Structured Low Rank Approximation

Lecture IV: Singular Value Assignment with Low Rank Matrices

Moody T. Chu

North Carolina State University

presented at

XXII School of Computational Mathematics
Numerical Linear Algebra and Its Applications
September 16, 2004

Syllabus

- Objectives:
  - ◇ To provide some preliminaries.
  - ◇ To treat some mathematics.
  - ◇ To point out some applications.
  - ◇ To describe some algorithms.
- Topics:
  - ◇ Lecture I: Introduction
  - ◇ Lecture II: General Approach
  - ◇ Lecture III: Distance Geometry and Protein Structure
  - ◇ Lecture IV: Singular Value Assignment with Low Rank Matrices
  - ⧫ Lecture V: Nonnegative Matrix Factorization
- Assignments:
  - ◇ Many numerical techniques, but none is superior.
  - ◇ Proper interpretation of the factorization is needed.
  - ◇ Perhaps more constraints need to be imposed.

Lecture V

Nonnegative Matrix Factorization

Joint Work with Fasma Diele, Robert Plemmons, and Stefania Ragni

Outline

- Some Real Data
  - ◇ EPA Data on Air Pollution
  - ◇ Pixels of Irises
- Linear Model
  - ◇ Mass Balance Equation
  - ◇ Principle Component Retrieval
- First Order Optimality Condition
  - ◇ Kuhn-Tucker Condition
  - ◇ Lagrangian Multiplier
- Numerical Methods
  - ◇ Newton-type Approach
  - ◇ Reduced Quadratic Model Approach
  - ◇ Gradient Approach
- Numerical Experiments

Air Pollution Data

	1970	1975	1980	1985	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Carbon Monoxide	129444	116756	117434	117013	106438	99119	101797	99307	99790	103713	94057	101294	101459	96872	97441
Lead	221	160	74	23	5	5	4	4	4	4	4	4	4	4	4
Nitrogen Oxides	20928	22632	24384	23197	23892	24170	24338	24732	25115	25474	25052	26053	26353	26020	25393
Volatile Organic	30982	26080	26336	24428	22513	21052	21249	11862	21100	21682	20919	19464	19732	18614	18145
PM_{10}	13165	7677	7109	41397	40963	27881	27486	27249	27502	28756	25931	25690	25900	26040	23679
Sulfur Dioxide	31161	28011	25906	23658	23294	23678	23045	22814	22475	21875	19188	18859	19366	19491	18867
$PM_{2.5}$						7429	7317	7254	7654	7012	6909	7267	7065	6773	6773
Ammonia						4355	4412	4483	4553	4628	4662	4754	4851	4929	4963

Table 1: Annual pollutants estimates (in thousand short tons).

- Who should be blamed for emitting these pollutants?
- How much responsibility should each guilty party bear?

Iris Data

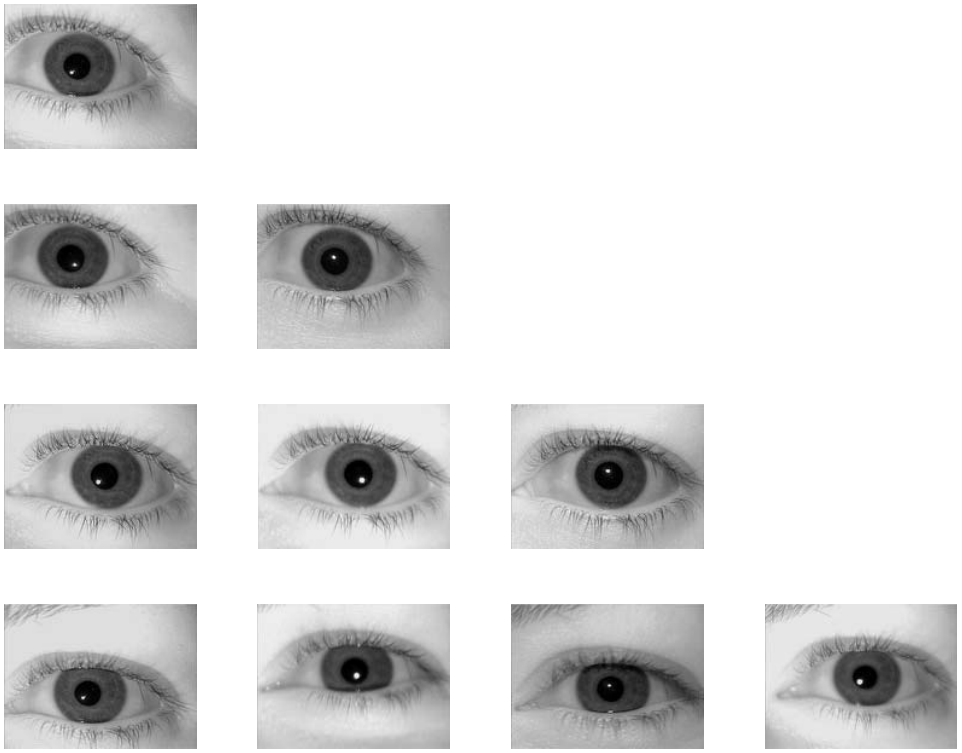

Figure 1: Intensity image of an iris

- Each iris is a 160×120 pixel matrix with entry values between 0 and 1.
- Are there any common features in these irises?
- Can any intrinsic parts that make up these poses be identifies?

Linear Model

- Let $Y = [y_{ij}] \in \mathbb{R}^{m \times n}$ denote the matrix of "observed" data.
  - $\diamond$ y_{ij} = the *score* obtained by entity j on variable i.
- Assume that y_{ij} is a linearly weighted score by entity j based on several factors.
  - $\diamond$ Temporarily assume that there are p factors, but it is precisely the point that the factors are to be retrieved in the mining process.
- Assumes the relationship

$$Y = \underbrace{A}_{m \times p} \underbrace{F}_{p \times n}.$$

  - $\diamond$ a_{ik} = *influence* of factor k on variable i.
  - $\diamond$ f_{kj} = the *response* of entity j to factor k.

Receptor Model

- An observational technique used within the air pollution research community.
- Assume that mass conservation.
  - ◇ A mass balance analysis can be used to identify and apportion sources of airborne particulate matter in the atmosphere.
- The relationships between p sources which contribute m chemical species to n samples lead to a *mass balance displaymath*,

$$y_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj}.$$

  - ◇ y_{ij} = the elemental concentration of the ith chemical measured in the jth sample.
  - ◇ a_{ik} = the gravimetric concentration of the ith chemical in the kth source.
  - ◇ f_{kj} = the airborne mass concentration that the kth source has contributed to the jth sample.

- A typical scenario,
  - ◇ Only values of y_{ij} are observable.
  - ◇ Neither the sources are known nor the compositions of the local particulate emissions are measured.
- A critical question,
  - ◇ Estimate the number p.
  - ◇ Determine the compositions a_{ik}, and the contributions f_{kj} of the sources.
  - ◇ The source compositions a_{ik} and the source contributions f_{kj} must all be nonnegative.
- Conventional tools such as principal component analysis, factor analysis, cluster analysis, and other multivariate statistical techniques cannot guarantee nonnegativity.

Image Articulation

- Image articulation libraries are made up of images showing a composite object in many articulations and poses.
- Could the factorization enable the identification and classification of intrinsic "parts" that make up the object being imaged by multiple observations?
  - ◇ Each column $\mathbf{y}_j$ of a nonnegative matrix Y represents m pixel values of one image.
  - ◇ The columns $\mathbf{a}_k$ of A are basis elements in $\mathbb{R}^m$.
  - ◇ The columns of F denote coefficient sequences representing the n images in the basis elements.

$$\mathbf{y}_j = \sum_{k=1}^{p} \mathbf{a}_k f_{kj},$$

- Nonnegativity requirement.
  - ◇ Those basic parts, being images themselves, are necessarily nonnegative.
  - ◇ The superposition coefficients, each part being present of absent, are also necessarily nonnegative.

Nonnegative Matrix Factorization

- Need to determine as fewer factors as possible and, hence, a low rank nonnegative matrix factorization of the data matrix Y.
- **(NNMF)**Given a nonnegative matrix $Y \in \mathbb{R}^{m\times n}$ and a positive integer $p < \min\{m, n\}$, find nonnegative matrices $U \in \mathbb{R}^{m\times p}$ and $V \in \mathbb{R}^{p\times n}$ so as to minimize the functional

$$f(U, V) := \frac{1}{2}\|Y - UV\|_F^2. \tag{1}$$

  - ⋄ Because $UV = (UD)(D^{-1}V)$ for any invertible matrix $D \in \mathbb{R}^{p\times p}$, it might be desirable to "normalize" columns of U.

Literature Search

- Quite a few numerical algorithms proposed in the literature for NNMF. (Hoyer'02, Lee and Seung'01, Liu and Yi'03, Donoho and Stodden'03).
- Current developments seem to lack a firm theoretical foundation in general. (Tropp'03)
- Nonnegative matrices form a cone with many facets.

Parameterization of Nonnegative Matrices

- Parameterize the cone $\mathbb{R}_+^{m\times p}$ of nonnegative matrices as

$$\mathbb{R}_+^{m\times p} = \{E. * E | E \in \mathbb{R}^{m\times p}\}.$$

  - $E. * E = [e_{ij}^2]$ denotes the Hadamard product.

- The NNMF can be expressed as the minimization of

$$g(E, F) := \frac{1}{2}\|Y - (E. * E)(F. * F)\|_F^2. \tag{2}$$

  - No constraints imposed on E and F.

Computing the Gradient

- Equip the space $\mathbb{R}^{m\times p}\times\mathbb{R}^{p\times n}$ with the product Frobenius inner product,

$$\langle (X_1,Y_1),(X_2,Y_2)\rangle = \langle X_1,X_2\rangle + \langle Y_1,Y_2\rangle.$$

- The Fréchet derivative of g can be calculated component by component.

$$\frac{\partial g}{\partial E}.H = \langle -2H, E.*\left(\delta(E,F)(F.*F)^\top\right)\rangle, \tag{3}$$

$$\frac{\partial g}{\partial F}.K = \langle -2K, F.*\left((E.*E)^\top\delta(E,F)\right)\rangle. \tag{4}$$

  - $\delta(E,F) := Y-(E.*E)(F.*F)$ denotes the residue.

- By the Riesz representation theorem, the gradient of g at (E,F) can be expressed as

$$\nabla g(E,F) = \left(-2E.*\left(\delta(E,F)(F.*F)^\top\right), -2F.*\left((E.*E)^\top\delta(E,F)\right)\right).$$

First Order Optimality Condition

- If (E, F) is a local minimizer of the objective functional g in (2), then necessarily the equations

$$E. * \left(\delta(E,F)(F. * F)^\top\right) = 0 \in \mathbb{R}^{m \times p}, \quad (5)$$
$$F. * \left((E. * E)^\top \delta(E,F)\right) = 0 \in \mathbb{R}^{p \times n}, \quad (6)$$

  are satisfied.

  - ◇ The corresponding stationary point to the nonnegative matrix factorization problem is given by $U = E. * E$ and $V = F. * F$.

- The necessary condition for $(U, V) \in \mathbb{R}_+^{m \times p} \times \mathbb{R}_+^{p \times n}$ to solve the nonnegative matrix factorization problem is

$$U. * \left((Y - UV)V^\top\right) = 0 \in \mathbb{R}^{m \times p}, \quad (7)$$
$$V. * \left(U^\top (Y - UV)\right) = 0 \in \mathbb{R}^{p \times n}. \quad (8)$$

  - ◇ Be aware of the complementarity condition.

Kuhn-Tucker Conditions

- Classical constrained optimization problem:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subect to} & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, n \\ & g_j(\mathbf{x}) \geq 0, \quad j = 1, \ldots, \ell. \end{array}$$

- Classical Kuhn-Tucker conditions: At an local minimizer $\mathbf{x}^*$, there exist values $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_\ell$ such that
  - $\nabla f(\mathbf{x}^*) - \sum_{i=1}^{n} \lambda_i \nabla h_i(\mathbf{x}^*) - \sum_{j=1}^{\ell} \mu_j \nabla g_i(\mathbf{x}^*) = 0$,
  - $\mu_j g_j(\mathbf{x}) = 0$,
  - $\mu_j \geq 0$.
- In our case,
  - The two matrices $-(Y - UV)V^\top$ and $-U^\top(Y - UV)$ are precisely the Lagrangian multipliers specified in the Kuhn-Tucker condition.
  - At a solution (U, V) of the nonnegative matrix factorization problem, it is necessary that

$$(Y - UV)V^\top \leq 0, \tag{9}$$
$$U^\top(Y - UV) \leq 0. \tag{10}$$

Numerical Methods

- All numerical methods essentially center around satisfying the first order optimality condition.
- Various additional mechanisms, such as the Hessian information or some descent properties, are built in the different schemes to ensure that a critical point is a solution to (1).
- Methods are plenty, but no absolutely superior algorithm.
  - ◇ Newton-type approach.
  - ◇ Reduced quadratic model approach.
  - ◇ Gradient approach.
- There is much room for further improvement of any of these methods for the NNMF problem.

Newton-type Approach

- Centered around implementing the Kuhn-Tucker conditions.
- There has been plenty software developed— 27 computer codes are compared in (Hock and Schittkowski'83).
- Fast convergence, but usually more expensive.

Constrained Quasi-Newton Methods

- Sequential Quadratic Programming (SQP) methods.
  - ◇ One of the most efficient techniques.
  - ◇ Solve successively a sequence of quadratic programming subproblems obtained by linearizing the original nonlinear problems at various approximate solutions.
  - ◇ Accumulate second order information via a quasi-Newton updating procedure.
  - ◇ Superlinear convergence. (Fletcher'87, Gill, Murray and Wright'81).
- In the NNMF application, the Kuhn-Tucker conditions are explicitly given by (7), (8), (9) and (10).
  - ◇ How to take advantage of the underlying structure?

ADI Newton Iteration

- Alternating direction iteration.
  - ◇ Start by fixing V in (7) and solve the system,

$$U. * [B - UC] = 0,$$

    for a nonnegative matrix $U \in \mathbb{R}_+^{m \times p}$.
    - ▷ $B = YV^\top \in \mathbb{R}^{m \times p}$.
    - ▷ $C = VV^\top \in \mathbb{R}^{p \times p}$.
  - ◇ Fix U in (8) and solve next the system

$$V. * [R - SV] = 0,$$

    for a nonnegative matrix $V \in \mathbb{R}_+^{p \times n}$.
    - ▷ $R = U^\top Y \in \mathbb{R}^{p \times n}$.
    - ▷ $S = U^\top U \in \mathbf{R}^{p \times p}$.
- Because p is low, the sizes of the square matrices C and S are relatively small.
  - ◇ Need to guarantee nonnegativity.
  - ◇ Need to satisfy (9) and (10).
  - ◇ Need to show convergence.

Inner Loop Iteration

- In each sweep of the outer loop iteration, the solution U and V could be solved row by row and column by column, respectively.
  - ◇ Each *row* of U gives rise to a nonlinear system of equations of the form

$$\mathbf{u}^\top .* [\mathbf{b}^\top - \mathbf{u}^\top C] = \mathbf{0}. \tag{11}$$

  - ◇ There are m rows for U and n columns for V to be solved, respectively,
  - ◇ Coefficient matrices $C = VV^\top$ or $S = U^\top U$ need to be updated once per sweep.
- To guarantee the nonnegativity of $\mathbf{u}^\top$, rewrite (11) as the equation

$$\psi(\mathbf{e}) = (C(\mathbf{e}. * \mathbf{e}) - \mathbf{b}). * \mathbf{e} = \mathbf{0}.$$

  - ◇ $\mathbf{e}. * \mathbf{e} = \mathbf{u}$ in nonnegative.
  - ◇ The Frèchet of ψ acting on an arbitrary vector $\mathbf{h} \in \mathbb{R}^p$ can be calculated as

$$\psi'(\mathbf{e}).\mathbf{h} = \{2\text{diag}(\mathbf{e})C\text{diag}(\mathbf{e}) + \text{diag}(C(\mathbf{e}. * \mathbf{e}) - \mathbf{b})\}\,\mathbf{h}.$$

- The inner loop algorithm.
  - ◇ Given $\mathbf{e}^{(0)}$ such that $C(\mathbf{e}^{(0)}. * \mathbf{e}^{(0)}) - \mathbf{b} \geq 0$, do until convergence:
    1. Compute $\mathbf{r}^{(k)} = C(\mathbf{e}^{(k)}. * \mathbf{e}^{(k)}) - \mathbf{b}$.
    2. Solve for $\mathbf{h}$ from the linear system
$$\left\{2\text{diag}(\mathbf{e}^{(k)})C\text{diag}(\mathbf{e}^{(k)}) + \text{diag}(\mathbf{r}^{(k)})\right\}\mathbf{h} = -\mathbf{r}^{(k)}. * \mathbf{e}^{(k)};$$
    3. Update $\mathbf{e}^{(k+1)} = \mathbf{e}^{(k)} + \alpha^{(k)}\mathbf{h}$.

Projected Newton Method

- The objective function (1) is separable.

$$f(U,V) = \frac{1}{2}\|Y - UV\|_F^2 = \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{y}_i - U\mathbf{v}_i\|_2^2.$$

  - ◇ For a fixed $U \in \mathbb{R}^{m\times p}$, each single column of V is a nonnegative least squares problem

$$\begin{aligned} \text{minimize} \quad & \phi(\mathbf{v}) := \frac{1}{2}\|\mathbf{y} - U\mathbf{v}\|_2^2, \\ \text{subject to} \quad & \mathbf{v} \geq \mathbf{0} \in \mathbb{R}^p. \end{aligned} \tag{12}$$

  - ◇ The MATLAB routine LSQNONNEG is readily available.
  - ◇ Use a notion of the projected Newton method. (Lawson and Hansen'74)

- Alternate between U and V

  - ◇ Employ the projected Newton method or the existing LSQNONNEG for each column of V and each row of U.

Reduced Quadratic Model Approach

- The reduced quadratic model approach is considerably simple to use.
- Replace the quadratic function $\phi(\mathbf{v})$ defined in (12) by a sequence of simpler quadratic functions.
  - ◇ Near any given $\mathbf{v}^c$, rewrite $\phi(\mathbf{v})$ as
$$\phi(\mathbf{v}) = \phi(\mathbf{v}^c) + (\mathbf{v} - \mathbf{v}^c)^\top \nabla\phi(\mathbf{v}^c) + \frac{1}{2}(\mathbf{v} - \mathbf{v}^c)^\top U^\top U(\mathbf{v} - \mathbf{v}^c).$$
  - ◇ Approximate $\phi(\mathbf{v})$ by a *simpler* quadratic model
$$\varphi(\mathbf{v}; \mathbf{v}^c) = \phi(\mathbf{v}^c) + (\mathbf{v} - \mathbf{v}^c)^\top \nabla\phi(\mathbf{v}^c) + \frac{1}{2}(\mathbf{v} - \mathbf{v}^c)^\top D(\mathbf{v}^c)(\mathbf{v} - \mathbf{v}^c), \tag{13}$$
    - ▷ $D(\mathbf{v}^c)$ is a diagonal matrix depending on $\mathbf{v}^c$.
  - ◇ The minimizer of $\phi(\mathbf{v})$ is approximated by the minimizer $\mathbf{v}^+$ of $\varphi(\mathbf{v}; \mathbf{v}^c)$, near which a new quadratic model is created.
- The definition of $D(\mathbf{v}^{(c)})$ is quite intriguing.

Lee and Seung Method

- Denote $\mathbf{v}^c = [v_i^c] \in \mathbb{R}^p$, $D(\mathbf{v}^c) = \text{diag}\{d_1(\mathbf{v}^c), \ldots, d_p(\mathbf{v}^c)\}$, and so on.
- Define the diagonal entries is by (Lee and Seung'01)

$$d_i(\mathbf{v}^c) := \frac{(U^\top U \mathbf{v}^c)_i}{v_i^c}, \quad i = 1, \ldots, p.$$

- Four important consequences:
  1. $D(\mathbf{v}^c) - U^\top U$ is positive semi-definite.

$$(\mathbf{v} - \mathbf{v}^c)^\top \left(D(\mathbf{v}^c) - U^\top U\right)(\mathbf{v} - \mathbf{v}^c) \geq 0 \text{for all } \mathbf{v}.$$

  - $\phi(\mathbf{v}) \leq \varphi(\mathbf{v}; \mathbf{v}^c)$ for all $\mathbf{v}$.

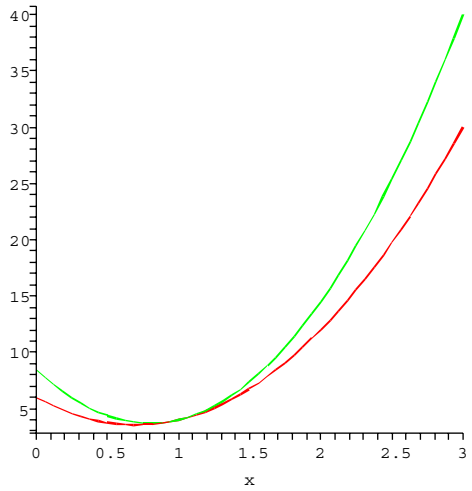

Figure 2: Reduced quadratic model.

2. The minimum of any quadratic function always has a closed form solution.
   - ◇ With $D(\mathbf{v}^c)$ being diagonal the close form solution is easy.
   - ◇ The minimum $\mathbf{v}^+$ of $\varphi(\mathbf{v};\mathbf{v}^c)$ is given by
   $$\mathbf{v}^+ := \mathbf{v}^c - D^{-1}(\mathbf{v}^c)(U^\top U\mathbf{v}^c - U^\top \mathbf{y}).$$
3. The entries of $\mathbf{v}^+$ are precisely
   $$v_i^+ = v_i^c \frac{(U^\top \mathbf{y})_i}{(U^\top U\mathbf{v}^c)_i}, \quad i = 1,\ldots,p.$$
   - ◇ $\mathbf{v}^+$ remains nonnegative if $\mathbf{v}^c$ is nonnegative.
4. Note that
   $$\phi(\mathbf{v}^+) \le \varphi(\mathbf{v}^+;\mathbf{v}^c) \le \varphi(\mathbf{v}^c;\mathbf{v}^c) = \phi(\mathbf{v}^c),$$
   - ◇ $\mathbf{v}^+$ is an improved update from $\mathbf{v}^c$.

Multiplicatie Rule

- Repeat the above process for each individual column of V and each row of U.
- The updated matrix $V^+ = [v_{ij}^+]$ from a given nonnegative matrix $V^c = [v_{ij}^c]$ and a fixed nonnegative matrix U can be defined by the multiplicative rule:

$$v_{ij}^+ := v_{ij}^c \frac{(U^\top Y)_{ij}}{(U^\top U V^c)_{ij}}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n.$$

  - ◇ In short,

$$V^+ := V^c .* (U^\top Y)./(U^\top U V^c).$$

- Similarly, the update $U^+ = [u_{ij}^+]$ from a given nonnegative matrix $U^c = [u_{ij}^c]$ and a fixed nonnegative matrix V can be defined by the rule:

$$U^+ := U^c .* (Y V^\top)./(U^c V V^\top).$$

- Alternating these multiplicative update rules between U and V.
- The objective function $f(U, V)$ is nonincreasing under the update rules.

Ellipsoid Method

- There are many other ways to set forth the simpler model (13).
  - ◇ If all diagonal entries of D are larger than the spectral radius of $U^\top U$, then $D - U^\top U$ is positive definite.
  - ◇ The larger the D, the smaller the D^{-1} and, hence, the less difference between $\mathbf{v}^+$ and $\mathbf{v}^c$ according to (2).
- The challenge thus lies in finding a diagonal matrix D that is
  - ◇ Large enough to make $D - U^\top U$ positive definite.
  - ◇ Small enough to signify the difference between $\mathbf{v}^+$ and $\mathbf{v}^c$.
- Use the semidefinite programming (SDP) technique?

Barrier Function

- Notation:
  - $S = U^\top U \in \mathbb{R}^{p\times p}$.
  - $\mathbf{g}^c = [g_i^c] := U^\top U\mathbf{v}^c - U^\top \mathbf{y} \in \mathbb{R}^p$.
  - $\lambda_i(D) =$ the i^{th} eigenvalue of $D - S$.
- Barrier function

$$\omega(D) := \sum_{i=1}^{p} \ln \frac{1}{\lambda_i(D)} + \sum_{i=1}^{p} \ln \frac{1}{d_i} + \sum_{i=1}^{p} \ln \frac{1}{d_i v_i^c - g_i^c}.$$

  - $\omega(D)$ can be defined only for diagonal matrices D such that $D - S$ is positive definite, D has positive diagonal entries, and $D\mathbf{v}^c - \mathbf{g}^c$ is a positive vector.
  - The level curves of $\omega(D)$ serve as reasonable approximations to the boundary of the desirable feasible domain.

Differential Properties of Barrier Function

- The gradient vector of $\omega(D)$ with $D = \text{diag}\{d_1, \ldots, d_p\}$ is given by

$$\nabla\omega(D) = \text{diag}\left((D-S)^{-1} - D^{-1}\right) - \begin{bmatrix} \frac{v_1^c}{d_1 v_1^c - g_1^c} \\ \vdots \\ \frac{v_k^c}{d_k v_k^c - g_k^c} \end{bmatrix}.$$

- The Hessian matrix $H(D)$ of $\omega(D)$ is given by

$$H(D) = (D-S)^{-1}.*(D-S)^{-1} + D^{-1}.*D^{-1} + \text{diag}\left\{\left(\frac{v_1^c}{d_1 v_1^c - g_1^c}\right)^2, \ldots, \left(\frac{v_k^c}{d_k v_k^c - g_k^c}\right)^2\right\}.$$

  - ◇ By the Schur product theorem, $H(D)$ is positive definite if D is feasible.
  - ◇ $\omega(D)$ is strictly convex over its feasible domain.

Ellipsoid

- An ellipsoid $\mathcal{E} \subset \mathbb{R}^p$ can best be characterized by its center $\boldsymbol{\gamma} \in \mathbb{R}^p$ and a symmetric and positive definite matrix $\Gamma \in \mathbb{R}^{p \times p}$ via

$$\mathcal{E} = \mathcal{E}(\Gamma, \boldsymbol{\gamma}) := \left\{ \mathbf{x} \in \mathbb{R}^p | (\mathbf{x} - \boldsymbol{\gamma})^T \Gamma^{-1} (\mathbf{x} - \boldsymbol{\gamma}) \leq 1 \right\}.$$

- Suppose $D^c = \text{diag}(\mathbf{d}^c)$ is a strictly feasible point. Then every $D^+ = \text{diag}(\mathbf{d}^+)$ with $\mathbf{d}^+ \in \mathcal{E}(H(D^c)^{-1}, \mathbf{d}^c)$ is also strictly feasible.
  - ◇ From within the feasible domain of ω, approximate its level curves by a sequence of inscribed ellipsoids.

Choice of D^c

- Given a feasible D^c, any point from the ellipsoid $\mathcal{E}(H(D^c)^{-1}, d^c)$ will carry the four properties that Lee and Seung's choice possesses.
  - ◇ Which point $\mathbf{d}^+$ on $\mathcal{E}(H(D^c)^{-1}, d^c)$ will serve the "goal" better?
- Some possible choices:
  - ◇ In order to make D^+ small, one possible objective is to minimize the trace of D^+, i.e.,

$$\begin{aligned} \text{minimize} \quad & \mathbf{1}^\top \mathbf{d}, && (14) \\ \text{subject to} \quad & \mathbf{d} \in \mathbf{E}(H(D^c)^{-1}, \mathbf{d}^c). && (15) \end{aligned}$$

  - ◇ Weight the diagonal entries of D differently and end up with different linear objective functional.
- Optimization of linear objective functional over ellipsoids has a closed form solution:
  - ◇ For $\mathbf{p} \neq 0$, the minimal value of the linear functional $\mathbf{p}^\top \mathbf{x}$ subject to the condition $\mathbf{x} \in \mathcal{E}(\Gamma, \boldsymbol{\gamma})$ occurs at

$$\mathbf{x}^* := \boldsymbol{\gamma} - \frac{1}{\sqrt{\mathbf{p}^T \Gamma \mathbf{p}}} \Gamma \mathbf{p}.$$

Dikin's Algoritm

- Given $\mathbf{d}^{(0)} \in \mathbb{R}^p$ strictly feasible, do until convergence:
  1. If $D^{(k)} - S$ is singular, then stop; otherwise
  2. Solve $H(D^{(k)})\delta\mathbf{d} = \mathbf{1}$ for $\delta\mathbf{d}$;
  3. Update $\mathbf{d}^{(k+1)} := \mathbf{d}^{(k)} - \frac{1}{\sqrt{\mathbf{1}^\top \delta\mathbf{d}}}\delta\mathbf{d}$.
- $D_{Lee\&Seung}$ is *always* on the boundary of the feasible domain because $D_{Lee\&Seung} - S$ has a zero eigenvalue with eigenvector $\mathbf{v}^c$.
- While the Dikin algorithm produces a diagonal matrix that has minimal trace, the Lee and Seung algorithm is remarkably cheap for computation.

Gradient Approach

- Gradient information is known explicitly.
- Easy to use, but slow convergence.

Gradient Flow

- The dynamical system

$$\frac{dE}{dt} = E. * \left(\delta(E,F)(F. * F)^\top\right) \in \mathbb{R}^{m\times p}, \tag{16}$$

$$\frac{dF}{dt} = F. * \left((E. * E)^\top \delta(E,F)\right) \in \mathbb{R}^{p\times n}, \tag{17}$$

  defines an analytic continuous flow.

  - ◇ Moves in the space $\mathbb{R}^{m\times p} \times \mathbb{R}^{p\times n}$ along the steepest descent direction of the objective functional g.
  - ◇ Along the solution flow $(E(t), F(t))$,

$$\begin{aligned}\frac{dg(E(t),F(t))}{dt} &= -\langle\left(\delta(E,F)(F. * F)^\top\right) . * E, \left(\delta(E,F)(F. * F)^\top\right) . * E\rangle \\ &\quad -\langle F. * \left((E. * E)^\top \delta(E,F)\right), F. * \left((E. * E)^\top \delta(E,F)\right)\rangle \leq 0.\end{aligned}$$

  - ◇ Analytic gradient flow converges to a single point of equilibrium. (Łojasiewicz'63, Simon'83)

- Employing any available ODE solvers to integrate the system (16) and (17) constitutes a ready-to-go numerical method.

Steepest Descent Method

- The Euler method with appropriate step size selection is an easy way of utilizing the gradient information.
  - ◇ Steepest descent update:

$$E^{(k+1)} := E^{(k)} + \mu_k E^{(k)} .* \left(\delta(E^{(k)}, F^{(k)})(F^{(k)} .* F^{(k)})^\top\right), \tag{18}$$

$$F^{(k+1)} := F^{(k)} + \mu_k F^{(k)} .* ((F^{(k)} .* F^{(k)})^\top \delta(E^{(k)}, F^{(k)})). \tag{19}$$

  - ◇ Shepherd update:

$$U^{(k+1)} = U^{(k+1)}(\mu_k) := \max\left\{0, U^{(k)} + \mu_k (Y - U^{(k)}V^{(k)})(V^{(k)})^\top\right\}, \tag{20}$$

$$V^{(k+1)} = V^{(k+1)}(\mu_k) := V^{(k)} + \mu_k (U^{(k)})^\top (Y - U^{(k)}V^{(k)}). \tag{21}$$

- The selection of μ_k is critical.
  - ◇ In general practice, a backtracking line search using, say, a cubic interpolation and a merit function, is performed to determine the step length μ_k.
  - ◇ For the NNMF problem, the selection of step length is easier.
    - ▷ The function,

$$\Theta(\mu) := F(U^{(k+1)}(\mu), V^{(k+1)}(\mu)),$$

      with $U^{(k+1)}(\mu)$ and $V^{(k+1)}(\mu)$ defined by (21) is a quartic polynomial in μ.
    - ▷ It was suggested to use the Tartaglia formula to compute directly the roots of $\Theta'(\mu)$ and hence locate the optimal μ.

Numerical Experiments

- Many possible numerical methods.
- Each approach has unique features and wide-ranging degrees of complexities.
  - ◇ Not easy to make a fair comparison of their performance.
- We shall demonstrate the limits and difficulties in interpreting the factorizations.

Iris Decomposition

- The NNMF applied to the iris data is meant to seek and identify any intrinsic parts that make up these poses.
  - ◇ We do not know a priori the number p of parts.
    - ▷ Can only experiment with different numbers of p.
  - ◇ Once a factorization UV is found,
    - ▷ Columns of U are normalized to unit length for uniformity.
    - ▷ Columns of the normalized U will be considered as the bases of these images.
- The results:
  - ◇ $p = 2$ suggests quite clearly that there are two "positions" of the pupils.
  - ◇ $p = 4$ indicates that there are two basic images overlaying each other.
  - ◇ The basic "parts" that make up these irises remain disappointedly complicated.

Figure 3: Basis images for $p = 2$.

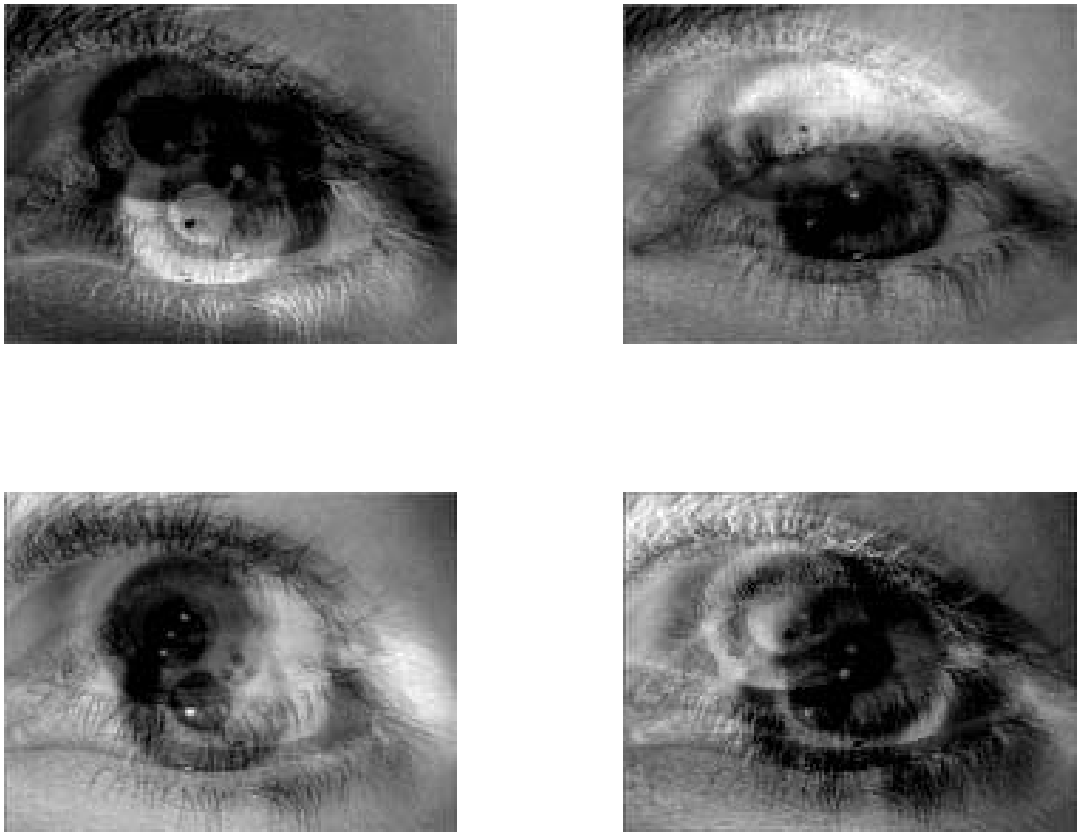

Figure 4: Basis images for $p = 4$.

Pollutant Decompositin

- Assume four principal sectors across the national economy.
  - ◇ Fuel combustion
  - ◇ Industrial Processes:
    - ▷ Chemical and allied product manufacturing
      - · Organic chemical manufacturing
      - · Inorganic chemical manufacturing
      - · Polymer and resin manufacturing
      - · Pharmaceutical manufacturing
      - · ⋮
    - ▷ Metals processing
    - ▷ Petroleum and related industries
    - ▷ Other industrial processes
    - ▷ Solvent utilization
    - ▷ Storage and transport
    - ▷ Waste disposal and recycling
  - ◇ Transportation
  - ◇ Miscellaneous
- Each subsector contributes certain degree of pollution.

Scenario I: Who Is Doing What Damages?

- Assume total emissions F from each sector is available.

	1970	1975	1980	1985	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Fuel	41754	40544	43512	41661	40659	39815	39605	40051	38926	38447	36138	36018	35507	34885	34187
Industrial	48222	32364	29615	22389	21909	21120	20900	21102	21438	21467	21190	17469	17988	17868	20460
Transportation	125637	121674	117527	119116	107978	100877	106571	105114	106328	108125	99642	106069	104748	103523	100783
Miscellaneous	10289	6733	10589	46550	46560	45877	42572	40438	41501	45105	39752	43829	46487	42467	39836

Table 2: Annual emissions estimates (in thousand short tons) from four sectors.

- Determine a nonnegative matrix A of size 8×4 that solves the optimization problem:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|Y - AF\|_F^2, \\ &\text{subject to} && A \geq 0, \quad \text{and} \quad \textstyle\sum_{i=1}^{8} a_{ij} = 1, \quad j = 1, \ldots 4. \end{aligned} \tag{22}$$

  - ◇ Each column of A represents the best fitting percentage distribution of pollutants from the emission of the corresponding sector.
  - ◇ This is a convex programming problem and the global minimizer is unique.

Comparing NNMF and Averaging Results

- Using existing software, such as FMINCON in MATLAB, to find the best fitting distribution A_{opt} to Problem (22).
- The average distribution A_{avg} would have to be obtained by extensive efforts in gathering *itemized* pollutant emissions of each sector per year.

	Fuel	Industrial	Transportation	Miscellaneous
Carbon Monoxide	0.1535	0.3116	0.7667	0.3223
Lead	0.0001	0.0002	0.0002	0
Nitrogen Oxides	0.2754	0.0417	0.1177	0.0113
Volatile Organic	0.0265	0.4314	0.0908	0.0347
PM_{10}	0.0368	0.0768	0.0074	0.4911
Sulfur Dioxide	0.4923	0.0996	0.0112	0.0012
$PM_{2.5}$	0.0148	0.0272	0.0043	0.0761
Ammonia	0.0007	0.0115	0.0016	0.0634

Table 3: Average distribution of pollutants from sectors.

	Fuel	Industrial	Transportation	Miscellaneous
Carbon Monoxide	0.1925	0.3400	0.8226	0.0090
Lead	0	0.0000	0	0.0000
Nitrogen Oxides	0.0631	0	0.1503	0.1524
Volatile Organic	0.3270	0.2759	0.0272	0
PM_{10}	0.0000	0.1070	0.0000	0.6198
Sulfur Dioxide	0.4174	0.2771	0.0000	0
$PM_{2.5}$	0.0000	0.0000	0	0.1326
Ammonia	0.0000	0	0	0.0862

Table 4: Optimal distribution of pollutants from sectors with fixed emission estimates.

- Several serious discrepancies:
  - ◇ In A_{opt} that 32.70% emissions from the fuel burning contribute to the volatile organic compounds whereas A_{avg} counts only 2.65%.
  - ◇ In A_{opt} that only 6.31% emissions from the fuel goes to the nitrogen oxides whereas A_{avg} count 27.54%.
  - ◇ Estimates from the best fitting A_{opt} is inconsistent with the scientific truth.

Scenario II: Total Factorization

- Assume only Y is available.
  - ◇ Determine four sectors, *not necessarily in any order or any definition.*
  - ◇ The corresponding percentage distributions U.
  - ◇ The total emission estimates per year V.
- Results from the Lee and Seung algorithm.

	Sector 1	Sector 2	Sector 3	Sector 4
Carbon Monoxide	0.2468	0.0002	0.7969	0.0001
Lead	0	0.0008	0	0.0000
Nitrogen Oxides	0.0000	0	0.1641	0.1690
Volatile Organic	0.3281	0.2129	0.0391	0
PM_{10}	0.0000	0.5104	0.0000	0.5532
Sulfur Dioxide	0.4251	0.2757	0.0000	0
$PM_{2.5}$	0.0000	0.0000	0	0.1680
Ammonia	0.0000	0	0	0.1097

Table 5: NNMF distribution estimates of pollutants from sectors (Lee and Seung algorithm)

	1970	1975	1980	1985	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Sector 1	58705	57455	57162	3718	4974	47464	46314	47175	47864	43630	44643	42657	43578	42926	43585
Sector 2	25487	11755	7431	81042	75327	10313	10784	8313	6848	12613	4069	3403	3541	3159	1
Sector 3	143614	128945	130225	145512	132349	109442	113118	109881	110295	116521	104440	113926	113910	108437	108828
Sector 4	0	3139	6254	2	4702	40785	39618	41539	43358	40302	43236	43319	43599	44239	42832

Table 6: NNMF emission estimates (in thousand short tons) (Lee and Seung algorithm).

- We do not know what each column of U really stand for.
- It requires a careful interpretation to identify what *factor* is being represented.
  - ◇ It is likely that a single column could represent a mixture of two or more known economy sectors.
- Improvement in the objective functions.

$$\frac{1}{2}\|Y - UV\|_F^2 = 1.5873 \times 10^7 < \|Y - A_{opt}F\|_F^2 = 2.7017 \times 10^8 < \frac{1}{2}\|Y - A_{avg}F\|^2 = 7.1548 \times 10^8.$$

  - ◇ Unevenness in the NNMF emission estimates per sector, making it more difficult to predict the estimate.

- Results from the constrained quasi-Newton method.
  - ◇ A different percentage distribution of pollutants from sectors.

	Sector 1	Sector 2	Sector 3	Sector 4
Carbon Monoxide	0.3124	0.4468	0.5426	0.6113
Lead	0	0	0.0000	0.0007
Nitrogen Oxides	0.1971	0.1299	0.0366	0.1412
Volatile Organic	0.0239	0.0654	0.1720	0.1191
PM_{10}	0.1936	0.3101	0.0401	0.0220
Sulfur Dioxide	0.0287	0.0477	0.2087	0.1058
$PM_{2.5}$	0.1480	0.0000	0	0
Ammonia	0.0963	0	0.0000	0.0000

Table 7: NNMF distribution estimates of pollutants from sectors (constrained quasi-Newton method).

- Computationally more expensive.
  - ◇ Able to find local solutions that give smaller objective values (1.0645×10^7).
  - ◇ Not clear how to identify the sectors and to interpret the distributions of pollutants.

Conclusion

- The nonnegative matrix factorization has been desired by many important applications.
- We have suggested a number of numerical procedures that can be employed to obtain a factorization that is at least locally optimal.
  - ◇ Not clear which method is superior.
- We have demonstrated by two real-world problems that the factorization itself does not necessarily provide immediate interpretation of the real data
  - ◇ The basic parts of the irises are themselves complicated images (and sometimes with overlapped irises).
  - ◇ The percentage distributions of pollutants from economical sectors are not always consistent with data obtained by other means (and could represent mixtures across several sectors.)
- Proper interpretations or additional constraints on the factors are needed for NNMF applications.