Gaussian low-rank approximations with Gaussian priors

Jack Poulson (Hodge Star Scientific Computing)

Aussois, France, June 21, 2019

June 21, 2019

Bayesian interpretation of SVD

An SVD approach to fitting a low-rank matrix XY^T to a matrix A would be to minimize

$$\frac{1}{2}\|A - XY^T\|_F^2 = \frac{1}{2}\sum_{i,j}(A_{i,j} - (XY^T)_{i,j})^2.$$

Recall that the log-likelihood of a Gaussian of mean x_0 and variance σ^2 is

$$\log((2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2\sigma^2}(x - x_0)^2)) = -\frac{1}{2\sigma^2}(x - x_0)^2 + C.$$

We can thus interpret minimizing the SVD loss as maximizing the likelihood if the entries of A are independent standard normals centered about the corresponding entry of XY^T.

But what do we do if we only have observations for a small number of entries of A?

Running SVD with the unobserved entries replaced with zero would spend most of its skill fitting the zeros if there were only a few observations – the homoscedastic assumption would be inappropriate.

Bayesian interpretation of SVD

An SVD approach to fitting a low-rank matrix XY^T to a matrix A would be to minimize

$$\frac{1}{2}\|A - XY^T\|_F^2 = \frac{1}{2}\sum_{i,j}(A_{i,j} - (XY^T)_{i,j})^2.$$

Recall that the log-likelihood of a Gaussian of mean x_0 and variance σ^2 is

$$\log((2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2\sigma^2}(x - x_0)^2)) = -\frac{1}{2\sigma^2}(x - x_0)^2 + C.$$

We can thus interpret minimizing the SVD loss as maximizing the likelihood if the entries of A are independent standard normals centered about the corresponding entry of XY^T.

But what do we do if we only have observations for a small number of entries of A?

Running SVD with the unobserved entries replaced with zero would spend most of its skill fitting the zeros if there were only a few observations – the homoscedastic assumption would be inappropriate.

Bayesian interpretation of SVD

An SVD approach to fitting a low-rank matrix XY^T to a matrix A would be to minimize

$$\frac{1}{2}\|A - XY^T\|_F^2 = \frac{1}{2}\sum_{i,j}(A_{i,j} - (XY^T)_{i,j})^2.$$

Recall that the log-likelihood of a Gaussian of mean x_0 and variance σ^2 is

$$\log((2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2\sigma^2}(x - x_0)^2)) = -\frac{1}{2\sigma^2}(x - x_0)^2 + C.$$

We can thus interpret minimizing the SVD loss as maximizing the likelihood if the entries of A are independent standard normals centered about the corresponding entry of XY^T.

But what do we do if we only have observations for a small number of entries of A?

Running SVD with the unobserved entries replaced with zero would spend most of its skill fitting the zeros if there were only a few observations – the homoscedastic assumption would be inappropriate.

Bayesian interpretation of SVD

An SVD approach to fitting a low-rank matrix XY^T to a matrix A would be to minimize

$$\frac{1}{2}\|A - XY^T\|_F^2 = \frac{1}{2}\sum_{i,j}(A_{i,j} - (XY^T)_{i,j})^2.$$

Recall that the log-likelihood of a Gaussian of mean x_0 and variance σ^2 is

$$\log((2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2\sigma^2}(x - x_0)^2)) = -\frac{1}{2\sigma^2}(x - x_0)^2 + C.$$

We can thus interpret minimizing the SVD loss as maximizing the likelihood if the entries of A are independent standard normals centered about the corresponding entry of XY^T.

But what do we do if we only have observations for a small number of entries of A?

Running SVD with the unobserved entries replaced with zero would spend most of its skill fitting the zeros if there were only a few observations – the homoscedastic assumption would be inappropriate.

Bayesian interp'n of SVD on observeds

If we only observed a collection of entries with indices $\mathcal{A} \subseteq [n] \times [n]$, we could replace the objective with

$$\frac{1}{2}\|\mathbb{1}_{\mathcal{A}} \circ (A - XY^T)\|_F^2 = \frac{1}{2} \sum_{(i,j)\in\mathcal{A}} (A_{i,j} - (XY^T)_{i,j})^2,$$

where $\circ$ here represents the Hadamard product, and $\mathbb{1}_{\mathcal{A}}$ is the binary indicator over the sparsity pattern of the observed entries.

This model is thus equivalent to maximizing the likelihood of a low-rank matrix XY^T if the **observed** entries of A are independent standard normals centered about the corresponding entry of XY^T.

This is standard fare for compressed sensing, where the typical assumption is that the observation locations are uniformly random.

But **there is typically signal in an entry not being observed**. Consider a cooccurrence score for two words based upon the number of times (and distance apart) they are in different passages from wikipedia.

It is useful information that two words were never observed in the same context. But it does not mean that it would never happen.

Bayesian interp'n of SVD on observeds

If we only observed a collection of entries with indices $\mathcal{A} \subseteq [n] \times [n]$, we could replace the objective with

$$\frac{1}{2}\|\mathbb{1}_{\mathcal{A}} \circ (A - XY^T)\|_F^2 = \frac{1}{2} \sum_{(i,j)\in\mathcal{A}} (A_{i,j} - (XY^T)_{i,j})^2,$$

where $\circ$ here represents the Hadamard product, and $\mathbb{1}_{\mathcal{A}}$ is the binary indicator over the sparsity pattern of the observed entries.

This model is thus equivalent to maximizing the likelihood of a low-rank matrix XY^T if the **observed** entries of A are independent standard normals centered about the corresponding entry of XY^T.

This is standard fare for compressed sensing, where the typical assumption is that the observation locations are uniformly random.

But **there is typically signal in an entry not being observed**. Consider a cooccurrence score for two words based upon the number of times (and distance apart) they are in different passages from wikipedia.

It is useful information that two words were never observed in the same context. But it does not mean that it would never happen.

Bayesian interp'n of SVD on observeds

If we only observed a collection of entries with indices $\mathcal{A} \subseteq [n] \times [n]$, we could replace the objective with

$$\frac{1}{2}\|\mathbb{1}_{\mathcal{A}} \circ (A - XY^T)\|_F^2 = \frac{1}{2} \sum_{(i,j)\in\mathcal{A}} (A_{i,j} - (XY^T)_{i,j})^2,$$

where $\circ$ here represents the Hadamard product, and $\mathbb{1}_{\mathcal{A}}$ is the binary indicator over the sparsity pattern of the observed entries.

This model is thus equivalent to maximizing the likelihood of a low-rank matrix XY^T if the **observed** entries of A are independent standard normals centered about the corresponding entry of XY^T.

This is standard fare for compressed sensing, where the typical assumption is that the observation locations are uniformly random.

But **there is typically signal in an entry not being observed**. Consider a cooccurrence score for two words based upon the number of times (and distance apart) they are in different passages from wikipedia.

It is useful information that two words were never observed in the same context. But it does not mean that it would never happen.

Bayesian interp'n of SVD on observeds

If we only observed a collection of entries with indices $\mathcal{A} \subseteq [n] \times [n]$, we could replace the objective with

$$\frac{1}{2}\|\mathbb{1}_{\mathcal{A}} \circ (A - XY^T)\|_F^2 = \frac{1}{2}\sum_{(i,j)\in\mathcal{A}} (A_{i,j} - (XY^T)_{i,j})^2,$$

where $\circ$ here represents the Hadamard product, and $\mathbb{1}_{\mathcal{A}}$ is the binary indicator over the sparsity pattern of the observed entries.

This model is thus equivalent to maximizing the likelihood of a low-rank matrix XY^T if the **observed** entries of A are independent standard normals centered about the corresponding entry of XY^T.

This is standard fare for compressed sensing, where the typical assumption is that the observation locations are uniformly random.

But **there is typically signal in an entry not being observed**. Consider a cooccurrence score for two words based upon the number of times (and distance apart) they are in different passages from wikipedia.

It is useful information that two words were never observed in the same context. But it does not mean that it would never happen.

Bayesian interp'n of SVD on observeds

If we only observed a collection of entries with indices $\mathcal{A} \subseteq [n] \times [n]$, we could replace the objective with

$$\frac{1}{2}\|\mathbb{1}_{\mathcal{A}} \circ (A - XY^T)\|_F^2 = \frac{1}{2} \sum_{(i,j)\in\mathcal{A}} (A_{i,j} - (XY^T)_{i,j})^2,$$

where $\circ$ here represents the Hadamard product, and $\mathbb{1}_{\mathcal{A}}$ is the binary indicator over the sparsity pattern of the observed entries.

This model is thus equivalent to maximizing the likelihood of a low-rank matrix XY^T if the **observed** entries of A are independent standard normals centered about the corresponding entry of XY^T.

This is standard fare for compressed sensing, where the typical assumption is that the observation locations are uniformly random.

But **there is typically signal in an entry not being observed**. Consider a cooccurrence score for two words based upon the number of times (and distance apart) they are in different passages from wikipedia.

It is useful information that two words were never observed in the same context. But it does not mean that it would never happen.

Gap SVD

Generalizing from an indicator entry weighting, we might consider

$$\frac{1}{2}\|\sqrt{W}\circ(A-XY^T)\|_F^2+\frac{\lambda_X}{2}\|X\|_F^2+\frac{\lambda_Y}{2}\|Y\|_F^2,$$

where we can interpret as a maximum-likelihood model where each $A_{i,j}\sim\mathcal{N}(0,W_{i,j}^{-1})$ is independent, with $\sqrt{W}$ being an entry-wise square-root, plus the prior of each entry of X being independently drawn from $\mathcal{N}(0,1/\lambda_X)$ and Y from $\mathcal{N}(0,1/\lambda_Y)$.

The simplest such extension is to set observed entries at a fixed variance, say, 1, and the unobserved entries at a lower variance, $\delta\ll 1$, with a predetermined value (e.g., 0).

Then, generally, both W and A would be sparse plus rank-one, and an efficient Alternating Weighted Least Squares (AWLS) solver tailored to it was proposed in [Pan/Scholz-2009].

Gap SVD

Generalizing from an indicator entry weighting, we might consider

$$\frac{1}{2}\|\sqrt{W} \circ (A - XY^T)\|_F^2 + \frac{\lambda_X}{2}\|X\|_F^2 + \frac{\lambda_Y}{2}\|Y\|_F^2,$$

where we can interpret as a maximum-likelihood model where each $A_{i,j} \sim \mathcal{N}(0, W_{i,j}^{-1})$ is independent, with $\sqrt{W}$ being an entry-wise square-root, plus the prior of each entry of X being independently drawn from $\mathcal{N}(0, 1/\lambda_X)$ and Y from $\mathcal{N}(0, 1/\lambda_Y)$.

The simplest such extension is to set observed entries at a fixed variance, say, 1, and the unobserved entries at a lower variance, $\delta \ll 1$, with a predetermined value (e.g., 0).

Then, generally, both W and A would be sparse plus rank-one, and an efficient Alternating Weighted Least Squares (AWLS) solver tailored to it was proposed in [Pan/Scholz-2009].

Gap SVD

Generalizing from an indicator entry weighting, we might consider

$$\frac{1}{2}\|\sqrt{W}\circ(A-XY^T)\|_F^2+\frac{\lambda_X}{2}\|X\|_F^2+\frac{\lambda_Y}{2}\|Y\|_F^2,$$

where we can interpret as a maximum-likelihood model where each $A_{i,j}\sim\mathcal{N}(0,W_{i,j}^{-1})$ is independent, with $\sqrt{W}$ being an entry-wise square-root, plus the prior of each entry of X being independently drawn from $\mathcal{N}(0,1/\lambda_X)$ and Y from $\mathcal{N}(0,1/\lambda_Y)$.

The simplest such extension is to set observed entries at a fixed variance, say, 1, and the unobserved entries at a lower variance, $\delta \ll 1$, with a predetermined value (e.g., 0).

Then, generally, both W and A would be sparse plus rank-one, and an efficient Alternating Weighted Least Squares (AWLS) solver tailored to it was proposed in [Pan/Scholz-2009].

AWLS for Gap SVD

Given the Lagrangian

$$\mathcal{L}(X, Y) = \frac{1}{2}\|\sqrt{W} \circ (A - XY^T)\|_F^2 + \frac{\lambda_X}{2}\|X\|_F^2 + \frac{\lambda_Y}{2}\|Y\|_F^2,$$

let us consider the first-order optimality conditions if we froze X about its current estimate to define $\mathcal{L}_X(Y) = \mathcal{L}(X, Y)$.

$$\begin{aligned}
(d_{Y_{a,b}})(\mathcal{L}_X) &= (d_{Y_{a,b}})(\frac{1}{2}\sum_{i,j} W_{i,j}(A_{i,j} - \sum_k X_{i,k}Y_{j,k})^2) + \lambda_Y Y_{a,b} \\
&= \frac{1}{2}\sum_{i} W_{i,a}(d_{Y_{a,b}})((A_{i,a} - \sum_k X_{i,k}Y_{a,k})^2)) + \lambda_Y Y_{a,b} \\
&= -\sum_{i} W_{i,a}X_{i,b}(A_{i,a} - \sum_k X_{i,k}Y_{a,k}) + \lambda Y_{a,b} = 0.
\end{aligned}$$

Or, for row a of Y, y_a,

$$(X^T\text{diag}(W_{:,a})X + \lambda_Y I)y_a^T = X^T\text{diag}(W_{:,a})A_{:,a}$$

AWLS for Gap SVD

Given the Lagrangian

$$\mathcal{L}(X,Y) = \frac{1}{2}\|\sqrt{W} \circ (A - XY^T)\|_F^2 + \frac{\lambda_X}{2}\|X\|_F^2 + \frac{\lambda_Y}{2}\|Y\|_F^2,$$

let us consider the first-order optimality conditions if we froze X about its current estimate to define $\mathcal{L}_X(Y) = \mathcal{L}(X,Y)$.

$$\begin{aligned}(d_{Y_{a,b}})(\mathcal{L}_X) &= (d_{Y_{a,b}})(\frac{1}{2}\sum_{i,j} W_{i,j}(A_{i,j} - \sum_k X_{i,k}Y_{j,k})^2) + \lambda_Y Y_{a,b} \\ &= \frac{1}{2}\sum_i W_{i,a}(d_{Y_{a,b}})((A_{i,a} - \sum_k X_{i,k}Y_{a,k})^2)) + \lambda_Y Y_{a,b} \\ &= -\sum_i W_{i,a}X_{i,b}(A_{i,a} - \sum_k X_{i,k}Y_{a,k}) + \lambda Y_{a,b} = 0.\end{aligned}$$

Or, for row a of Y, y_a,

$$(X^T \mathrm{diag}(W_{:,a})X + \lambda_Y I)y_a^T = X^T \mathrm{diag}(W_{:,a})A_{:,a}$$

AWLS for Gap SVD

Given the Lagrangian

$$\mathcal{L}(X, Y) = \frac{1}{2}\|\sqrt{W} \circ (A - XY^T)\|_F^2 + \frac{\lambda_X}{2}\|X\|_F^2 + \frac{\lambda_Y}{2}\|Y\|_F^2,$$

let us consider the first-order optimality conditions if we froze X about its current estimate to define $\mathcal{L}_X(Y) = \mathcal{L}(X, Y)$.

$$\begin{aligned}(d_{Y_{a,b}})(\mathcal{L}_X) &= (d_{Y_{a,b}})(\frac{1}{2}\sum_{i,j} W_{i,j}(A_{i,j} - \sum_k X_{i,k}Y_{j,k})^2) + \lambda_Y Y_{a,b} \\ &= \frac{1}{2}\sum_i W_{i,a}(d_{Y_{a,b}})((A_{i,a} - \sum_k X_{i,k}Y_{a,k})^2)) + \lambda_Y Y_{a,b} \\ &= -\sum_i W_{i,a}X_{i,b}(A_{i,a} - \sum_k X_{i,k}Y_{a,k}) + \lambda Y_{a,b} = 0.\end{aligned}$$

Or, for row a of Y, y_a,

$$(X^T \operatorname{diag}(W_{:,a})X + \lambda_Y I)y_a^T = X^T \operatorname{diag}(W_{:,a})A_{:,a}$$

AWLS for Gap SVD

If freezing X to update Y, the update equation

$$(X^T \mathrm{diag}(W_{:,a})X + \lambda_Y I)y_a^T = X^T \mathrm{diag}(W_{:,a})A_{:,a}$$

is equivalent to the Weighted Least Squares problem

$$\arg\min_{y_a} \left\| \begin{pmatrix} \sqrt{\mathrm{diag}(W_{:,a})}X \\ \sqrt{\lambda_Y} I \end{pmatrix} y_a - \begin{bmatrix} \sqrt{\mathrm{diag}(W_{:,a})}A_{:,a} \\ 0 \end{bmatrix} \right\|_F^2 .$$

Since W is decomposable as $\delta 1_{m,n} + \tilde{W}$, where $\tilde{W}$ has the same sparsity pattern as the observations of A, we can reuse the **background Gramian** $G_X := \delta X^T X + \lambda_Y I$ to save the bulk of the work of forming the separate row Gramians.

The Gramian for row a is then:

$$G_X^a = G_X + \sum_{i:(i,a)\in\mathcal{A}} \tilde{W}_{i,a} x_i^T x_i,$$

where the sum maps to a symmetric rank-k update (`syrk`) if we absorb the square-root of $\tilde{W}_{i,a}$ into each x_i.

AWLS for Gap SVD

If freezing X to update Y, the update equation

$$(X^T \mathrm{diag}(W_{:,a})X + \lambda_Y I)y_a^T = X^T \mathrm{diag}(W_{:,a})A_{:,a}$$

is equivalent to the Weighted Least Squares problem

$$\arg\min_{y_a} \left\| \begin{pmatrix} \sqrt{\mathrm{diag}(W_{:,a})}X \\ \sqrt{\lambda_Y} I \end{pmatrix} y_a - \begin{bmatrix} \sqrt{\mathrm{diag}(W_{:,a})}A_{:,a} \\ 0 \end{bmatrix} \right\|_F^2 .$$

Since W is decomposable as $\delta 1_{m,n} + \tilde{W}$, where $\tilde{W}$ has the same sparsity pattern as the observations of A, we can reuse the **background Gramian** $G_X := \delta X^T X + \lambda_Y I$ to save the bulk of the work of forming the separate row Gramians.

The Gramian for row a is then:

$$G_X^a = G_X + \sum_{i:(i,a)\in\mathcal{A}} \tilde{W}_{i,a} x_i^T x_i,$$

where the sum maps to a symmetric rank-k update (`syrk`) if we absorb the square-root of $\tilde{W}_{i,a}$ into each x_i.

AWLS for Gap SVD

If freezing X to update Y, the update equation

$$(X^T \mathrm{diag}(W_{:,a})X + \lambda_Y I)y_a^T = X^T \mathrm{diag}(W_{:,a})A_{:,a}$$

is equivalent to the Weighted Least Squares problem

$$\arg\min_{y_a} \left\| \begin{pmatrix} \sqrt{\mathrm{diag}(W_{:,a})}X \\ \sqrt{\lambda_Y} I \end{pmatrix} y_a - \begin{bmatrix} \sqrt{\mathrm{diag}(W_{:,a})}A_{:,a} \\ 0 \end{bmatrix} \right\|_F^2 .$$

Since W is decomposable as $\delta 1_{m,n} + \tilde{W}$, where $\tilde{W}$ has the same sparsity pattern as the observations of A, we can reuse the **background Gramian** $G_X := \delta X^T X + \lambda_Y I$ to save the bulk of the work of forming the separate row Gramians.

The Gramian for row a is then:

$$G_X^a = G_X + \sum_{i:(i,a)\in\mathcal{A}} \tilde{W}_{i,a} x_i^T x_i,$$

where the sum maps to a symmetric rank-k update (`syrk`) if we absorb the square-root of $\tilde{W}_{i,a}$ into each x_i.

AWLS for Gap SVD

The update for x_a (while Y is frozen) is similar:

$$\begin{aligned} G_Y &= \delta Y^T Y + \lambda_X I \\ G_Y^a &= G_Y + \sum_{j:(a,j)\in\mathcal{A}} \tilde{W}_{a,j} y_j^T y_j \\ G_Y^a x_a &= Y^T \operatorname{diag}(W_{a,:}) A_{a,:}^T. \end{aligned}$$

In practice, one can randomly initialize and perform 10 pairs of alternating updates.

Note that the $\lambda_X I$ and $\lambda_Y I$ Gramian contributions provide a lower bound on the smallest singular value of the Gramians due to the zero-centered prior.

We could alternatively provide stabilization of the normal equations via **proximal regulazation** with moving objective terms of the form $\frac{\lambda_P}{2}\|X - X_0\|_F^2$ and $\frac{\lambda_P}{2}\|Y - Y_0\|_F^2$.

AWLS for Gap SVD

The update for x_a (while Y is frozen) is similar:

$$\begin{aligned} G_Y &= \delta Y^T Y + \lambda_X I \\ G_Y^a &= G_Y + \sum_{j:(a,j)\in\mathcal{A}} \tilde{W}_{a,j} y_j^T y_j \\ G_Y^a x_a &= Y^T \mathrm{diag}(W_{a,:}) A_{a,:}^T. \end{aligned}$$

In practice, one can randomly initialize and perform 10 pairs of alternating updates.

Note that the $\lambda_X I$ and $\lambda_Y I$ Gramian contributions provide a lower bound on the smallest singular value of the Gramians due to the zero-centered prior.

We could alternatively provide stabilization of the normal equations via **proximal regulazation** with moving objective terms of the form $\frac{\lambda_P}{2}\|X - X_0\|_F^2$ and $\frac{\lambda_P}{2}\|Y - Y_0\|_F^2$.

AWLS for Gap SVD

The update for x_a (while Y is frozen) is similar:

$$\begin{aligned} G_Y &= \delta Y^T Y + \lambda_X I \\ G_Y^a &= G_Y + \sum_{j:(a,j)\in\mathcal{A}} \tilde{W}_{a,j} y_j^T y_j \\ G_Y^a x_a &= Y^T \mathrm{diag}(W_{a,:}) A_{a,:}^T. \end{aligned}$$

In practice, one can randomly initialize and perform 10 pairs of alternating updates.

Note that the $\lambda_X I$ and $\lambda_Y I$ Gramian contributions provide a lower bound on the smallest singular value of the Gramians due to the zero-centered prior.

We could alternatively provide stabilization of the normal equations via **proximal regulazation** with moving objective terms of the form $\frac{\lambda_p}{2}\|X - X_0\|_F^2$ and $\frac{\lambda_p}{2}\|Y - Y_0\|_F^2$.

AWLS for Gap SVD

The update for x_a (while Y is frozen) is similar:

$$\begin{aligned} G_Y &= \delta Y^T Y + \lambda_X I \\ G_Y^a &= G_Y + \sum_{j:(a,j)\in\mathcal{A}} \tilde{W}_{a,j} y_j^T y_j \\ G_Y^a x_a &= Y^T \operatorname{diag}(W_{a,:}) A_{a,:}^T. \end{aligned}$$

In practice, one can randomly initialize and perform 10 pairs of alternating updates.

Note that the $\lambda_X I$ and $\lambda_Y I$ Gramian contributions provide a lower bound on the smallest singular value of the Gramians due to the zero-centered prior.

We could alternatively provide stabilization of the normal equations via **proximal regulazation** with moving objective terms of the form $\frac{\lambda_p}{2}\|X - X_0\|_F^2$ and $\frac{\lambda_p}{2}\|Y - Y_0\|_F^2$.

Column rescalings

For matrices with highly non-uniform row/column degrees, it can be important to downweight high degree rows.

The easiest way to do so is to generalize our weight matrix:

$$W = \delta 1_{m,n} + \tilde{W} \mapsto \delta 1_{m,n} + \mathrm{diag}(r)\tilde{W}\mathrm{diag}(c).$$

We simply need to use $\mathrm{diag}(r)\tilde{W}\mathrm{diag}(c)$ instead of $\tilde{W}$ for the row-specific Gramian correction weights..

For a matrix like

$$\begin{pmatrix} 0 & 0 & 0 & x & 0 & x \\ x & x & x & x & x & x \\ 0 & x & 0 & 0 & 0 & 0 \end{pmatrix}$$

we may want to set the second entry of r lower than the others so that the low-rank model does not overemphasize said row.

Column rescalings

For matrices with highly non-uniform row/column degrees, it can be important to downweight high degree rows.

The easiest way to do so is to generalize our weight matrix:

$$W = \delta 1_{m,n} + \tilde{W} \mapsto \delta 1_{m,n} + \mathrm{diag}(r)\tilde{W}\mathrm{diag}(c).$$

We simply need to use $\mathrm{diag}(r)\tilde{W}\mathrm{diag}(c)$ instead of $\tilde{W}$ for the row-specific Gramian correction weights..

For a matrix like

$$\begin{pmatrix} 0 & 0 & 0 & x & 0 & x \\ x & x & x & x & x & x \\ 0 & x & 0 & 0 & 0 & 0 \end{pmatrix}$$

we may want to set the second entry of r lower than the others so that the low-rank model does not overemphasize said row.

Column rescalings

For matrices with highly non-uniform row/column degrees, it can be important to downweight high degree rows.

The easiest way to do so is to generalize our weight matrix:

$$W = \delta 1_{m,n} + \tilde{W} \mapsto \delta 1_{m,n} + \mathrm{diag}(r)\tilde{W}\mathrm{diag}(c).$$

We simply need to use $\mathrm{diag}(r)\tilde{W}\mathrm{diag}(c)$ instead of $\tilde{W}$ for the row-specific Gramian correction weights..

For a matrix like

$$\begin{pmatrix} 0 & 0 & 0 & x & 0 & x \\ x & x & x & x & x & x \\ 0 & x & 0 & 0 & 0 & 0 \end{pmatrix}$$

we may want to set the second entry of r lower than the others so that the low-rank model does not overemphasize said row.

Relaxing off sphere

The resulting factors X and Y will generally not have unit row norms, and they roughly encode the popularity of their corresponding item.

Cosine similarity,

$$\cos(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

is scale invariant, so vectors are often normalized to live on the unit sphere.

But we may want to at least preserve some small portion of the norms, e.g., by dividing by $\|x\|_2^p$, $0 < p < 1$, so that we may penalize recommendations of rare items from common items:

$$\frac{x^T y}{\|x\|_2 \max(\|x\|_2, \|y\|_2)}.$$

But one could argue this is performing a reranker function in the retrieval.

Relaxing off sphere

The resulting factors X and Y will generally not have unit row norms, and they roughly encode the popularity of their corresponding item.

Cosine similarity,

$$\cos(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

is scale invariant, so vectors are often normalized to live on the unit sphere.

But we may want to at least preserve some small portion of the norms, e.g., by dividing by $\|x\|_2^p$, $0 < p < 1$, so that we may penalize recommendations of rare items from common items:

$$\frac{x^T y}{\|x\|_2 \max(\|x\|_2, \|y\|_2)}.$$

But one could argue this is performing a reranker function in the retrieval.

Relaxing off sphere

The resulting factors X and Y will generally not have unit row norms, and they roughly encode the popularity of their corresponding item.

Cosine similarity,

$$\cos(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

is scale invariant, so vectors are often normalized to live on the unit sphere.

But we may want to at least preserve some small portion of the norms, e.g., by dividing by $\|x\|_2^p$, $0 < p < 1$, so that we may penalize recommendations of rare items from common items:

$$\frac{x^T y}{\|x\|_2 \max(\|x\|_2, \|y\|_2)}.$$

But one could argue this is performing a reranker function in the retrieval.

Relaxing off sphere

The resulting factors X and Y will generally not have unit row norms, and they roughly encode the popularity of their corresponding item.

Cosine similarity,

$$\cos(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

is scale invariant, so vectors are often normalized to live on the unit sphere.

But we may want to at least preserve some small portion of the norms, e.g., by dividing by $\|x\|_2^p$, $0 < p < 1$, so that we may penalize recommendations of rare items from common items:

$$\frac{x^T y}{\|x\|_2 \max(\|x\|_2, \|y\|_2)}.$$

But one could argue this is performing a reranker function in the retrieval.

Convex cones

In order to prep for the next lecture, we will talk about some of the background mathematics on symmetric cones and Jordan algebras.

Convex cones

Definition 1. A **cone** is a subset K of a vector space V which, for any $x \in K$ and $\alpha > 0$, satisfies $\alpha x \in K$.

Proposition 1. A **convex cone** is a subset K of a vector space V which, for any $x, y \in K$ and $\alpha, \beta > 0$, satisfies $\alpha x + \beta y \in K$.

Proof.

Suppose $x, y \in K$ and $\alpha, \beta > 0$. Then, choosing $t \in (0, 1)$,

$$\alpha x + \beta y = t\left(\frac{\alpha}{t}x\right) + (1-t)\left(\frac{\beta}{1-t}y\right)$$

is a convex combination of members of the convex cone, so it is also in the cone. □

Question: What is the simplest example you can think of which is a nonconvex cone?

Proposition 2. The sets of symmetric positive-definite and symmetric positive semidefinite matrices of given order are convex cones.

Convex cones

Definition 1. A **cone** is a subset K of a vector space V which, for any $x \in K$ and $\alpha > 0$, satisfies $\alpha x \in K$.

Proposition 1. A **convex cone** is a subset K of a vector space V which, for any $x, y \in K$ and $\alpha, \beta > 0$, satisfies $\alpha x + \beta y \in K$.

Proof.

Suppose $x, y \in K$ and $\alpha, \beta > 0$. Then, choosing $t \in (0, 1)$,

$$\alpha x + \beta y = t \left(\frac{\alpha}{t} x \right) + (1 - t) \left(\frac{\beta}{1 - t} y \right)$$

is a convex combination of members of the convex cone, so it is also in the cone. □

Question: What is the simplest example you can think of which is a nonconvex cone?

Proposition 2. The sets of symmetric positive-definite and symmetric positive semidefinite matrices of given order are convex cones.

Convex cones

Definition 1. A **cone** is a subset K of a vector space V which, for any $x \in K$ and $\alpha > 0$, satisfies $\alpha x \in K$.

Proposition 1. A **convex cone** is a subset K of a vector space V which, for any $x, y \in K$ and $\alpha, \beta > 0$, satisfies $\alpha x + \beta y \in K$.

Proof.

Suppose $x, y \in K$ and $\alpha, \beta > 0$. Then, choosing $t \in (0, 1)$,

$$\alpha x + \beta y = t\left(\frac{\alpha}{t}x\right) + (1-t)\left(\frac{\beta}{1-t}y\right)$$

is a convex combination of members of the convex cone, so it is also in the cone. □

Question: What is the simplest example you can think of which is a nonconvex cone?

Proposition 2. The sets of symmetric positive-definite and symmetric positive semidefinite matrices of given order are convex cones.

Convex cones

Definition 1. A **cone** is a subset K of a vector space V which, for any $x \in K$ and $\alpha > 0$, satisfies $\alpha x \in K$.

Proposition 1. A **convex cone** is a subset K of a vector space V which, for any $x, y \in K$ and $\alpha, \beta > 0$, satisfies $\alpha x + \beta y \in K$.

Proof.

Suppose $x, y \in K$ and $\alpha, \beta > 0$. Then, choosing $t \in (0, 1)$,

$$\alpha x + \beta y = t\left(\frac{\alpha}{t}x\right) + (1-t)\left(\frac{\beta}{1-t}y\right)$$

is a convex combination of members of the convex cone, so it is also in the cone. □

Question: What is the simplest example you can think of which is a nonconvex cone?

Proposition 2. The sets of symmetric positive-definite and symmetric positive semidefinite matrices of given order are convex cones.

Dual cones

Definition 2. The dual of a convex cone K in a vector space V is

$$K^* = \{v \in V : \langle w, v \rangle > 0, \ \forall\, w \in K\}\,.$$

Question: What is the dual of the convex cone $\mathbb{R}_+$?

Question: What is the dual of the convex cone $\mathbb{R}^n_+$?

Question: What is the dual of the closed upper-half plane of $\mathbb{R}^2$?

Question: What are the duals of the convex cones S^n_+ and S^n_{++}?

Dual cones

Definition 2. The dual of a convex cone K in a vector space V is

$$K^* = \{v \in V : \langle w, v \rangle > 0, \ \forall\, w \in K\}\,.$$

Question: What is the dual of the convex cone $\mathbb{R}_+$?

Question: What is the dual of the convex cone $\mathbb{R}^n_+$?

Question: What is the dual of the closed upper-half plane of $\mathbb{R}^2$?

Question: What are the duals of the convex cones S^n_+ and S^n_{++}?

Dual cones

Definition 2. The dual of a convex cone K in a vector space V is

$$K^* = \{v \in V : \langle w, v \rangle > 0, \ \forall\, w \in K\}.$$

Question: What is the dual of the convex cone $\mathbb{R}_+$?

Question: What is the dual of the convex cone $\mathbb{R}^n_+$?

Question: What is the dual of the closed upper-half plane of $\mathbb{R}^2$?

Question: What are the duals of the convex cones S^n_+ and S^n_{++}?

Dual cones

Definition 2. The dual of a convex cone K in a vector space V is

$$K^* = \{v \in V : \langle w, v \rangle > 0, \ \forall\, w \in K\}\,.$$

Question: What is the dual of the convex cone $\mathbb{R}_+$?

Question: What is the dual of the convex cone $\mathbb{R}^n_+$?

Question: What is the dual of the closed upper-half plane of $\mathbb{R}^2$?

Question: What are the duals of the convex cones S^n_+ and S^n_{++}?

Dual cones

Definition 2. The dual of a convex cone K in a vector space V is

$$K^* = \{v \in V : \langle w, v \rangle > 0, \ \forall\, w \in K\}.$$

Question: What is the dual of the convex cone $\mathbb{R}_+$?

Question: What is the dual of the convex cone $\mathbb{R}^n_+$?

Question: What is the dual of the closed upper-half plane of $\mathbb{R}^2$?

Question: What are the duals of the convex cones S^n_+ and S^n_{++}?

Symmetric cones

Definition 3. The **automorphism group** of an open convex cone K in a vector space V is

$$\mathrm{Aut}(K) = \{g \in \mathrm{GL}(V) | gK = K\}.$$

Definition 4. A group G is said to be **transitive** on a set $F \subseteq G$ if F is non-empty and, for each $x, y \in F$, there exists $g \in G$ such that $g \cdot x = y$.

Definition 5. An open convex cone is said to be **homogeneous** if $\mathrm{Aut}(K)$ acts transitively on K.

Definition 6. An open convex cone is said to be **symmetric** if it is homogeneous and self-dual.

Proposition 3. The SPD and HPD cones are symmetric.

Symmetric cones

Definition 3. The **automorphism group** of an open convex cone K in a vector space V is

$$\mathrm{Aut}(K) = \{g \in \mathrm{GL}(V) | gK = K\}.$$

Definition 4. A group G is said to be **transitive** on a set $F \subseteq G$ if F is non-empty and, for each $x, y \in F$, there exists $g \in G$ such that $g \cdot x = y$.

Definition 5. An open convex cone is said to be **homogeneous** if $\mathrm{Aut}(K)$ acts transitively on K.

Definition 6. An open convex cone is said to be **symmetric** if it is homogeneous and self-dual.

Proposition 3. The SPD and HPD cones are symmetric.

Symmetric cones

Definition 3. The **automorphism group** of an open convex cone K in a vector space V is

$$\mathrm{Aut}(K) = \{g \in \mathrm{GL}(V) | gK = K\}.$$

Definition 4. A group G is said to be **transitive** on a set $F \subseteq G$ if F is non-empty and, for each $x, y \in F$, there exists $g \in G$ such that $g \cdot x = y$.

Definition 5. An open convex cone is said to be **homogeneous** if $\mathrm{Aut}(K)$ acts transitively on K.

Definition 6. An open convex cone is said to be **symmetric** if it is homogeneous and self-dual.

Proposition 3. The SPD and HPD cones are symmetric.

Symmetric cones

Definition 3. The **automorphism group** of an open convex cone K in a vector space V is

$$\mathrm{Aut}(K) = \{g \in \mathrm{GL}(V) | gK = K\}.$$

Definition 4. A group G is said to be **transitive** on a set $F \subseteq G$ if F is non-empty and, for each $x, y \in F$, there exists $g \in G$ such that $g \cdot x = y$.

Definition 5. An open convex cone is said to be **homogeneous** if $\mathrm{Aut}(K)$ acts transitively on K.

Definition 6. An open convex cone is said to be **symmetric** if it is homogeneous and self-dual.

Proposition 3. The SPD and HPD cones are symmetric.

Symmetric cones

Definition 3. The **automorphism group** of an open convex cone K in a vector space V is

$$\mathrm{Aut}(K) = \{g \in \mathrm{GL}(V) | gK = K\}.$$

Definition 4. A group G is said to be **transitive** on a set $F \subseteq G$ if F is non-empty and, for each $x, y \in F$, there exists $g \in G$ such that $g \cdot x = y$.

Definition 5. An open convex cone is said to be **homogeneous** if $\mathrm{Aut}(K)$ acts transitively on K.

Definition 6. An open convex cone is said to be **symmetric** if it is homogeneous and self-dual.

Proposition 3. The SPD and HPD cones are symmetric.

Jordan algebra

Definition 7. A **Jordan algebra** J is a nonassociative algebra over a field which, for any $x, y \in J$, and Jordan product $\circ$, satisfies:

- $x \circ y = y \circ x$ [commutativity],
- $(x \circ y)(x \circ x) = x \circ (y \circ (x \circ x))$ [the Jordan identity].

Definition 8. A Jordan algebra J is called **formally real**, or **Euclidean**, if, for $x_1, ..., x_k \in J$, $x_1^2 + ... + x_k^2 = 0$ if and only if $x_j = 0$ for all j.

The standard example of a formally real Euclidean Jordan algebra is the set of Hermitian $n \times n$ matrices equipped with standard addition and the Jordan product:

$$A \circ B = \frac{1}{2}(AB + BA).$$

Proposition 4. Given a member x of a formally real Jordan algebra J, there exists a unique **spectral decomposition**

$$x = \sum_i \lambda_i P_i,$$

where each P_i is **idempotent** – i.e., $P_i^2 = P_i$ – and the λ_i's are real and distinct. See [Faraut/Koranyi-1994] for proof.

Jordan algebra

Definition 7. A **Jordan algebra** J is a nonassociative algebra over a field which, for any $x, y \in J$, and Jordan product $\circ$, satisfies:

- $x \circ y = y \circ x$ [commutativity],
- $(x \circ y)(x \circ x) = x \circ (y \circ (x \circ x))$ [the Jordan identity].

Definition 8. A Jordan algebra J is called **formally real**, or **Euclidean**, if, for $x_1, ..., x_k \in J$, $x_1^2 + ... + x_k^2 = 0$ if and only if $x_j = 0$ for all j.

The standard example of a formally real Euclidean Jordan algebra is the set of Hermitian $n \times n$ matrices equipped with standard addition and the Jordan product:

$$A \circ B = \frac{1}{2}(AB + BA).$$

Proposition 4. Given a member x of a formally real Jordan algebra J, there exists a unique **spectral decomposition**

$$x = \sum_i \lambda_i P_i,$$

where each P_i is **idempotent** – i.e., $P_i^2 = P_i$ – and the λ_i's are real and distinct. See [Faraut/Koranyi-1994] for proof.

Jordan algebra

Definition 7. A **Jordan algebra** J is a nonassociative algebra over a field which, for any $x, y \in J$, and Jordan product $\circ$, satisfies:

- $x \circ y = y \circ x$ [commutativity],
- $(x \circ y)(x \circ x) = x \circ (y \circ (x \circ x))$ [the Jordan identity].

Definition 8. A Jordan algebra J is called **formally real**, or **Euclidean**, if, for $x_1, ..., x_k \in J$, $x_1^2 + ... + x_k^2 = 0$ if and only if $x_j = 0$ for all j.

The standard example of a formally real Euclidean Jordan algebra is the set of Hermitian $n \times n$ matrices equipped with standard addition and the Jordan product:

$$A \circ B = \frac{1}{2}(AB + BA).$$

Proposition 4. Given a member x of a formally real Jordan algebra J, there exists a unique **spectral decomposition**

$$x = \sum_i \lambda_i P_i,$$

where each P_i is **idempotent** – i.e., $P_i^2 = P_i$ – and the λ_i's are real and distinct. See [Faraut/Koranyi-1994] for proof.

Jordan algebra

Definition 7. A **Jordan algebra** J is a nonassociative algebra over a field which, for any $x, y \in J$, and Jordan product $\circ$, satisfies:

- $x \circ y = y \circ x$ [commutativity],
- $(x \circ y)(x \circ x) = x \circ (y \circ (x \circ x))$ [the Jordan identity].

Definition 8. A Jordan algebra J is called **formally real**, or **Euclidean**, if, for $x_1, ..., x_k \in J$, $x_1^2 + ... + x_k^2 = 0$ if and only if $x_j = 0$ for all j.

The standard example of a formally real Euclidean Jordan algebra is the set of Hermitian $n \times n$ matrices equipped with standard addition and the Jordan product:

$$A \circ B = \frac{1}{2}(AB + BA).$$

Proposition 4. Given a member x of a formally real Jordan algebra J, there exists a unique **spectral decomposition**

$$x = \sum_i \lambda_i P_i,$$

where each P_i is **idempotent** – i.e., $P_i^2 = P_i$ – and the λ_i's are real and distinct. See [Faraut/Koranyi-1994] for proof.

Jordan algebra

Proposition 5. Every symmetric cone is the subset of members of a particular Jordan algebra with all positive eigenvalues. Again, see [Faraut/Koranyi-1994] for proof.

This is trivial for the Hermitian positive-definite cone and the Jordan algebra of Hermitian matrices, where the spectral decomposition of the Jordan algebra is the standard one.

This is, by reduction, also trivial for the positive orthant, where the Jordan algebra is the Cartesian product of the 1x1 Hermitian case and the spectral decomposition of a vector $x \in \mathbb{R}^n_{++}$, whose entries take on the values $\Lambda(x) = \{x_i : i = 0, ..., n-1\}$,

$$x = \sum_{\lambda \in \Lambda(x)} \lambda (\sum_{j : x_j = \lambda} e_j).$$

Definition 9. The **second-order cone** of order $n \geq 1$ is the set

$$\mathcal{Q}_n = \{(\chi_0, x_1) \subseteq \mathbb{R}_+ \times \mathbb{R}^{n-1} : \chi_0 \geq \|x_1\|_2\}.$$

There is an associated Jordan algebra, of **spin** or **Clifford** type, $(\chi_0, x_1) \circ (\eta_0, y_1) = (\chi_0 \eta_0 + x_1^T y_1, \chi_0 y_1 + \eta_0 x_1)$, whose positive components yield the second-order cone (which is symmetric).

Jordan algebra

Proposition 5. Every symmetric cone is the subset of members of a particular Jordan algebra with all positive eigenvalues. Again, see [Faraut/Koranyi-1994] for proof.

This is trivial for the Hermitian positive-definite cone and the Jordan algebra of Hermitian matrices, where the spectral decomposition of the Jordan algebra is the standard one.

This is, by reduction, also trivial for the positive orthant, where the Jordan algebra is the Cartesian product of the 1x1 Hermitian case and the spectral decomposition of a vector $x \in \mathbb{R}^n_{++}$, whose entries take on the values $\Lambda(x) = \{x_i : i = 0, ..., n-1\}$,

$$x = \sum_{\lambda \in \Lambda(x)} \lambda (\sum_{j : x_j = \lambda} e_j).$$

Definition 9. The **second-order cone** of order $n \geq 1$ is the set

$$\mathcal{Q}_n = \{(\chi_0, x_1) \subseteq \mathbb{R}_+ \times \mathbb{R}^{n-1} : \chi_0 \geq \|x_1\|_2\}.$$

There is an associated Jordan algebra, of **spin** or **Clifford** type, $(\chi_0, x_1) \circ (\eta_0, y_1) = (\chi_0 \eta_0 + x_1^T y_1, \chi_0 y_1 + \eta_0 x_1)$, whose positive components yield the second-order cone (which is symmetric).

Jordan algebra

Proposition 5. Every symmetric cone is the subset of members of a particular Jordan algebra with all positive eigenvalues. Again, see [Faraut/Koranyi-1994] for proof.

This is trivial for the Hermitian positive-definite cone and the Jordan algebra of Hermitian matrices, where the spectral decomposition of the Jordan algebra is the standard one.

This is, by reduction, also trivial for the positive orthant, where the Jordan algebra is the Cartesian product of the 1x1 Hermitian case and the spectral decomposition of a vector $x \in \mathbb{R}^n_{++}$, whose entries take on the values $\Lambda(x) = \{x_i : i = 0, ..., n-1\}$,

$$x = \sum_{\lambda \in \Lambda(x)} \lambda(\sum_{j : x_j = \lambda} e_j).$$

Definition 9. The **second-order cone** of order $n \geq 1$ is the set

$$\mathcal{Q}_n = \{(\chi_0, x_1) \subseteq \mathbb{R}_+ \times \mathbb{R}^{n-1} : \chi_0 \geq \|x_1\|_2\}.$$

There is an associated Jordan algebra, of **spin** or **Clifford** type, $(\chi_0, x_1) \circ (\eta_0, y_1) = (\chi_0 \eta_0 + x_1^T y_1, \chi_0 y_1 + \eta_0 x_1)$, whose positive components yield the second-order cone (which is symmetric).

Jordan algebra

Proposition 5. Every symmetric cone is the subset of members of a particular Jordan algebra with all positive eigenvalues. Again, see [Faraut/Koranyi-1994] for proof.

This is trivial for the Hermitian positive-definite cone and the Jordan algebra of Hermitian matrices, where the spectral decomposition of the Jordan algebra is the standard one.

This is, by reduction, also trivial for the positive orthant, where the Jordan algebra is the Cartesian product of the 1x1 Hermitian case and the spectral decomposition of a vector $x \in \mathbb{R}^n_{++}$, whose entries take on the values $\Lambda(x) = \{x_i : i = 0, ..., n-1\}$,

$$x = \sum_{\lambda \in \Lambda(x)} \lambda (\sum_{j : x_j = \lambda} e_j).$$

Definition 9. The **second-order cone** of order $n \geq 1$ is the set

$$\mathcal{Q}_n = \{ (\chi_0, x_1) \subseteq \mathbb{R}_+ \times \mathbb{R}^{n-1} : \chi_0 \geq \|x_1\|_2 \}.$$

There is an associated Jordan algebra, of **spin** or **Clifford** type, $(\chi_0, x_1) \circ (\eta_0, y_1) = (\chi_0 \eta_0 + x_1^T y_1, \chi_0 y_1 + \eta_0 x_1)$, whose positive components yield the second-order cone (which is symmetric).

Discussion

These slides are available at:
hodgestar.com/G2S3/

Questions/comments?
Chatroom at:

`https://gitter.im/hodge_star/G2S3`

Lab 2: Word embeddings

1. Generate dictionary keyed on pairs of terms with values equal to the sum of interaction scores within each sentence of the normalized terms – convert all letters to lowercase and replace strings with digits, e.g., “50th“ becomes “DDth“.
2. Compute dictionary keyed on source (target) term with values equal to the sum of interaction terms with said source (target).
3. Truncate down to the top `MAX_SOURCE_TERMS` and `MAX_TARGET_TERMS` source and target terms.
4. Build a sparse matrix on the remaining sources and targets by transforming the interaction sources via $\log(c_{i,j} + 1)$.
5. Perform 10 pairs of iterations of randomly-initialized rank 100 AWLS, printing the objective function after each of the 20 updates.
6. Return the nearest neighbors – via cosine similarity – of: “france”, “music”, “holiday”, “summer”, and “mountain”.