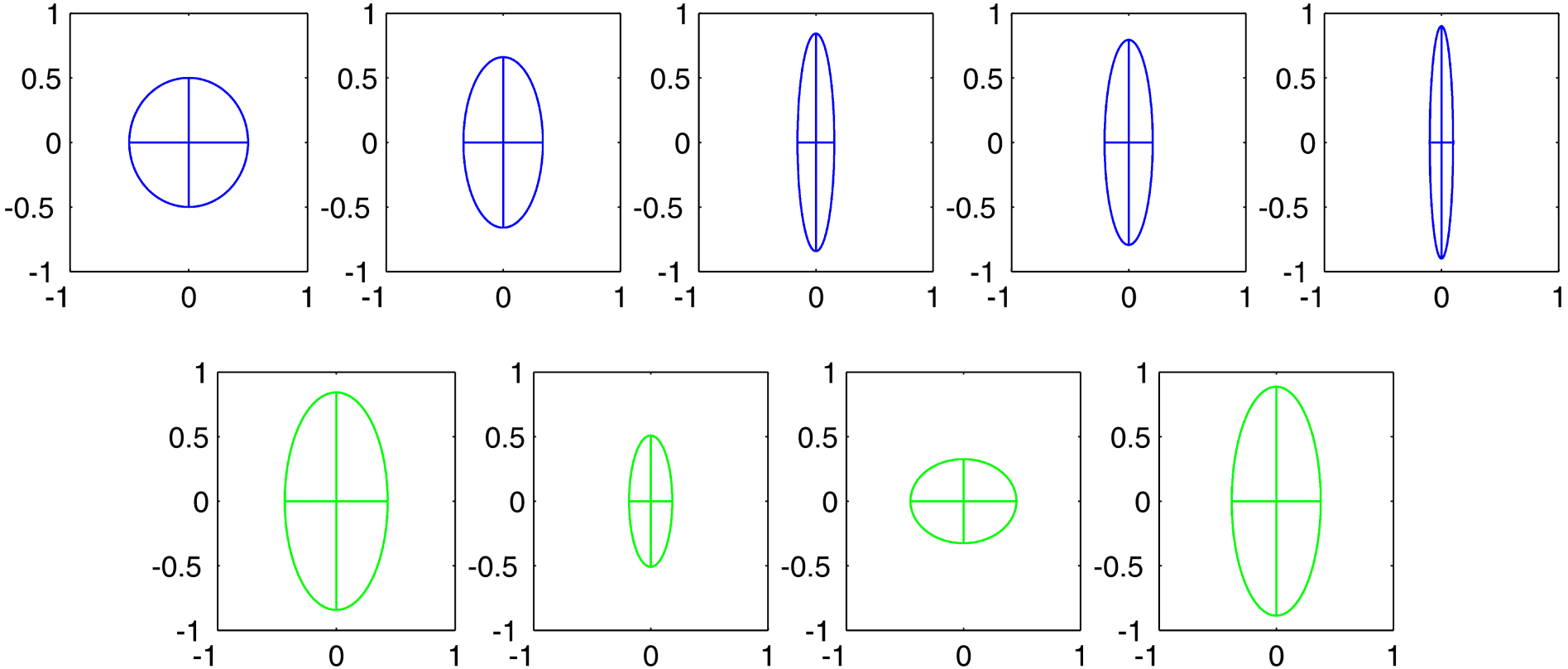
- If $\mathbf{D}(\mathbb{M})$ and $\mathbf{D}(\mathbf{y}|\mathbb{M})$ have the same eigensystem, then generalized Bayes rule specializes the the conventional case

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

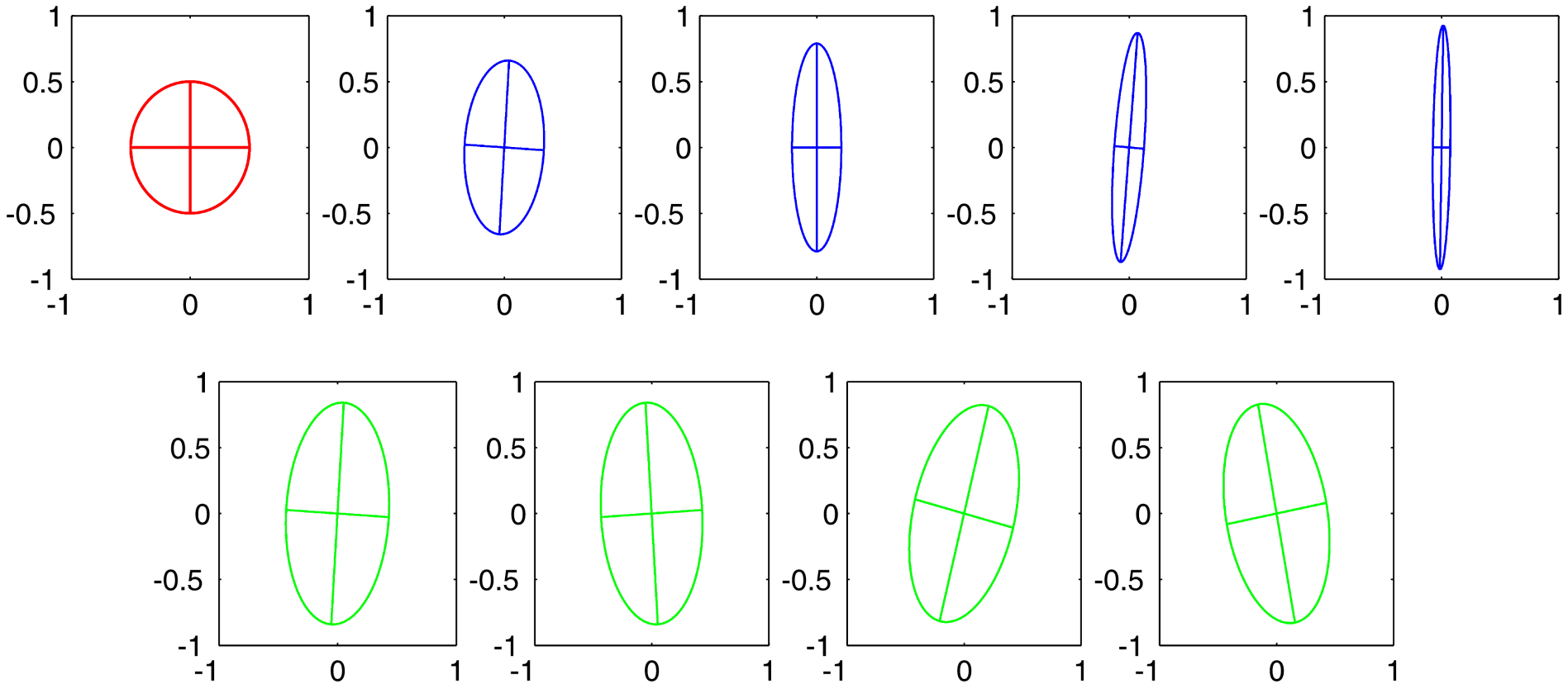
$$\frac{\begin{pmatrix} \ddots & & 0 \\ & P(M_i) & \\ 0 & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & 0 \\ & P(y|M_i) & \\ 0 & & \ddots \end{pmatrix}}{\mathrm{tr}(\ \cdots\ )}$$

Diagonal case in dimension 2

Sequence of 2-dim ellipses for diagonal case



General case in dimension 2



What's the upshot

- Nifty generalization of Bayes rule
- Lots of evidence that it is the right generalization
- We don't have an application yet where the calculus is needed for the analysis

Properties $\mathbf{D}(\mathbb{M})$ and $\mathbf{D}(\mathbf{y}|\mathbb{M})$

$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\mathbf{exp}\left(\mathbf{log}\,\mathbf{D}(\mathbb{M}) + \mathbf{log}\,\mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$

$\mathbf{D}(\mathbb{M})$

- symmetric positive definite
- eigenvalues sum to one

$\mathbf{D}(\mathbf{y}|\mathbb{M})$

- symmetric positive definite
- all eigenvalues in range [0..1]

Later we discuss where these matrices come from

Non-commutative Bayes Rule?

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

$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\overbrace{\mathbf{exp}\Big(\ \mathbf{log}\ \overbrace{\mathbf{D}(\mathbb{M})}^{\text{sym.pos.def}}\ +\ \mathbf{log}\ \overbrace{\mathbf{D}(\mathbf{y}|\mathbb{M})}^{\text{sym.pos.def}}\ \Big)}^{\text{sym.pos.def}}}{\mathrm{tr}\,(\cdots)}$$

Intersection properties of Bayes

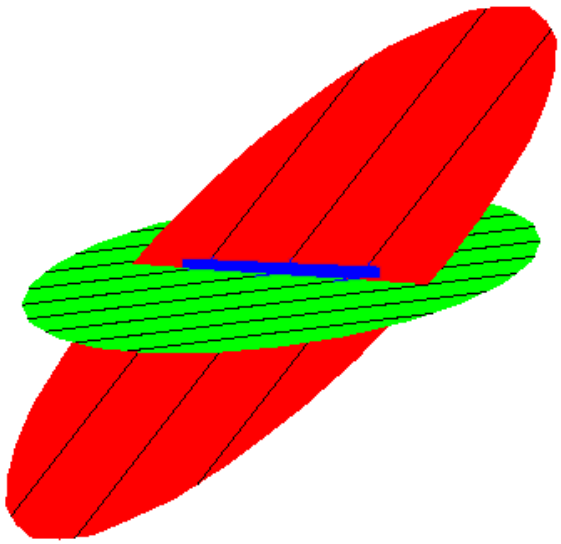
Conventional Bayes:

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

- Computes intersection of two sets

Intersection properties on new Bayes Rule

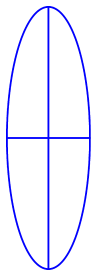
$$\underbrace{\mathbf{exp}(\mathbf{log}\,\mathbf{A} + \mathbf{log}\,\mathbf{B})}$$

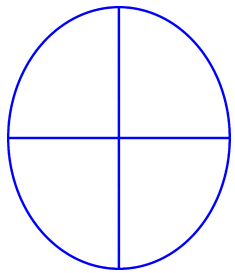


- Result lies in intersection of both spans:
  here a degenerate ellipse of dimension one

Same eigensystem, then trace dot product of eigenvectors

- **W** and **S** have the same eigensystem

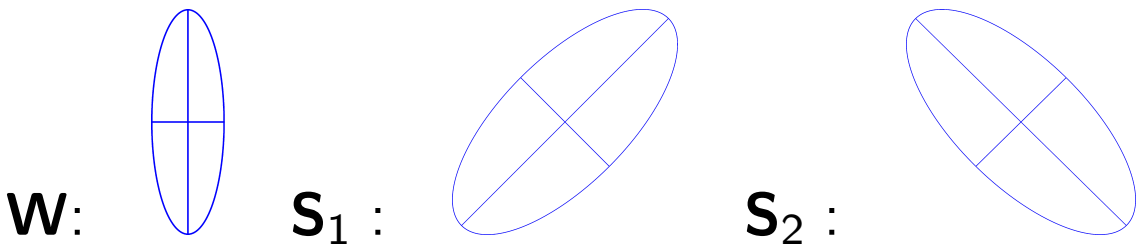




$$\mathrm{tr}(\mathbf{WS}) = \sum_i \omega_i \sigma_i$$

The ω_i can track the high σ_i

Rotated matrices are "invisible"



- $\mathbf{W}$ is any matrix with eigensystem $\mathbf{I}$
- $\mathbf{S}_1, \mathbf{S}_2$ are any matrices with rotated eigensystem and $\mathrm{tr}(\mathbf{S}_1) = \mathrm{tr}(\mathbf{S}_2)$, then

$$\mathrm{tr}(\mathbf{W}\mathbf{S}_1) = \mathrm{tr}(\mathbf{W}\mathbf{S}_2)$$

- So $\mathbf{W}$ cannot distinguish $\mathbf{S}_1$ and $\mathbf{S}_2$

Hadamard matrices

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \qquad \mathbf{H}^n = \begin{pmatrix} \mathbf{H}^{n-1} & \mathbf{H}^{n-1} \\ \mathbf{H}^{n-1} & -\mathbf{H}^{n-1} \end{pmatrix}$$

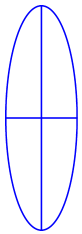
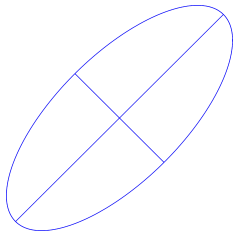
- Columns orthogonal
- $\frac{1}{\sqrt{n}}\mathbf{H}$ orthonormal - here called rotated eigensystem
- Dyads $\mathbf{h}\mathbf{h}^\top$ formed by any column $\mathbf{h}$ of $\frac{1}{\sqrt{n}}\mathbf{H}$ has $\frac{1}{n}$ along diagonal

$$\frac{1}{n}\begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix}$$

  - Diagonal of $\mathbf{h}\mathbf{h}^\top$ consists of all ones

Rotated versus unrotated in n dimensions

- If $\mathbf{W}$ has eigensystem $\mathbf{I}$
  and $\mathbf{S}$ has rotated eigensystem:

$\mathbf{W}$:  $\mathbf{S}$: 

$$\mathrm{tr}(\underbrace{\sum_i \omega_i \mathbf{e}_i \mathbf{e}_i^\top}_{\mathbf{W}} \underbrace{\sum_j \sigma_j \frac{1}{n} \mathbf{h}_j \mathbf{h}_j^\top}_{\mathbf{S}}) = \frac{1}{n} \sum_{i,j} \omega_i \sigma_j \underbrace{\mathrm{tr}(\mathbf{e}_i \mathbf{e}_i^\top \mathbf{h}_j \mathbf{h}_j^\top)}_{1} = \frac{1}{n} \mathrm{tr}(\mathbf{W}) \mathrm{tr}(\mathbf{S})$$

- A density matrix $\mathbf{W}$ with unit eigen system does not see much detail about matrices in the rotated system

Avoiding logs of zeros

Rewrite

$$\exp(\log \mathbf{A} + \log \mathbf{B}) \quad \text{as} \quad \lim_{n\to\infty} (\mathbf{A}^{1/n}\mathbf{B}^{1/n})^n$$

- Lie-Trotter Formula
- Limit always exists and well behaved
- Short hand for new product

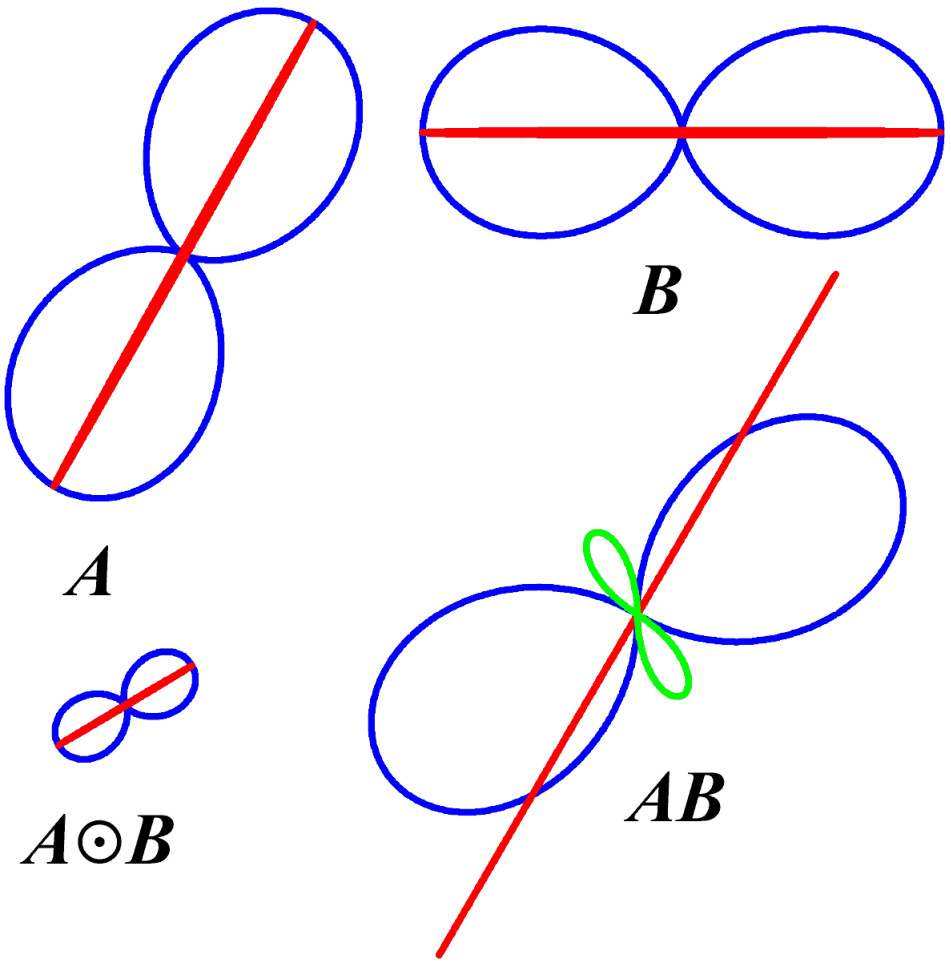
$$\mathbf{A} \odot \mathbf{B} := \lim_{n\to\infty} (\mathbf{A}^{1/n}\mathbf{B}^{1/n})^n$$

- Generalized Bayes Rule becomes

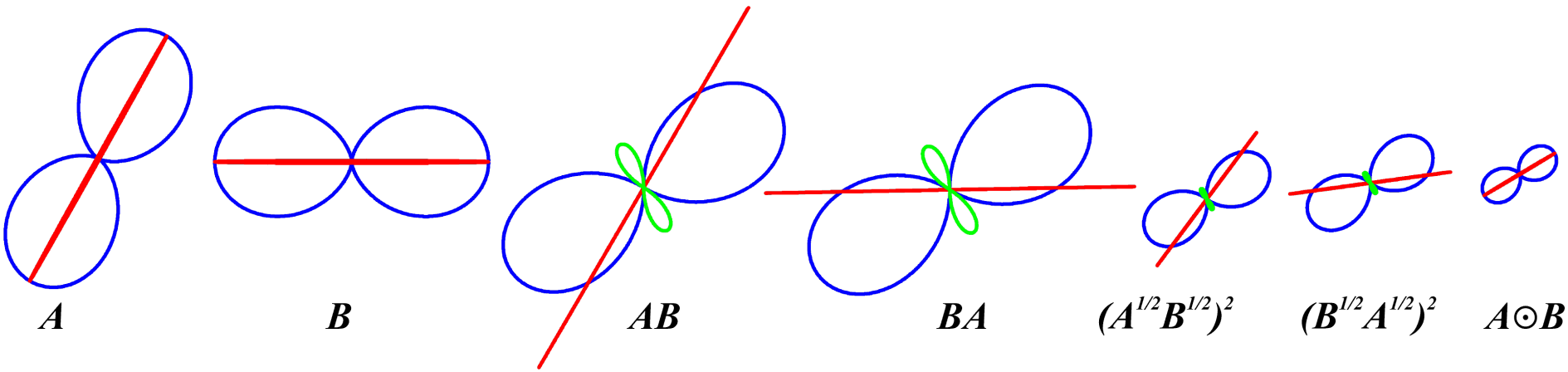
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}|\mathbb{M})}{\mathrm{tr}(\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}|\mathbb{M}))}$$

$\odot$ - commutative matrix product

Plain matrix product is non-commutative and can violate symmetry and positive definiteness. $\odot$ does not have these drawbacks.



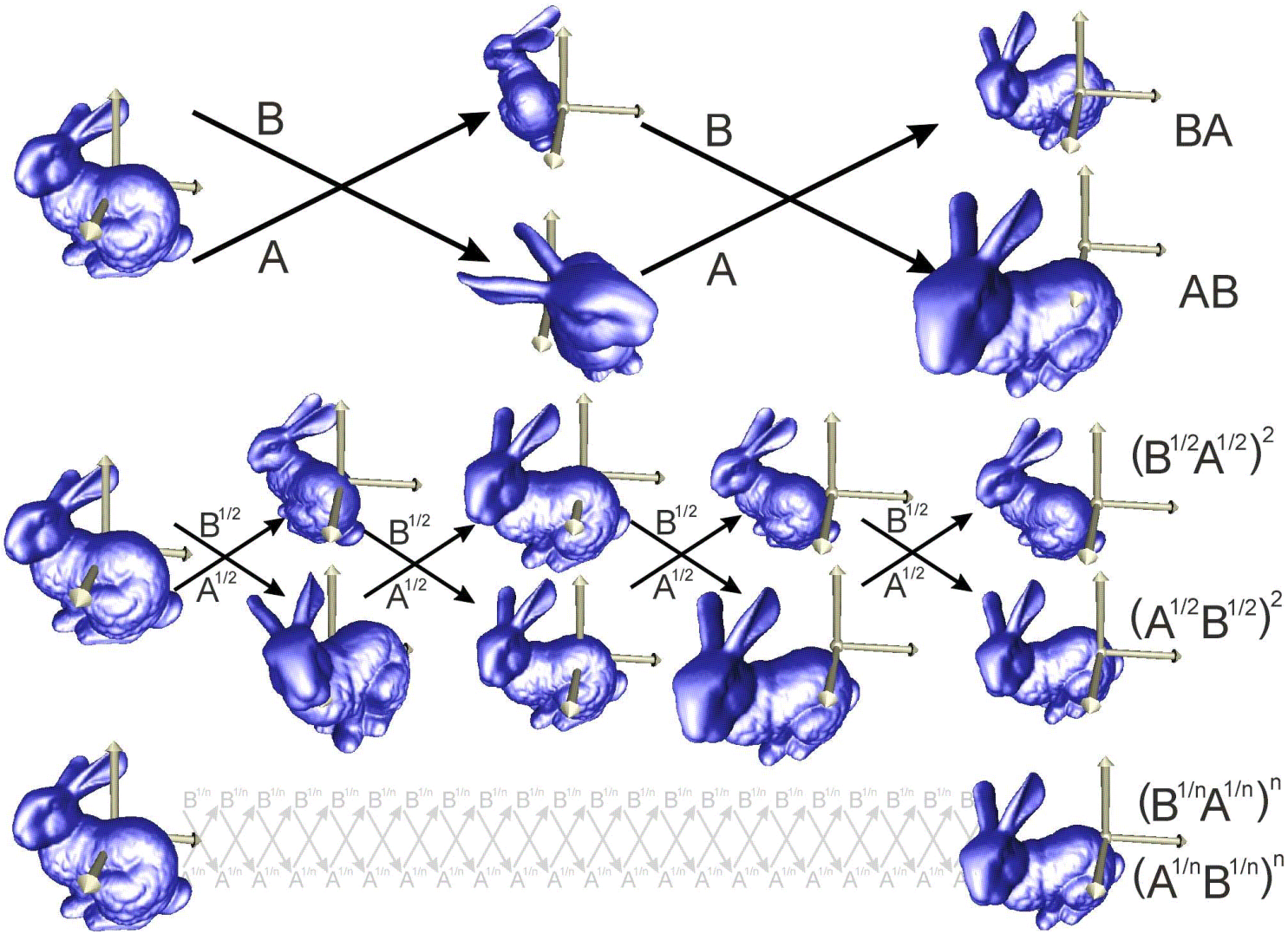
Behaviour of the limit for $\odot$



- "Ears" indicating negative definiteness are smaller for $(\mathbf{A}^{1/2}\mathbf{B}^{1/2})^2$ compared to $\mathbf{AB}$
- Non-commuting part shrinks as well

Operating on bunnies with $\odot$

Commutative combination of linear transformations



©Marc Alexa

$\odot$ Properties

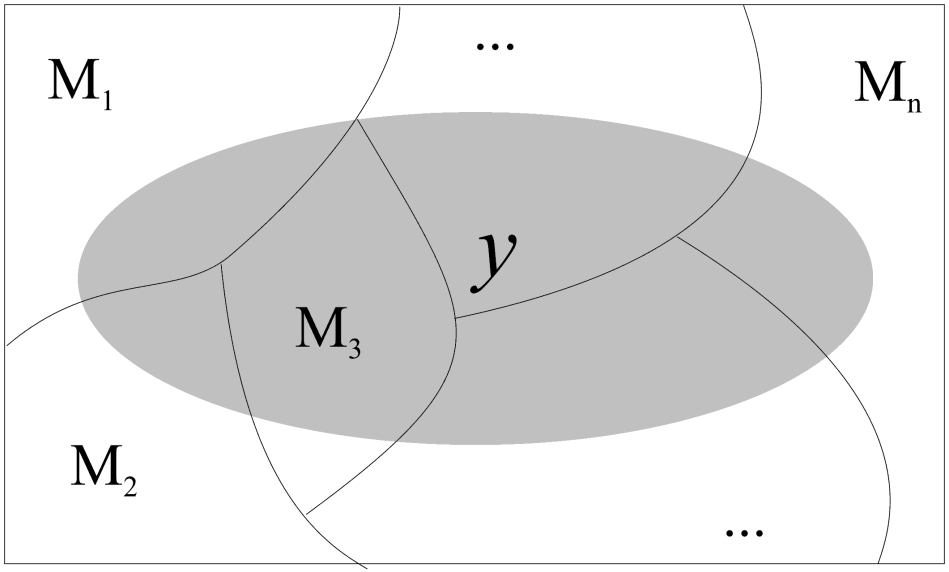
1. Commutative, associative, identity matrix as neutral elmt, preserves symmetry and positive definiteness
2. $\mathbf{A} \odot \mathbf{B} = \mathbf{AB}$ iff $\mathbf{A}$ and $\mathbf{B}$ commute
3. $\mathrm{range}(\mathbf{A} \odot \mathbf{B}) = \mathrm{range}(\mathbf{A}) \cap \mathrm{range}(\mathbf{B})$
4. $\mathrm{tr}(\mathbf{A} \odot \mathbf{B}) \leq \mathrm{tr}(\mathbf{AB})$ with equality when $\mathbf{A}$ and $\mathbf{B}$ commute
5. For any unit direction $\mathbf{u} \in \mathrm{range}(\mathbf{A})$, $\mathbf{u}\mathbf{u}^\top \odot \mathbf{A} = e^{\mathbf{u}^\top (\log^+ \mathbf{A})\mathbf{u}}\, \mathbf{u}\mathbf{u}^\top$
6. $\det(\mathbf{A} \odot \mathbf{B}) = \det(\mathbf{A}) \det(\mathbf{B})$, as for the regular matrix product
7. Typically $\mathbf{A} \odot (\mathbf{B} + \mathbf{C}) \neq \mathbf{A} \odot \mathbf{B} + \mathbf{A} \odot \mathbf{C}$

Conventional setup again

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

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

- Theorem of Total Probability



Generalized setup

$$\mathbf{D}(\mathbf{y}) = \mathrm{tr}(\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}|\mathbb{M})) \leq \mathrm{tr}(\mathbf{D}(\mathbb{M})\mathbf{D}(\mathbf{y}|\mathbb{M}))$$

$$= \underbrace{\sum_i \overbrace{\delta_i}^{\text{probability}} \overbrace{\mathbf{d}_i^\top \mathbf{D}(\mathbf{y}|\mathbb{M})\mathbf{d}_i}^{\text{variance}}}_{\text{expected variance}}$$

$$= \underbrace{\sum_i \overbrace{\mathbf{u}_i^\top \mathbf{D}(\mathbb{M})\mathbf{u}_i}^{\text{probability}} \overbrace{\lambda_i}^{\text{outcome}}}_{\text{expected measurement}}$$

where $\mathbf{D}(\mathbf{y}|\mathbb{M}) = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$

- Upper bounds similar to Theorem of Total Probability
- Only decouples when $\mathbf{D}(\mathbb{M})$ and $\mathbf{D}(\mathbf{y}|\mathbb{M})$ have same eigensystem
  - $\leq$ becomes $=$
  - Probabilities $\rightarrow P(M_i)$ and variance/outcome $\rightarrow P(y|M_i)$

Outline

1. Multiplicative updates - the genesis of this research
2. Conventional and Generalized Probability Distributions
3. Conventional and generalized Bayes rule
4. Bounds, derivation, calculus

Conventional bound ito MAP

- Probability domain

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

- Log domain

$$\begin{aligned} -\log P(y) &= -\log \sum_i P(y|M_i)P(M_i) \\ &\leq \min_i(-\log P(y|M_i) - \log P(M_i)) \end{aligned}$$

Generalized bound ito MAP

Only in the log domain

- $-\log \mathbf{m}^\top \mathbf{A}\mathbf{m} \leq -\mathbf{m}^\top \mathbf{log}(\mathbf{A})\mathbf{m}$, for any unit vector $\mathbf{m}$ and symmetric positive definite matrix $\mathbf{A}$

$$
\begin{aligned}
-\log \mathbf{D}(\mathbf{y}) = & \\
= & -\log \mathrm{tr}(\mathbf{D}(\mathbf{y}|\mathbb{M}) \odot \mathbf{D}(\mathbb{M})) \\
\leq & \min_{\mathbf{m}}(-\log \mathbf{m}^\top (\mathbf{D}(\mathbf{y}|\mathbb{M}) \odot \mathbf{D}(\mathbb{M}))\mathbf{m}) \\
\leq & \min_{\mathbf{m}}(-\mathbf{m}^\top \mathbf{log}(\mathbf{D}(\mathbf{y}|\mathbb{M}) \odot \mathbf{D}(\mathbb{M}))\mathbf{m}) \\
= & \min_{\mathbf{m}}(-\mathbf{m}^\top \mathbf{log}\, \mathbf{D}(\mathbf{y}|\mathbb{M})\mathbf{m} - \mathbf{m}^\top \mathbf{log}\, \mathbf{D}(\mathbb{M})\mathbf{m})
\end{aligned}
$$

Derivation of updates [KW97]

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

- 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 coefficients ω_i
- Prior $P(M_i)$

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

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

Minimization of $\boldsymbol{\omega}$

Lagrangian:

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

$$\frac{\partial L(\boldsymbol{\omega})}{\partial \omega_i} = \frac{1}{\eta}\left(\log\frac{\omega_i}{P(M_i)} + 1\right) - \log P(y|M_i) + \lambda$$

Setting partials zero:

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

Enforcing sum constraint:

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

$\eta = 1$: Conventional Bayes rule

Conventional Bayes again

$$\inf_{\sum_i \omega = 1} \quad \underbrace{\frac{1}{\eta} \sum_i \omega_i \log \omega_i}_{\text{neg.entropy}} - \underbrace{\frac{1}{\eta} \sum_i \omega_i \log P(M_i)}_{\text{initial data}} - \sum_i \omega_i \log P(y|M_i)$$

- Prior and data treated the same when $\eta = 1$
- Commutativity

Bayes Rule for density matrices

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

$$\inf_{\mathrm{tr}(\mathbf{W})=1} \quad \frac{1}{\eta} \underbrace{\mathrm{tr}(\mathbf{W}(\log \mathbf{W} - \log \mathbf{D}(\mathbb{M}))}_{\text{Quantum rel. entr.}} - \underbrace{\mathrm{tr}(\mathbf{W} \log \mathbf{D}(\mathbf{y}|\mathbb{M}))}_{\text{Fancier mixture loss}}$$

- $\eta = \infty$: minimized when $\mathbf{W}$ is dyad $\mathbf{u}\mathbf{u}^\top$ and $\mathbf{u}$ is the eigenvector belonging to a minimum eigenvalue of $-\log \mathbf{D}(\mathbf{y}|\mathbb{M})$
- $\eta = 1$: Generalized Bayes Rule
  - Soft maximum eigenvalue calculation
- Special case of Matrix Exponentiated Gradient update

Generalized Bayes Rule again

$$\inf_{\mathrm{tr}(\mathbf{W})=1} \quad \frac{1}{\eta}\underbrace{\mathrm{tr}(\mathbf{W}(\log \mathbf{W}))}_{\text{quantum entropy}} - \frac{1}{\eta}\underbrace{\mathrm{tr}(\mathbf{W}\log \mathbf{D}(\mathbb{M}))}_{\text{initial data}} - \mathrm{tr}(\mathbf{W}\log \mathbf{D}(\mathbf{y}|\mathbb{M}))$$

- Von Neumann Entropy is just entropy of eigenvalues
- Prior and all data treated the same when $\eta = 1$
- Commutativity

$\mathbf{D}(\mathbf{y}|\mathbb{M})$?

- Where does data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$ come from?
- From a joint distribution on space $(\mathbb{Y}, \mathbb{M})$

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: $\mathbb{A}$ and $\mathbb{B}$ of dimension $n_{\mathbb{A}}$ and $n_{\mathbb{B}}$
- Joint space: tensor product $\mathbb{A} \otimes \mathbb{B}$ - real space of dimension $n_{\mathbb{A}} n_{\mathbb{B}}$
- Elementary events are dyads of joint space
- Joint distribution is a density matrix over joint space

Joint probability?

Given joint density matrix $\mathbf{D}(\mathbb{A}, \mathbb{B})$
a dyad $\mathbf{a}\mathbf{a}^\top$ from space $\mathbb{A}$
a dyad $\mathbf{b}\mathbf{b}^\top$ from space $\mathbb{B}$
What's the joint probability of $\mathbf{a}\mathbf{a}^\top$ and $\mathbf{b}\mathbf{b}^\top$?

- $\mathbf{D}(\mathbf{a}, \mathbf{b}) =$?
- Recall $\mathbf{D}(\mathbf{a}) = \mathrm{tr}(\mathbf{D}(\mathbb{A})\, \mathbf{a}\mathbf{a}^\top)$.
- Thus $\mathbf{D}(\mathbf{a}, \mathbf{b}) = \mathrm{tr}(\mathbf{D}(\mathbb{A}, \mathbb{B})\ ?)$

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

What is a jointly specified dyad?

We will use Kronecker product

Kronecker Product

Kronecker product of $n \times m$ matrix $\mathbf{E}$ and $p \times q$ matrix $\mathbf{F}$ is a $np \times mq$ matrix $\mathbf{E} \otimes \mathbf{F}$ which in block form is given as:

$$\mathbf{E} \otimes \mathbf{F} = \begin{pmatrix} e_{11}\mathbf{F} & e_{12}\mathbf{F} & \dots & e_{1m}\mathbf{F} \\ e_{21}\mathbf{F} & e_{22}\mathbf{F} & \dots & e_{2m}\mathbf{F} \\ \dots & \dots & \dots & \dots \\ e_{n1}\mathbf{F} & e_{n2}\mathbf{F} & \dots & e_{nm}\mathbf{F} \end{pmatrix}$$

Properties:

- $(\mathbf{E} \otimes \mathbf{F})(\mathbf{G} \otimes \mathbf{H}) = \mathbf{EG} \otimes \mathbf{FH}$
- $\mathrm{tr}(\mathbf{E} \otimes \mathbf{F}) = \mathrm{tr}(\mathbf{E})\mathrm{tr}(\mathbf{F})$
- If $\mathbf{D}(\mathbb{A})$ and $\mathbf{D}(\mathbb{B})$ are density matrices, then so is $\mathbf{D}(\mathbb{A}) \otimes \mathbf{D}(\mathbb{B})$
  - $(\mathbf{a} \otimes \mathbf{b})(\mathbf{a} \otimes \mathbf{b})^\top = \mathbf{a}\mathbf{a}^\top \otimes \mathbf{b}\mathbf{b}^\top$ is a dyad of space $(\mathbb{A}, \mathbb{B})$

Joint probability

Use $\mathbf{a}\mathbf{a}^\top \otimes \mathbf{b}\mathbf{b}^\top$ as *jointly specified dyad*

Joint probability: $\mathbf{D}(\mathbf{a}, \mathbf{b}) = \mathrm{tr}(\mathbf{D}(\mathbb{A}, \mathbb{B})(\mathbf{a}\mathbf{a}^\top \otimes \mathbf{b}\mathbf{b}^\top))$

Not every dyad on the joint space can be written as $\mathbf{a}\mathbf{a}^\top \otimes \mathbf{b}\mathbf{b}^\top$!!!

This issue in quantum physics is known as *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

- $\mathbf{D}(\mathbb{A}) = \mathrm{tr}_{\mathbb{B}}(\mathbf{D}(\mathbb{A}, \mathbb{B}))$
  $\mathbf{D}(\mathbb{A}, \mathbf{b}) = \mathrm{tr}_{\mathbb{B}}(\mathbf{D}(\mathbb{A}, \mathbb{B})(\mathbf{I}_{\mathbb{A}} \otimes \mathbf{b}\mathbf{b}^{\top}))$
  Marginalization
- $\mathbf{D}(\mathbb{A}|\mathbb{B}) = \mathbf{D}(\mathbb{A}, \mathbb{B}) \odot (\mathbf{I}_{\mathbb{A}} \otimes \mathbf{D}(\mathbb{B}))^{-1}$
  Conditional in terms of the joint
  Introduced by Cerf and Adami
- $\mathbf{D}(\mathbb{A}) = \mathrm{tr}_{\mathbb{B}}(\mathbf{D}(\mathbb{A}|\mathbb{B}) \odot (\mathbf{I}_{\mathbb{A}} \otimes \mathbf{D}(\mathbb{B})))$
  Theorem of total probability
- $\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}|\mathbb{M})}{\mathrm{tr}(\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}|\mathbb{M}))}$
  Our Bayes rule
- $\mathbf{D}(\mathbf{b}|\mathbb{A}) = \mathbf{D}(\mathbf{b})\mathbf{D}(\mathbb{A}|\mathbf{b}) \odot (\mathbf{D}(\mathbb{A}|\mathbb{B}) \odot (\mathbf{I}_{\mathbb{A}} \otimes \mathbf{D}(\mathbb{B})))^{-1}$
  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 :-)
- Applications that need the calculus
- Connections to quantum computation

Open problem 1

- You can implement the conventional Bayes Rule in the tube - in vitro selection with 10^{20} variable

Open problem 1

- You can implement the conventional Bayes Rule in the tube - in vitro selection with 10^{20} variable

- Recall Generalized Bayes Rule?

$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\mathbf{exp}\left(\mathbf{log}\,\mathbf{D}(\mathbb{M}) + \mathbf{log}\,\mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$

Open problem 1

- You can implement the conventional Bayes Rule in the tube - in vitro selection with 10^{20} variable

- Recall Generalized Bayes Rule?

$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\mathbf{exp}\left(\mathbf{log}\,\mathbf{D}(\mathbb{M}) + \mathbf{log}\,\mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$

- Is there a physical realization?

Open problem 2

- The Generalized Bayes Rule retains the Conventional Bayes Rule as the hardest case

- Learning the eigensystem for free

Open problem 2

- The Generalized Bayes Rule retains the Conventional Bayes Rule as the hardest case
- Learning the eigensystem for free
- Find other cases where there is a "free matrix lunch"