Matrix Nearness Problems using Bregman Divergences

Inderjit S. Dhillon

The University of Texas at Austin

Householder Symposium XVI

Seven Springs, PA

May 24, 2005

joint work with A.Banerjee, H.Cho, J.Ghosh, Y.Guan, S.Merugu, D.Modha, S.Sra and J.Tropp

Bregman Divergences

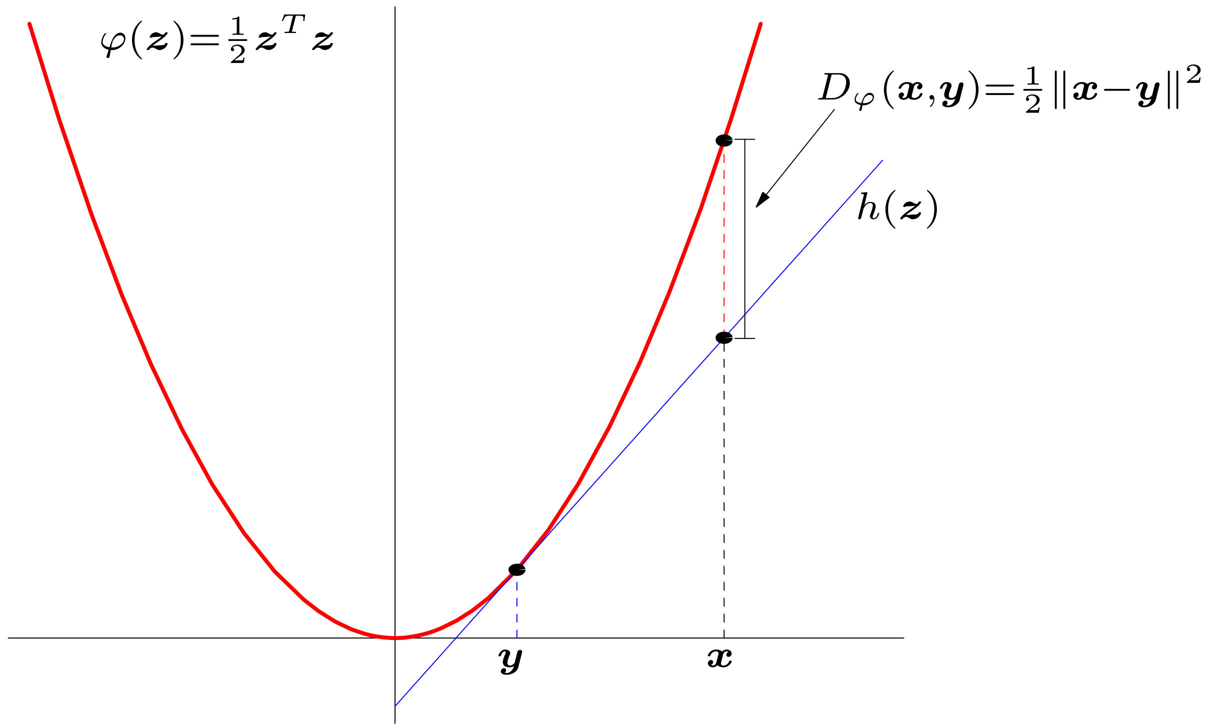
- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of “Legendre type” ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{int}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^T \nabla\varphi(\boldsymbol{y})$$

Bregman Divergences

- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{int}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^T \nabla \varphi(\boldsymbol{y})$$

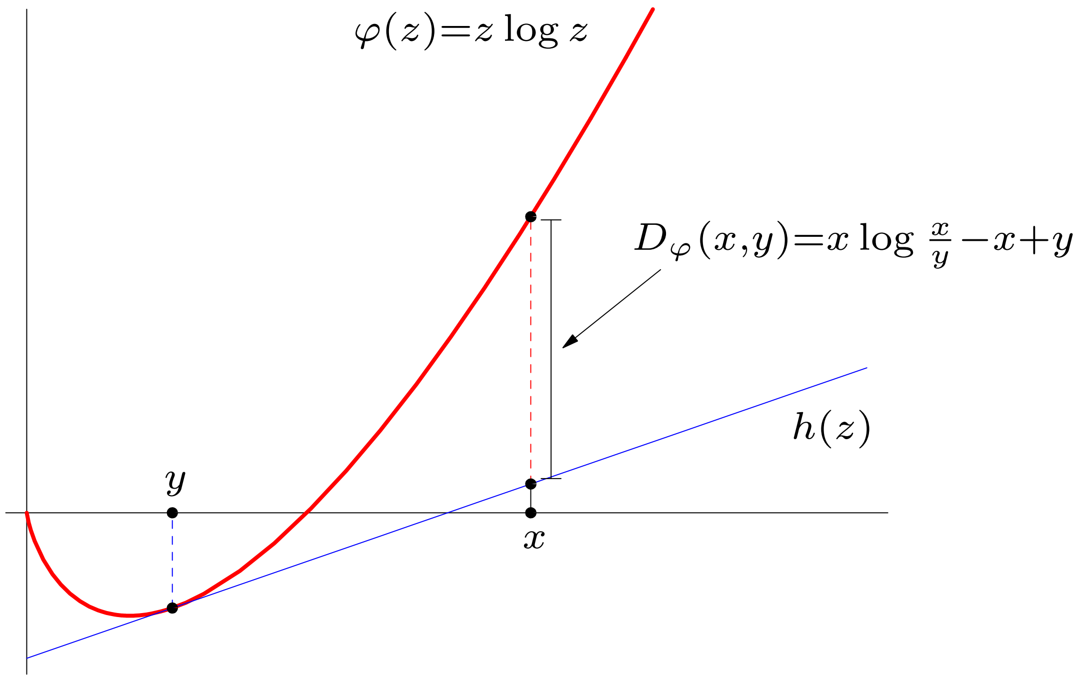


Squared Euclidean distance is a Bregman divergence

Bregman Divergences

- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{int}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^T \nabla\varphi(\boldsymbol{y})$$

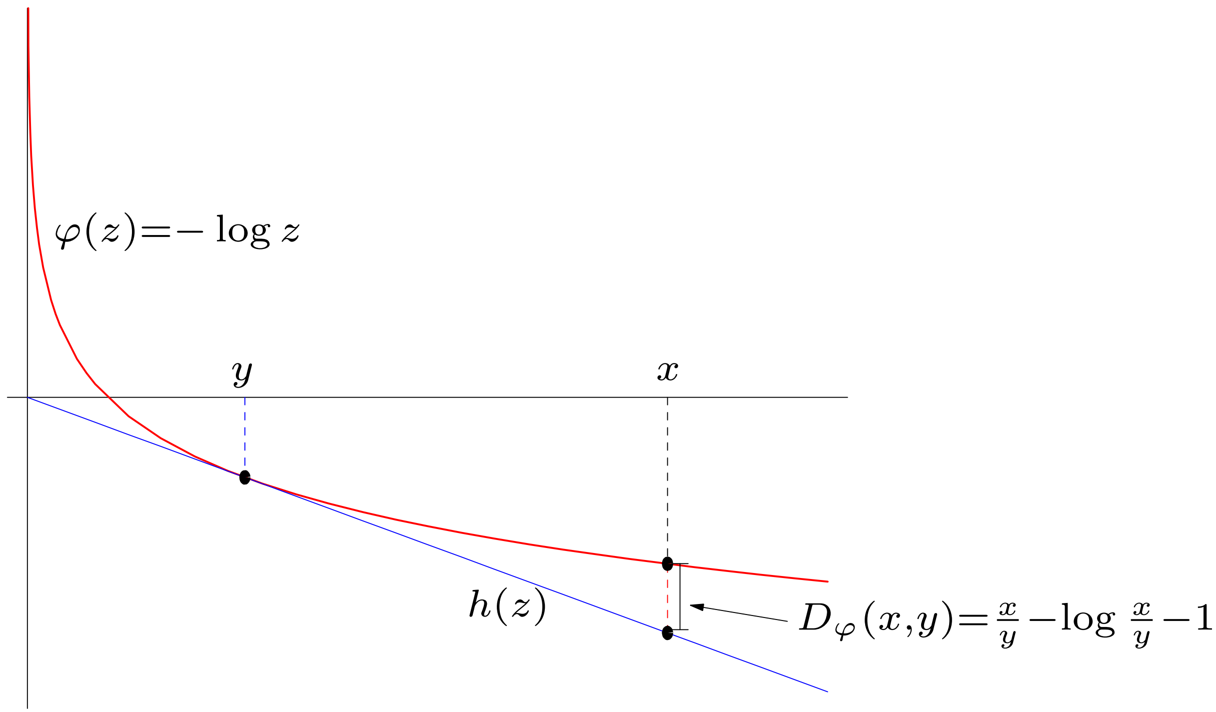


Relative Entropy (also called KL-divergence) is another Bregman divergence

Bregman Divergences

- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{int}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^T \nabla\varphi(\boldsymbol{y})$$



Itakura-Saito Distance (used in signal processing) is another Bregman divergence

Bregman Divergences

Function Name	$\varphi(x)$	$\mathrm{dom}\,\varphi$	$D_\varphi(x,y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x - x$	$[0,+\infty)$	$x\log\frac{x}{y} - x + y$
Bit entropy	$x\log x + (1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y} - \log\frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \quad (0<p<1)$	$[0,+\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p \quad (1<p<\infty)$	$(-\infty,+\infty)$	$\lvert x\rvert^p - p\,x\,\mathrm{sgn}\,y\lvert y\rvert^{p-1} + (p-1)\lvert y\rvert^p$
Exponential	e^x	$(-\infty,+\infty)$	$e^x - (x-y+1)e^y$
Inverse	$1/x$	$(0,+\infty)$	$1/x + x/y^2 - 2/y$

Properties of Bregman Divergences

- $D_\varphi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, and equals 0 iff $\boldsymbol{x} = \boldsymbol{y}$

Properties of Bregman Divergences

- $D_\varphi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, and equals 0 iff $\boldsymbol{x} = \boldsymbol{y}$
- Not a metric (symmetry, triangle inequality do not hold)

Properties of Bregman Divergences

- $D_\varphi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, and equals 0 iff $\boldsymbol{x} = \boldsymbol{y}$
- Not a metric (symmetry, triangle inequality do not hold)
- Strictly convex in the first argument, but not convex (in general) in the second argument

Properties of Bregman Divergences

- $D_\varphi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, and equals 0 iff $\boldsymbol{x} = \boldsymbol{y}$
- Not a metric (symmetry, triangle inequality do not hold)
- Strictly convex in the first argument, but not convex (in general) in the second argument
- Three-point property generalizes the “Law of cosines”:

$$D_\varphi(\boldsymbol{x}, \boldsymbol{y}) = D_\varphi(\boldsymbol{x}, \boldsymbol{z}) + D_\varphi(\boldsymbol{z}, \boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{z})^T (\nabla\varphi(\boldsymbol{y}) - \nabla\varphi(\boldsymbol{z}))$$

Bregman Projections

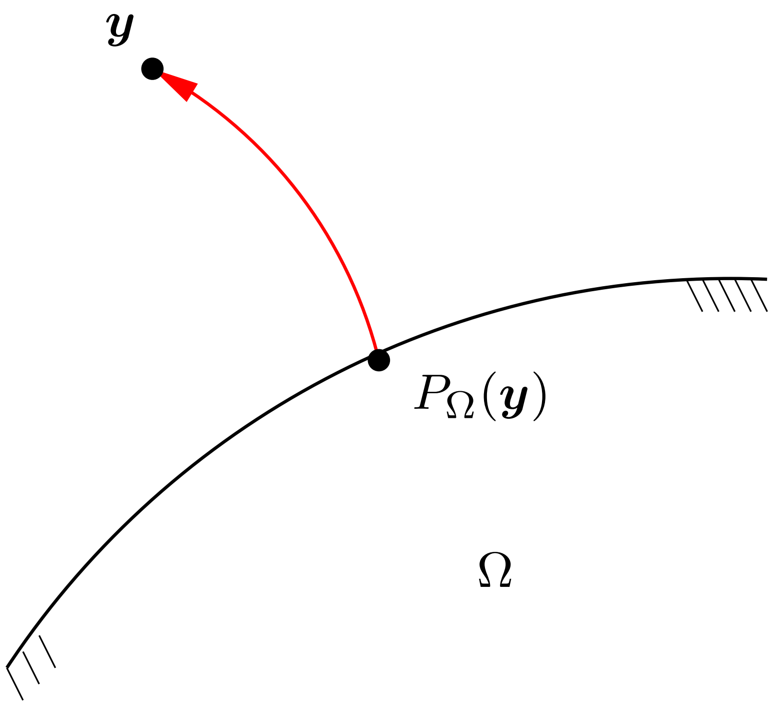
- Nearness in Bregman divergence: the “Bregman” projection of $\boldsymbol{y}$ onto a convex set Ω,

$$P_\Omega(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{\omega} \in \Omega} D_\varphi(\boldsymbol{\omega}, \boldsymbol{y})$$

Bregman Projections

- Nearness in Bregman divergence: the "Bregman" projection of $\boldsymbol{y}$ onto a convex set Ω,

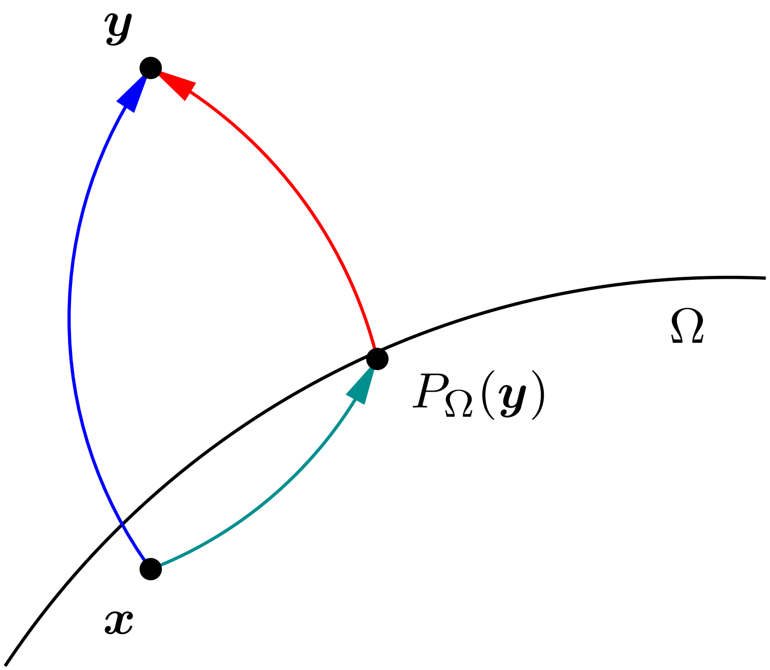
$$P_\Omega(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{\omega} \in \Omega} D_\varphi(\boldsymbol{\omega}, \boldsymbol{y})$$



Bregman Projections

- Nearness in Bregman divergence: the "Bregman" projection of $\boldsymbol{y}$ onto a convex set Ω,

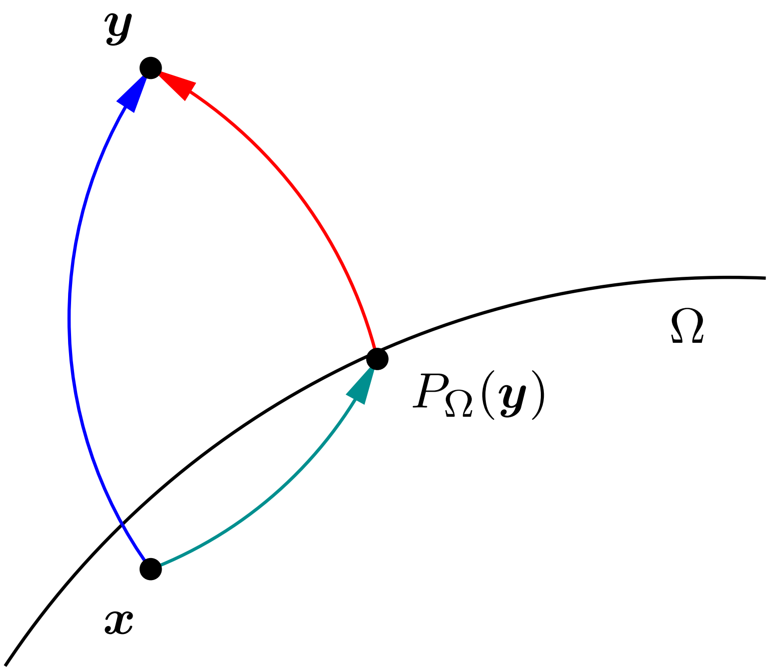
$$P_\Omega(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{\omega}\in\Omega} D_\varphi(\boldsymbol{\omega}, \boldsymbol{y})$$



Bregman Projections

- Nearness in Bregman divergence: the "Bregman" projection of $\boldsymbol{y}$ onto a convex set Ω,

$$P_\Omega(\boldsymbol{y}) = \underset{\boldsymbol{\omega}\in\Omega}{\operatorname{argmin}}\, D_\varphi(\boldsymbol{\omega}, \boldsymbol{y})$$



- Generalized Pythagoras Theorem:

$$D_\varphi(\boldsymbol{x}, \boldsymbol{y}) \geq D_\varphi(\boldsymbol{x}, P_\Omega(\boldsymbol{y})) + D_\varphi(P_\Omega(\boldsymbol{y}), \boldsymbol{y})$$

When Ω is an affine set, the above holds with equality

Historical References

- L. M. Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." *USSR Computational Mathematics and Physics*, 7:200-217, 1967.
  - Problem:

$$\min \varphi(\boldsymbol{x}) \qquad \text{subject to} \qquad \boldsymbol{a}_i^T \boldsymbol{x} = b_i, \;\; i = 0, \ldots, m-1$$

  - Bregman's cyclic projection method:
    1. Start with appropriate $\boldsymbol{x}^{(0)}$. Compute $\boldsymbol{x}^{(t+1)}$ to be the Bregman projection of $\boldsymbol{x}^{(t)}$ onto the i-th hyperplane ($i = t \mod m$) for $t = 0, 1, 2, \ldots$
  - Converges to globally optimal solution. This cyclic projection method can be extended to halfspace and convex constraints, where each projection is followed by a correction.

Question: What role can Bregman Divergences play in data analysis?

Exponential Families of Distributions

- Definition. A regular exponential family is a family of probability distributions on $\mathbb{R}^d$ with density function parameterized by $\boldsymbol{\theta}$:

$$p_\psi(\boldsymbol{x} \mid \boldsymbol{\theta}) = \exp\{\boldsymbol{x}^T\boldsymbol{\theta} - \psi(\boldsymbol{\theta}) - g_\psi(\boldsymbol{x})\}$$

  ψ is the so-called *cumulant function*, and is a convex function of Legendre type

Exponential Families of Distributions

- Definition. A regular exponential family is a family of probability distributions on $\mathbb{R}^d$ with density function parameterized by $\boldsymbol{\theta}$:

$$p_\psi(\boldsymbol{x} \mid \boldsymbol{\theta}) = \exp\{\boldsymbol{x}^T\boldsymbol{\theta} - \psi(\boldsymbol{\theta}) - g_\psi(\boldsymbol{x})\}$$

  ψ is the so-called *cumulant function*, and is a convex function of Legendre type

- Example — consider spherical Gaussians parameterized by mean $\boldsymbol{\mu}$ (with fixed variance σ):

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right\} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{\boldsymbol{x}^T\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right) - \frac{\sigma^2}{2}\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right)^2 - \frac{\boldsymbol{x}^T\boldsymbol{x}}{2\sigma^2}\right\} \end{aligned}$$

  Thus $\boldsymbol{\theta} = \dfrac{\boldsymbol{\mu}}{\sigma^2}$, and $\psi(\boldsymbol{\theta}) = \dfrac{\sigma^2}{2}\boldsymbol{\theta}^2$

Exponential Families of Distributions

- Definition. A regular exponential family is a family of probability distributions on $\mathbb{R}^d$ with density function parameterized by $\boldsymbol{\theta}$:

$$p_\psi(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \exp\{\boldsymbol{x}^T\boldsymbol{\theta} - \psi(\boldsymbol{\theta}) - g_\psi(\boldsymbol{x})\}$$

  ψ is the so-called *cumulant function*, and is a convex function of Legendre type

- Example — consider spherical Gaussians parameterized by mean $\boldsymbol{\mu}$ (with fixed variance σ):

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right\} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{\boldsymbol{x}^T\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right) - \frac{\sigma^2}{2}\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right)^2 - \frac{\boldsymbol{x}^T\boldsymbol{x}}{2\sigma^2}\right\} \end{aligned}$$

  Thus $\boldsymbol{\theta} = \dfrac{\boldsymbol{\mu}}{\sigma^2}$, and $\psi(\boldsymbol{\theta}) = \dfrac{\sigma^2}{2}\boldsymbol{\theta}^2$

- Note: Gaussian distribution $\longleftrightarrow$ Squared Loss

Example

- Poisson Distribution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{Z}_+$$

- The Poisson Distribution is a member of the exponential family
- Is there a Divergence associated with the Poisson Distribution?

Example

- Poisson Distribution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{Z}_+$$

- The Poisson Distribution is a member of the exponential family
- Is there a Divergence associated with the Poisson Distribution?
- YES — $p(x)$ can be written as

$$p(x) \quad = \quad \exp\{-D_\varphi(x,\mu) - g_\varphi(x)\},$$

where D_φ is the Relative Entropy, i.e., $D_\varphi(x,\mu) = x \log\left(\frac{x}{\mu}\right) - x + \mu$

Example

- Poisson Distribution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{Z}_+$$

- The Poisson Distribution is a member of the exponential family
- Is there a Divergence associated with the Poisson Distribution?
- YES — $p(x)$ can be written as

$$p(x) \quad = \quad \exp\{-D_\varphi(x, \mu) - g_\varphi(x)\},$$

where D_φ is the Relative Entropy, i.e., $D_\varphi(x, \mu) = x \log\left(\frac{x}{\mu}\right) - x + \mu$

- Implication: Poisson distribution $\longleftrightarrow$ Relative Entropy

Bregman Divergences and the Exponential Family

Theorem 1 *Suppose that φ and ψ are conjugate Legendre functions. Let D_φ be the Bregman divergence associated with φ, and let $p_\psi(\,\cdot\,|\,\boldsymbol{\theta})$ be a member of the regular exponential family with cumulant function ψ. Then*

$$p_\psi(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \exp\{-D_\varphi(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{\theta})) - g_\varphi(\boldsymbol{x})\},$$

where g_φ is a function uniquely determined by φ.

- Thus there is unique Bregman divergence associated with every member of the exponential family
- Implication: Member of Exponential Family $\longleftrightarrow$ unique Bregman Divergence.

[Banerjee, Merugu, Dhillon, Ghosh, 2005] — "Clustering with Bregman Divergences", *Journal of Machine Learning Research*.

Some Matrix Nearness Problems in Data Analysis

- Diagonal Scaling to Doubly Stochastic Form
  - Kruithof(1937), Sinkhorn(1964), Parlett & Landis(1982)
  - The iterative scaling procedure (of Sinkhorn) can be shown to compute the doubly stochastic matrix nearest in relative entropy

Some Matrix Nearness Problems in Data Analysis

- Diagonal Scaling to Doubly Stochastic Form
  - Kruithof(1937), Sinkhorn(1964), Parlett & Landis(1982)
  - The iterative scaling procedure (of Sinkhorn) can be shown to compute the doubly stochastic matrix nearest in relative entropy
- Clustering
  - Partition the columns of a data matrix, so that “similar” columns are in the same partition

Some Matrix Nearness Problems in Data Analysis

- Diagonal Scaling to Doubly Stochastic Form
  - Kruithof(1937), Sinkhorn(1964), Parlett & Landis(1982)
  - The iterative scaling procedure (of Sinkhorn) can be shown to compute the doubly stochastic matrix nearest in relative entropy
- Clustering
  - Partition the columns of a data matrix, so that “similar” columns are in the same partition
- Co-clustering
  - Simultaneously partition both the rows and columns of a data matrix

Some Matrix Nearness Problems in Data Analysis

- Diagonal Scaling to Doubly Stochastic Form
  - Kruithof(1937), Sinkhorn(1964), Parlett & Landis(1982)
  - The iterative scaling procedure (of Sinkhorn) can be shown to compute the doubly stochastic matrix nearest in relative entropy
- Clustering
  - Partition the columns of a data matrix, so that “similar” columns are in the same partition
- Co-clustering
  - Simultaneously partition both the rows and columns of a data matrix
- Low-Rank Matrix Approximation
  - Non-negative matrix factorization: Lee & Seung (2001)

Some Matrix Nearness Problems in Data Analysis

- Diagonal Scaling to Doubly Stochastic Form
  - Kruithof(1937), Sinkhorn(1964), Parlett & Landis(1982)
  - The iterative scaling procedure (of Sinkhorn) can be shown to compute the doubly stochastic matrix nearest in relative entropy
- Clustering
  - Partition the columns of a data matrix, so that “similar” columns are in the same partition
- Co-clustering
  - Simultaneously partition both the rows and columns of a data matrix
- Low-Rank Matrix Approximation
  - Non-negative matrix factorization: Lee & Seung (2001)
- Metric Nearness Problem
  - Given a matrix of “distances”, find the “nearest” matrix of distances such that all distances satisfy the triangle inequality
  - Dhillon, Sra & Tropp (2004)

Some Matrix Nearness Problems in Data Analysis

- Diagonal Scaling to Doubly Stochastic Form
  - Kruithof(1937), Sinkhorn(1964), Parlett & Landis(1982)
  - The iterative scaling procedure (of Sinkhorn) can be shown to compute the doubly stochastic matrix nearest in relative entropy
- Clustering
  - Partition the columns of a data matrix, so that “similar” columns are in the same partition
- Co-clustering
  - Simultaneously partition both the rows and columns of a data matrix
- Low-Rank Matrix Approximation
  - Non-negative matrix factorization: Lee & Seung (2001)
- Metric Nearness Problem
  - Given a matrix of “distances”, find the “nearest” matrix of distances such that all distances satisfy the triangle inequality
  - Dhillon, Sra & Tropp (2004)

Clustering with Bregman Divergences

- Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ be data vectors to be divided into k disjoint partitions $\gamma_1, \ldots, \gamma_k$
- The objective function for Bregman clustering

$$\min_{\gamma_1,\ldots,\gamma_k} \sum_{h=1}^{k} \sum_{\boldsymbol{a}_i \in \gamma_h} D_\varphi(\boldsymbol{a}_i, \boldsymbol{\mu}_h),$$

  where $\boldsymbol{\mu}_h$ is the representative of the h-th partition
- Lemma. Arithmetic mean is the optimal representative for all Bregman divergences, i.e.,

$$\boldsymbol{\mu}_h \equiv \frac{1}{|\gamma_h|} \sum_{\boldsymbol{a}_i \in \gamma_h} \boldsymbol{a}_i \;=\; \operatorname*{argmin}_{\boldsymbol{x}} \sum_{\boldsymbol{a}_i \in \gamma_h} D_\varphi(\boldsymbol{a}_i, \boldsymbol{x})$$

  - generalizes another property of squared Euclidean distance
- Algorithm: `KMeans`-type iterative re-partitioning algorithm decreases objective function at every iteration and converges to a local minimum (finding the globally optimal solution is NP-hard)

Co-clustering

- Co-clustering: Given a data matrix, partition the rows as well as columns

Co-clustering

- Co-clustering: Given a data matrix, partition the rows as well as columns

Original Matrix

$$
\begin{bmatrix}
z & x & z & - & - & x \\
+ & \circ & + & * & * & \circ \\
z & x & z & - & - & x \\
+ & \circ & + & * & * & \circ \\
+ & \circ & + & * & * & \circ
\end{bmatrix}
$$

Co-clustering

- Co-clustering: Given a data matrix, partition the rows as well as columns

Original Matrix

z	x	z	−	−	x
+	○	+	∗	∗	○
z	x	z	−	−	x
+	○	+	∗	∗	○
+	○	+	∗	∗	○

After co-clustering and permutation

x	x	−	−	z	z
x	x	−	−	z	z
○	○	∗	∗	+	+
○	○	∗	∗	+	+
○	○	∗	∗	+	+

Co-clustering & Matrix Approximation

- Co-clustering: Given a data matrix, partition the rows as well as columns
- Matrix approximation: Given a matrix, find an approximation determined by fewer parameters
- Can a co-clustering be associated with a matrix approximation?

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice

Knows input matrix A

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice

Bob

Knows input matrix A

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice	Bob
Knows input matrix A	Does not know A

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice

Bob

Knows input matrix A

Does not know A

Determines a co-clustering

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice $\xrightarrow[\text{\& summary statistics}]{\text{Transmits co-clustering}}$ Bob

Knows input matrix A

Does not know A

Determines a co-clustering

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice $\xrightarrow[\text{\& summary statistics}]{\text{Transmits co-clustering}}$ Bob

Knows input matrix $\boldsymbol{A}$

Does not know $\boldsymbol{A}$

Determines a co-clustering

Reconstructs
an approximation $\hat{\boldsymbol{A}}$ given
co-clustering $\&$ summary statistics

Minimum Bregman Information

- Matrix Approximation from a co-clustering:

Alice $\xrightarrow[\text{\& summary statistics}]{\text{Transmits co-clustering}}$ Bob

Knows input matrix $\boldsymbol{A}$

Does not know $\boldsymbol{A}$

Determines a co-clustering

Reconstructs
an approximation $\hat{\boldsymbol{A}}$ given
co-clustering $\&$ summary statistics

- Key Idea: Bob will reconstruct $\hat{\boldsymbol{A}}$ using the Minimum Bregman Information principle:

$$\hat{\boldsymbol{A}} = \underset{\substack{\boldsymbol{X}\ \text{satisfies} \\ \text{summary statistics}}}{\text{argmin}} \sum_{i=1}^{m} \sum_{j=1}^{n} D_{\varphi}(X_{ij}, \mu_{\boldsymbol{A}})$$

  - generalizes the maximum entropy approach

Example — Minimum Bregman Information (MBI)

Example — Minimum Bregman Information (MBI)

Original Matrix

$$\begin{bmatrix} 0 & 0 & 1 & 2 & 10 & 27 \\ 0 & 0 & 1 & 2 & 20 & 55 \\ 1 & 2 & 10 & 22 & 55 & 160 \\ 4 & 8 & 41 & 84 & 506 & 1720 \\ 1 & 2 & 10 & 20 & 56 & 180 \end{bmatrix}$$

Example — Minimum Bregman Information (MBI)

Original Matrix

$$
\left[\begin{array}{cc|cc|cc}
0 & 0 & 1 & 2 & 10 & 27 \\
0 & 0 & 1 & 2 & 20 & 55 \\
\hline
1 & 2 & 10 & 22 & 55 & 160 \\
4 & 8 & 41 & 84 & 506 & 1720 \\
1 & 2 & 10 & 20 & 56 & 180
\end{array}\right]
$$

Example — Minimum Bregman Information (MBI)

Original Matrix

0	0	1	2	10	27
0	0	1	2	20	55
1	2	10	22	55	160
4	8	41	84	506	1720
1	2	10	20	56	180

MBI matrix approximation from global mean (1 summary statistic)

100	100	100	100	100	100
100	100	100	100	100	100
100	100	100	100	100	100
100	100	100	100	100	100
100	100	100	100	100	100

Example — Minimum Bregman Information (MBI)

Original Matrix

0	0	1	2	10	27
0	0	1	2	20	55
1	2	10	22	55	160
4	8	41	84	506	1720
1	2	10	20	56	180

MBI matrix approximation from co-cluster means (6 summary statistics)

0	0	1.5	1.5	28	28
0	0	1.5	1.5	28	28
3	3	31.17	31.17	446.17	446.17
3	3	31.17	31.17	446.17	446.17
3	3	31.17	31.17	446.17	446.17

Example — Minimum Bregman Information (MBI)

Original Matrix

0	0	1	2	10	27
0	0	1	2	20	55
1	2	10	22	55	160
4	8	41	84	506	1720
1	2	10	20	56	180

MBI matrix approximation from row, column and co-cluster Means (5+6+6)

0	0	0.66	1.37	8.81	29.16
0	0	1.29	2.67	17.17	56.86
0.52	1.04	5.3	10.93	53.87	178.35
4.92	9.84	50.05	103.28	509.18	1685.73
0.56	1.12	5.7	11.76	57.96	191.9

Co-clustering & Matrix Approximation

- Main Idea: Judge co-clustering by goodness of the matrix approximation
- Objective Function for Co-clustering:

$$\min_{(\rho,\gamma)} D_{\varphi}(\boldsymbol{A}, \hat{\boldsymbol{A}}_{(\rho,\gamma)})$$

where $\hat{\boldsymbol{A}}_{(}\rho,\gamma)$ is the MBI matrix approximation corresponding to co-clustering (ρ,γ)

  - The problem is NP-hard
  - Algorithm: Iterative method alternates between row re-partitioning and column re-partitioning
  - Monotonically decreases objective function till convergence

Co-clustering Example

Original Matrix:

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	1	2	10	27
ρ_1	0	0	1	2	20	55
ρ_2	1	2	10	22	55	160
ρ_2	4	8	41	84	506	1720
ρ_2	1	2	10	20	56	180

Co-clustering Example

Original Matrix:

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	1	2	10	27
ρ_1	0	0	1	2	20	55
ρ_2	1	2	10	22	55	160
ρ_2	4	8	41	84	506	1720
ρ_2	1	2	10	20	56	180

Relative Entropy Co-clustering

	γ_1	γ_2	γ_2	γ_3	γ_3	γ_3
ρ_1	0	0.16	0.84	1.74	8.64	28.62
ρ_2	0.16	0.31	1.64	3.38	16.82	55.69
ρ_2	0.51	1	5.25	10.83	53.92	178.5
ρ_2	4.79	9.45	49.62	102.39	509.61	1687.14
ρ_2	0.55	1.08	5.65	11.66	58.01	192.06

Co-clustering Example

Original Matrix:

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	1	2	10	27
ρ_1	0	0	1	2	20	55
ρ_2	1	2	10	22	55	160
ρ_2	4	8	41	84	506	1720
ρ_2	1	2	10	20	56	180

Relative Entropy Co-clustering

	γ_1	γ_2	γ_2	γ_3	γ_3	γ_3
ρ_1	0	0.11	0.57	1.75	8.72	28.86
ρ_1	0	0.21	1.11	3.41	17	56.27
ρ_2	0.52	1.01	5.32	10.83	53.89	178.42
ρ_2	4.92	9.58	50.28	102.35	509.4	1686.47
ρ_2	0.56	1.09	5.72	11.65	57.99	191.98

Co-clustering Example

Original Matrix:

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	1	2	10	27
ρ_1	0	0	1	2	20	55
ρ_2	1	2	10	22	55	160
ρ_2	4	8	41	84	506	1720
ρ_2	1	2	10	20	56	180

Relative Entropy Co-clustering

	γ_1	γ_1	γ_2	γ_3	γ_2	γ_2
ρ_1	0	0	0.85	1.36	8.77	29.02
ρ_1	0	0	1.66	2.64	17.1	56.6
ρ_2	0.52	1.04	5.25	10.93	53.88	178.38
ρ_2	4.92	9.84	49.59	103.31	509.28	1686.06
ρ_2	0.56	1.12	5.65	11.76	57.98	191.94

Co-clustering Example

Original Matrix:

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	1	2	10	27
ρ_1	0	0	1	2	20	55
ρ_2	1	2	10	22	55	160
ρ_2	4	8	41	84	506	1720
ρ_2	1	2	10	20	56	180

Relative Entropy Co-clustering

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	0.66	1.37	8.81	29.16
ρ_1	0	0	1.29	2.67	17.17	56.86
ρ_2	0.52	1.04	5.3	10.93	53.87	178.35
ρ_2	4.92	9.84	50.05	103.28	509.18	1685.73
ρ_2	0.56	1.12	5.7	11.76	57.96	191.9

Co-clustering Example

Original Matrix:

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	1	2	10	27
ρ_1	0	0	1	2	20	55
ρ_2	1	2	10	22	55	160
ρ_2	4	8	41	84	506	1720
ρ_2	1	2	10	20	56	180

Relative Entropy Co-clustering

	γ_1	γ_1	γ_2	γ_2	γ_3	γ_3
ρ_1	0	0	0.66	1.37	8.81	29.16
ρ_1	0	0	1.29	2.67	17.17	56.86
ρ_2	0.52	1.04	5.3	10.93	53.87	178.35
ρ_2	4.92	9.84	50.05	103.28	509.18	1685.73
ρ_2	0.56	1.12	5.7	11.76	57.96	191.9

Squared Euclidean Co-clustering

	γ_1	γ_1	γ_1	γ_1	γ_2	γ_3
ρ_1	−24.6	−23.4	−13.2	0.2	15.38	85.63
ρ_1	−18.27	−17.07	−6.87	6.53	21.71	91.96
ρ_1	10.4	11.6	21.8	35.2	50.38	120.63
ρ_2	24.9	26.1	36.3	49.7	506	1720
ρ_1	13.57	14.77	24.97	38.37	53.54	123.79

Results — Document Clustering

- Document data set with 3 known clusters
- Co-clustering with Relative Entropy
  - superior performance as compared to just column clustering
  - performs implicit dimensionality reduction at each iteration

(3 doc;20 word)			(3 doc;500 word)			(3 doc;2500 word)		
1389	1	2	1364	3	18	920	49	292
9	1455	33	5	1446	21	31	1239	404
0	4	998	29	11	994	447	172	337

Confusion matrices for a document data set with different number of word clusters

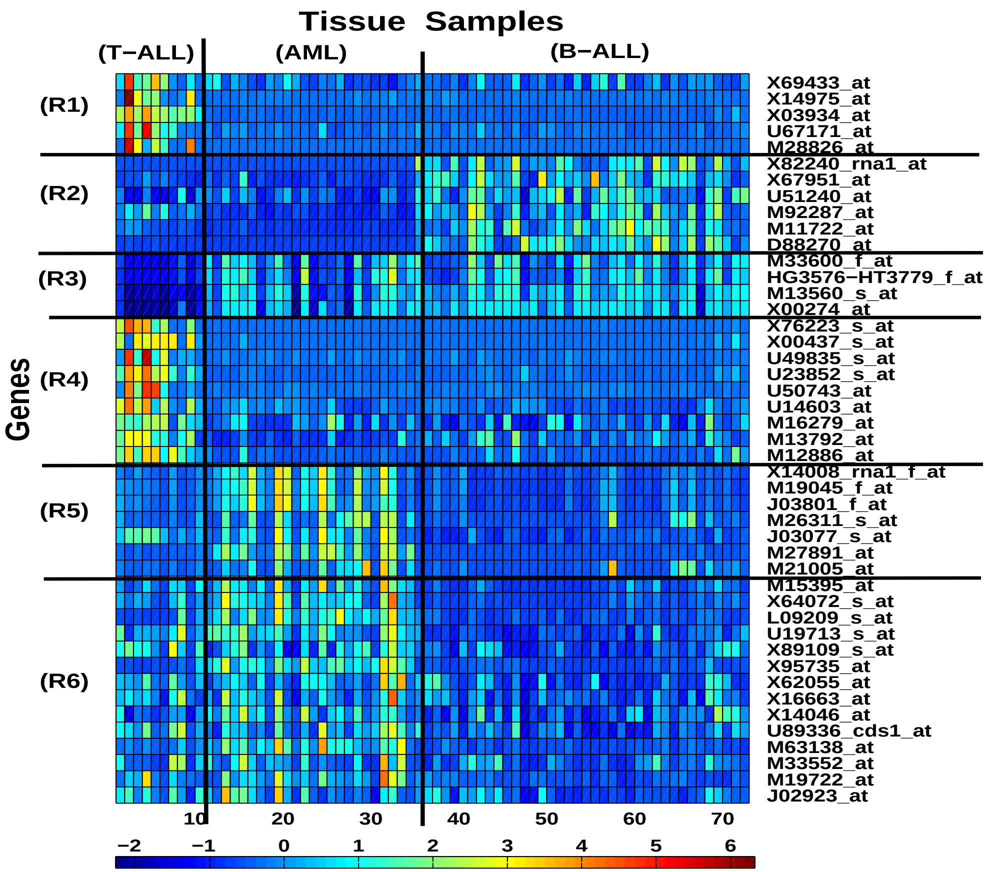
- Co-clustering with Relative Entropy — has also been applied to tasks in Natural Language Processing (Part-of-speech tagging) where rows correspond to “words” and columns to “senses” [Rowher & Freitag, 2004]

Results — Bioinformatics

- Gene Expression Leukemia Data – Matrix contains positive and negative numbers
- Squared Euclidean Distance works well

Results — Bioinformatics

- Gene Expression Leukemia Data – Matrix contains positive and negative numbers
- Squared Euclidean Distance works well
- Co-clustering is able to recover the cancer samples and functionally related genes



Matrix Divergences

- Non-separable matrix divergences obtained by applying φ to eigenvalues:
  - Let $\mathcal{H}$: space of $N \times N$ Hermitian matrices
  - Let $\boldsymbol{\lambda} : \mathcal{H} \to \mathbb{R}^N$ be the eigenvalue map

$$D_{\varphi\circ\boldsymbol{\lambda}}(\boldsymbol{A}, \boldsymbol{B}) = (\varphi\circ\boldsymbol{\lambda})(\boldsymbol{A}) - (\varphi\circ\boldsymbol{\lambda})(\boldsymbol{B}) - \langle \boldsymbol{A} - \boldsymbol{B}, \boldsymbol{U}\,\mathrm{diag}\,\{\nabla\varphi(\boldsymbol{\lambda}(\boldsymbol{A}))\}\boldsymbol{U}^*)\}\rangle$$

- Example: $\varphi(\boldsymbol{x}) = -\sum_k \log x_k$. Then $(\varphi \circ \boldsymbol{\lambda})(\boldsymbol{A}) = -\log\det \boldsymbol{A}$, and

$$D_{\varphi\circ\boldsymbol{\lambda}}(\boldsymbol{A}; \boldsymbol{B}) = \mathrm{trace}(\boldsymbol{A}\boldsymbol{B}^{-1}) - \log\det \boldsymbol{A}\boldsymbol{B}^{-1} - N$$

- Inequalities:

Hadamard: $\det \boldsymbol{A} \leq \prod_{i=1}^N a_{ii}$ for all positive definite $\boldsymbol{A}$

$$\sum_{i=1}^N \frac{A_{ii}}{\lambda_i} \geq N, \quad \text{and} \quad \sum_{i=1}^N \lambda_i(\boldsymbol{A}^{-1})_{ii} \geq N$$ for all positive definite $\boldsymbol{A}$

References

- Optimization: Bregman(1967), Censor & Zenios(1998)
- Convex Analysis: Rockafellar(1970), Bauschke & Borwein (1997)
- Exponential Families: Barndorff-Nielsen (1978)
- Data Analysis:
  - Banerjee, Merugu, Dhillon & Ghosh (2004)
  - Banerjee, Dhillon, Ghosh, Merugu & Modha (2004)
  - Dhillon, Sra & Tropp (2005)
  - Dhillon & Tropp (2005, working manuscript)

Conclusions

- Squared loss is used in many data inference problems
- When data is drawn from a member of the exponential family, the corresponding Bregman nearness problem needs to be solved
- Leads to various interesting matrix nearness problems
- Open questions:
  - How good is the matrix approximation from co-clustering?
  - Given an application, what is the appropriate divergence measure?