What is a low-rank approximation?

Juan Sebastián Rodríguez Lozano 201815823
Nicolás Torres Caicedo 201822010
Diego Andrés Gómez Polo 201713198

Suponga que la matriz de transición de una cadena de Markov tiene un bajo rango a pesar de tener una gran cantidad de estados. Para este tipo de procesos es posible aproximar de manera precisa el modelo de transición completo a partir de una trayectoria cuyo tamaño es proporcional al número de estados. Basados en el trabajo de Zhu, Li, Wang y Zhang [14] implementamos el primer algoritmo propuesto en este paper e intentamos reproducir sus resultados.

Una cadena de Markov describe una sucesión de eventos, en el cual la probabilidad de cada evento depende únicamente del estado obtenido del estado anterior. Para formalizar esta idea se deben enunciar las siguientes definiciones.

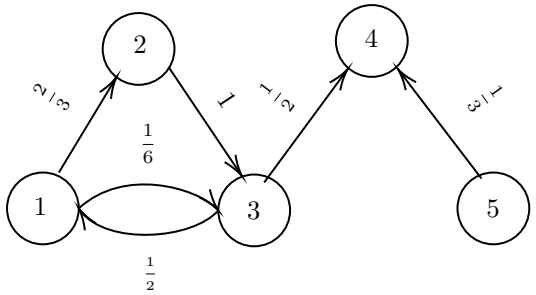
Definición 1. Medida. Sea $I = \{1, \dots, p\}$ un conjunto finito. Cada $i \in I$ se denomina estado y I se denomina espacio de estados. Una distribución sobre I es una tupla $\lambda = (\lambda_i : i \in I)$ de números reales no negativos tales que $\sum_{i \in I} \lambda_i = 1$.

Definición 2. (Cadena de Markov). Una sucesión de variables aleatorias $(X_k)_{k=0}^n$ se dice una cadena de Markov con distribución inicial λ y matriz de transición P si:

- X_0 tiene distribución λ.
- Para todo $k \in \{0, \dots, n-1\}$, condicionado a $X_k = i$, X_{k+1} tiene distribución $(p_{ij} : j \in I)$ y es independiente de $X_0, \dots, X_{k-1}$.

La entrada p_{ij} de la matriz de transición P corresponde a la probabilidad de pasar al estado j, partiendo del estado i en cualquier momento dado.

Ejemplo 3. Considere la siguiente cadena de Markov



Para esta cadena de Markov tenemos que su matriz de transición es la matriz 5×5

$$\begin{bmatrix} \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

En el ejemplo 3 la probabilidad de pasar de un estado a otro no siempre es diferente a 0, por ejemplo, la probabilidad de pasar del estado 2 al estado 4 es igual a 0. En el caso de que la probabilidad de pasar de cualquier estado a cualquier otro estado sea siempre positiva, esto es la entrada de la matriz de transición $p_i j > 0$, decimos que la cadena de Markov es ergódica.

Sea $\mathcal{X} = (X_0, \dots, X_n)$ una cadena de Markov ergódica en el conjunto de estados $\{1, \dots, p\}$, con matriz de transición $P \in \mathbb{R}^{p \times p}$. Contamos la cantidad de veces que se pasa del estado i al estado j en $\mathcal{X}$

$$n_{ij} = |\{1 \le k \le n : X_{k-1} = i \wedge X_k = j\}|$$

Con esto definimos la siguiente función la función de log-verosimilitud negativa promediada

$$\ell_n(P) = -\frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{p} n_{ij} \ln(P_{ij}).$$

donde $n_i = \sum_{j=1}^{p} n_{ij}$ y $n = \sum_{i=1}^{p} n_i$. Una vez tenemos esta función el estimador propuesto en [14] para aproximar la matriz de transición, es el estimador de máxima verosimilitud (MLE) para la norma nuclear regularizada el cual se define como la solución del problema de optimización convexo

$$\hat{P} = \arg\min \ell_n(Q) + \lambda \|Q\|_* \qquad (1)$$
$$s.a.\ \ Q1_p = 1_p \quad \frac{\alpha}{p} \le Q_{ij} \le \frac{\beta}{p}.$$

donde 1_p es el vector de unos de dimensión p y α y β son constantes no negativas y $\lambda > 0$ es un parámetro de penalización.

Nota: Para ver sobre las garantías estadísticas de este estimador ver Teorema 1 [14].

El problema de calcular el MLE (1), tomando $\alpha = 0, \beta = p, g(X) = -\ell_n(X) + \delta(X \ge 0), \mathcal{A}(X) = X1_p$ y $b = 1_p$, se puede ver como un caso particular de la familia de problemas de optimización

$$\min\{g(X) + c\|X\|_* : \mathcal{A}(X) = b\} \qquad (2)$$

donde $g : \mathbb{R}^{p \times p} \to (-\infty, \infty]$ es una función convexa y cerrada, $\mathcal{A} : \mathbb{R}^{p \times p} \to \mathbb{R}^m$ lineal, $b \in \mathbb{R}^m$ y $c > 0$.

A pesar de la convexidad del problema su solución es bastante no trivial, sobre todo por la falta de suavidad de la función g y la presencia de la norma nuclear. Por lo tanto, se recurre el la formulación dual de (1) (ver Apéndice B)

$$\min\ g^*(\Xi) - \langle b, y \rangle \qquad (3)$$
$$s.a.\ \ \Xi + \mathcal{A}^*(y) + S = 0 \quad \|S\|_2 \le c$$

donde $\|\cdot\|_2$ es la norma espectral y g^* es la función conjugada de g dada por

$$g^*(\Xi) = \sum_{(i,j)\in\Omega} \frac{n_{ij}}{n}(\ln\frac{n_{ij}}{n} - 1 - \ln(-\Xi_{ij})) + \delta(\Xi \le 0),$$

con $\Omega = \{(i,j) : n_{ij} \neq 0\}$.

El Lagrangiano aumentado dado $\sigma > 0$ (ver Apéndice A) asociado al problema (3) es:

$$\mathcal{L}_\sigma(\Xi, y, S; X) = g^*(-\Xi) - \langle b, y\rangle - \frac{1}{2\sigma}\|X\|^2 + \frac{\sigma}{2}\left\|\Xi + \mathcal{A}^*(y) + S + X/\sigma\right\|^2 \quad (4)$$

Dado que hay tres bloques separables en (3), más específicamente Ξ, y y S, el ADMM directo extendido no es aplicable. Como se muestra en [3] ADMM directo extendido no es necesariamente convergente para problemas de minimización convexa con multibloques. Sin embargo, como las funciones correspondientes al bloque y son lineales es posible aplicar el método Gauss-Sediel simétrico multibloque basado en ADMM (sGS-ADMM) [9].

Este es el algoritmo que implementamos y se expone en la sección subsiguiente.

1. Algoritmo

Tomamos un punto inicial $(\Xi^0,\ y^0,\ S^0,\ X^0)$, un parámetro de penalización $\sigma > 0$, un máximo número de iteración K y una longitud de paso $\gamma \in (0, \frac{1+\sqrt{5}}{2})$. Para $0 < k < K$ definimos

$$\begin{aligned}
y^{k+\frac{1}{2}} &= \arg\min_y \mathcal{L}_\sigma(\Xi^k, y, S^k; X^k)\\
\Xi^{k+1} &= \arg\min_\Xi \mathcal{L}_\sigma(\Xi, y^{k+\frac{1}{2}}, S^k; X^k)\\
y^{k+1} &= \arg\min_y \mathcal{L}_\sigma(\Xi^{k+1}, y, S^k; X^k)\\
S^{k+1} &= \arg\min_S \mathcal{L}_\sigma(\Xi^{k+1}, y^{k+1}, S; X^k)\\
X^{k+1} &= X^k + \gamma\sigma(\Xi^{k+1} + \mathcal{A}^*(y^{k+1}) + S^{k+1}).
\end{aligned}$$

Aunque las expresiones para las iteraciones de Ξ, y, S, X parecen complicadas en todos los casos se puede obtener una fórmula explícita de la k-ésima iteración. En los subsecuentes cálculos utilizaremos repetidamente el truco de ignorar los términos que no involucren la variable a minimizar.

Para calcular las iteraciones de y basta considerar la función

$$T(y) = \frac{\sigma}{2}\left\|A^*(y) + K\right\|^2 - \langle b, y\rangle,$$

donde $K = \Xi + S + \frac{X}{\sigma}$. Considere la función

$$\begin{aligned}
f(x) &= \langle Ax + c, Ax + c\rangle\\
&= \langle Ax, Ax\rangle + 2\langle Ax, c\rangle + \|c\|^2.
\end{aligned}$$

Note que

$$(x+h)^t A^t A(x+h) = x^t A^t Ax + 2(A^t Ax)^t h + \epsilon(h)$$

donde $\epsilon(h) = h^t A^t Ah$ y como $\langle Ax, c\rangle = (Ac)^t x$, se sigue que:

$$\nabla f(x) = 2(A^t Ax + A^t c)$$

Por lo anterior, tenemos que el gradiente de T es:

$$\nabla T(y) = \sigma(\mathcal{A}\mathcal{A}^* y + \mathcal{A}K) - b.$$

Para obtener el óptimo, igualamos este gradiente a 0. Con esto el óptimo es $y = (\mathcal{A}\mathcal{A}^*)^{-1}(\frac{b}{\sigma} - \mathcal{A}K)$. En conclusión

$$\begin{aligned}
y^{k+1} &= \frac{1}{\sigma}(\mathcal{A}\mathcal{A}^*)^{-1}(b - \mathcal{A}(X^k - \sigma(\Xi^{k+1} + S^k))),\\
y^{k+\frac{1}{2}} &= \frac{1}{\sigma}(\mathcal{A}\mathcal{A}^*)^{-1}(b - \mathcal{A}(X^k - \sigma(\Xi^k + S^k)))
\end{aligned}$$

Además se puede probar que $\mathcal{A}\mathcal{A}^* y = py$.

Las iteraciones de Ξ se calculan resolviendo este problema de optimización:

$$\min_\Xi\{g^*(-\Xi) + \frac{\sigma}{2}\left\|\Xi + R^k\right\|^2\},$$

donde $R^k = \mathcal{A}^*(y^{k+\frac{1}{2}} + S^k + X^k/\sigma)$

Teorema 4. *(Identidad de Moreau, Teorema 31.5 [13]) Sea f un función propia convexa cerrada sobre $\mathbb{R}^n$. Entonces*

$$\inf_x\{f(x) + \frac{1}{2}\|z - x\|^2\} + \inf_{x^*}\{f^*(x^*) + \frac{1}{2}\left\|z - x^*\right\|\} = \frac{1}{2}\|z\|^2$$

donde ambos ínfimos son finitos, únicos y se alcanzan. Los únicos vectores x y x^* en los cuales se alcanzan los ínfimos respectivamente para un z dado son lo únicos tales que

$$x + x^* = z \qquad x^* \in \partial f(x)$$

y están dados por

$$\begin{aligned}
x &= \nabla \inf_x\{f(x) + \frac{1}{2}\|z - x\|^2\}\\
x^* &= \nabla \inf_{x^*}\{f^*(x^*) + \frac{1}{2}\left\|z - x^*\right\|\}.
\end{aligned}$$

Utilizando la identidad de Moreau tenemos que

$$\Xi^{k+1} = \frac{1}{\sigma}(Z^k - \sigma R^k)$$

donde

$$Z^k = \min_Z\{\sigma g(Z) + \frac{1}{2}\left\|Z - \sigma R^k\right\|^2\}$$

para el cual se puede demostrar que

$$Z^k_{ij} = \begin{cases} \frac{\sigma R^k_{ij} + \sigma\sqrt{(R^k_{ij} + 4n_{ij}/(n\sigma))^2}}{2} & (i,j) \in \Omega\\ \sigma\max(R^k_{ij}, 0) & (i,j) \notin \Omega \end{cases}$$

Las iteraciones de S se pueden calcular resolviendo el problema de optimización

$$\arg\min\{\frac{\sigma}{2}\left\|S - W^k\right\| : \|S\|_2 \le c\}$$

donde $W^k = -(\Xi^{k+1} + \mathcal{A}^*(y^{k+1}) + X^k/\sigma)$.

Lema 5. *(Lema 2.1 [8]). Sea $Y \in \mathbb{R}^{n\times m}$ y $Y = U\Sigma_Y V^t$ su SVD. Entonces la única solución del problema*

$$arg\,\text{mín}\{\|X - Y\|^2 : \|X\|_2 \leq \rho\}$$

es $\hat{X} = U\,\text{mín}(\Sigma_Y, \rho)V^t$, *donde* $\text{mín}(\Sigma_Y, \rho) = \text{Diag}(\text{mín}\{\sigma_1, \rho\}, ..., \text{mín}\{\sigma_p, \rho\})$.

Dem: Claramente si el problema tiene una solución esta será única. Defina

$$A_1 = \{i : \sigma_i(Y) > \rho\} \text{ y } A_2 = \{i : \sigma_i(y) \leq \rho\}$$

Sea Z un valor factible con SVD $Z = U_1\Sigma_Z V_1^*$, por suposición tenemos que $\sigma_i(Z) \leq \rho$ para $1 \leq i \leq p$. Entonces por (Ejercicio IV.3.5,[1]) como $\|\cdot\|$ es unitariamente invariante tenemos

$$\begin{aligned}\|Z - Y\|^2 &\geq \|\Sigma_Z - \Sigma_Y\| \\ &\geq \sum_{i\in A_1}(\sigma_i(Y) - \sigma_i(Z)) + \sum_{i\in A_2}(\sigma_i(Y) - \Sigma_i(Z)) \\ &\geq \sum_{i\in A_1}(\sigma_i(Y) - \rho) = \left\|Y - \hat{X}\right\|^2\end{aligned}$$

■

Por el Lema 5 tenemos que la expresión explícita para S^{k+1} es:

$$S^{k+1} = U_k\,\text{mín}(\Sigma_k, c)V_k^t$$

donde $W_k = U_k\Sigma_k V_k^t$ y $\text{mín}(\Sigma_k, c) = \text{Diag}(\text{mín}(\alpha_1, c), ..., \text{mín}(\alpha_n, c))$ con $\alpha_1 \geq ... \geq \alpha_n$ son los valores singulares de W_k.

En general la solución del problema dual sólo da un cota inferior al problema de optimización. Sin embargo, en este caso las parejas de soluciones primal-dual asociadas a los problemas (1) y (3) satisfaces el sistema KKT (Apéndice B)

$$0 \in R(X, \Xi, S), \quad \mathcal{A}(X) = b, \quad \Xi + \mathcal{A}^*(y) + S = 0$$

con

$$R(X, \Xi, S) = \begin{bmatrix}\Xi + \partial g(X) \\ X + \partial\delta(\|S\|_2 \leq c)\end{bmatrix}$$

$$(X, \Xi, S) \in \text{dom}(g) \times \mathbb{R}^{p\times p} \times \{S \in \mathbb{R}^{p\times p} : \|S\|_2 \leq c\}$$

Además el siguiente resultado garantiza que el Algoritmo aproxima una solución tanto del problema dual como del problema primal.

Teorema 6. *(Teorema EC.3 [14]) Suponga que el conjunto de soluciones de* (1) *y* (3) *son no vacíos. Sea* $\{(\Xi^k, y^k, S^k, X^k)\}_{k\in\omega}$ *la sucesión generada en el Algoritmo. Si* $\gamma \in (0, (1 + \sqrt{5})/2)$ *entonces la sucesión* $\{(\Xi^k, y^k, S^k)\}_{k\in\omega}$ *converge a una solución óptima de* (3) *y* $\{X^k\}_{k\in\omega}$ *converge a una solución óptima de* (1).

2. Resultados

Para finalizar este proyecto hicimos pruebas numéricas y simulaciones con el fin de verificar el desempeño práctico del algoritmo antes descrito. La implementación del algoritmo se puede encontrar en el siguiente repositorio **github: diegommezp28/low_rank_opti_for_markov_chains**

El primer paso fue generar una matriz de transición aleatoria con dimensión y rango deseado. Para esto se tomó como dimensión deseada del espacio de estados $p = 1000$ y rango $r = 10$. Luego, de forma aleatoria, se generaron dos matrices U y V. Cada una con entradas positivas entre 0 y 1 tales que la suma de las filas fuese 1. Además, $U \in \mathbb{R}^{pxr}$ y $V \in \mathbb{R}^{rxp}$. Luego, $M = U \cdot P$ sería nuestra matriz de transición de dimensión pxp y rango r.

Esta matriz generada tiene probabilidad 0 de tener una entrada en 0, por ende, es ergódica con probabilidad 1. En la Figura 2, se muestra el heatmap de la matriz M. Se usó una semilla en Python para controlar la aleatoriedad y trabajar siempre con la misma matriz M, esto con el fin de que resultados para distintos parámetros del algoritmo fuesen comparables.

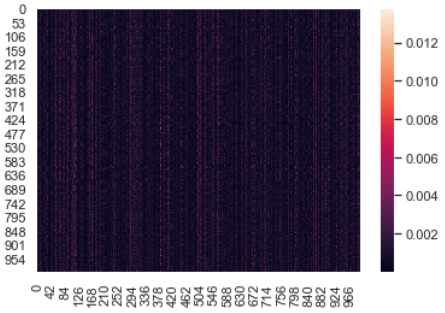


Figura 1: Heatmap matriz de transición original.

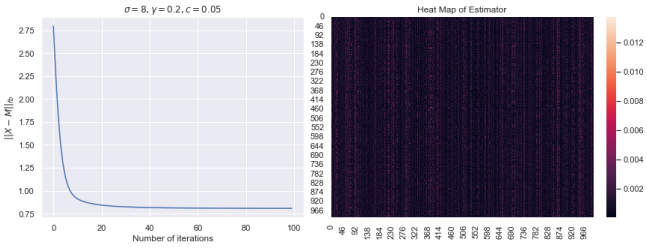


Figura 2: Gráfico de convergencia en norma de Frobenius (izquierda) y Heatmap estimador (derecha).

Luego de esto, se simuló una caminata aleatoria en el espacio de estados siguiendo las probabilidades de transición de M. Probamos con dos tamaños para esta caminata, el primero fue de $n = p^2$ y $n = 10rp\log p$. En ambos casos encontramos un buen comportamiento de la convergencia, en particular, en la Figura 2 se muestra el caso con $n = 10rp\log p$, los demás casos se pueden ver en el código fuente contenido en el repositorio del proyecto.

Cabe destacar que el punto inicial del algoritmo lo escogimos como el estimador de máxima verosimilitud de la matriz M basado en los datos de la caminata aleatoria de tamaño n. La motivación detrás de esto era acelerar un poco la convergencia del algoritmo.

Después de múltiples intentos con parámetros distintos encontramos que el mejor comportamiento se daba cuando la penalidad σ era, por lo menos, dos ordenes de

magnitud más grande que la constante c de regularización de la norma nuclear.

Incluso en las pruebas hechas para las cuales c y σ estaban cercanas, si bien la convergencia no fue tan buena y el algoritmo tendía a buscar una matriz de rango 1, de igual forma la norma de Frobenius de la diferencia entre el estimador y la matriz M siempre estuvo con valores menores a 1.

A. Lagrangiano aumentado

Sea $g : \mathbb{R}^n \to \mathbb{R}$ una función convexa, $A \in \mathbb{R}^{m\times n}$ y $b \in \mathbb{R}^m$. Considere el problema de optimización con restricciones

$$\begin{aligned} &\text{mín } g(x) \\ &s.t.\ Ax = b \end{aligned}$$

Definimos el Lagrangiano asociado a este problema como

$$\mathcal{L}(x;y) = g(x) + \langle y, Ax - b\rangle$$

y para $\sigma > 0$ definimos el Lagrangiano aumentado

$$\mathcal{L}_\sigma(x;y) = g(x) + \langle y, Ax - b\rangle + \frac{\sigma}{2}\|Ax - b\|^2$$

Nota: Los métodos para resolver problemas de optimización a través del Lagrangiano aumentado fueron introducidos por primera vez por Hestenes [7] y Powel [12].
Alternativamente se puede presentar el Lagrangiano aumentado como

$$\mathcal{L}_\sigma(x;y) = g(x) + \frac{\sigma}{2}\|Ax - b + y/\sigma\| - \frac{1}{2\sigma}\|y\|^2$$

con un par de cálculos sencillos se puede ver que ambas presentaciones son equivalentes. Note que si tomamos $y = 0$

$$\mathcal{L}_\sigma(x;0) = g(x) + \frac{\sigma}{2}\|Ax - b\|^2$$

obtenemos el funcional de penalización clásico sujeto a la restricción $Ax = b$.
La ventaja de usar el Lagrangiano aumentado es que, gracias a la presencia del término $\langle y, Ax - b\rangle$, la solución exacta del problema de optimización se puede hallar sin necesidad de mandar σ al infinito, contrario a lo que ocurre con la penalización clásica la cual ocasiona un deterioro en el condicionamiento del sistema a resolver.

Teorema 7. *(Teorema 2.1 [6]) Un punto (u, λ) es un punto de silla de $\mathcal{L}$ si y sólo si es un punto de silla de $\mathcal{L}_\sigma$ para todo $\sigma > 0$. Además u es una solución del problema de optimización y $Au = b$.*

B. Dual y condiciones KKT

Sea g una función $g : \mathbb{R}^n \to \overline{\mathbb{R}}$. Definimos su conjugado dual g^* por

$$g^* : \mathbb{R}^n \to \overline{\mathbb{R}},$$

$$g^*(x) = \sup_{v\in\mathbb{R}^n}\{\langle v, x\rangle - g(v)\}.$$

Para una función $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ dada, considere el siguiente problema de optimización:

$$\min_{x\in\mathbb{R}^n} g(x).$$

Con el lagrangiano $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^k$ obtenemos la siguiente familia de problemas de optimización, uno para cada $y \in \mathbb{R}^k$:

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(x, y).$$

Considere el problema de optimización

$$\min_{x\in\mathbb{R}^n} f(x)$$

sujeto a

$$h_i(x) \le 0, i = 1, \dots, m,$$

$$l_j(x) = 0, j = 1, \dots, r.$$

Para que la solución del problema dual coincida con la solución del problema original, es necesario y suficiente que se cumplan las condiciones de Karush-Kuhn-Tucker:

- $0 \in \partial f(x) + \sum_{i=1}^m u_i \partial h_i(x) + \sum_{j=1}^r v_j \partial l_j(x)$.
- $u_i \cdot h_i(x) = 0$ para i.
- $h_i(x) \le 0$, $l_j(x) = 0$ para todo i, j.
- $u_i \ge 0$ para todo i.

Estas condiciones se conocen como estacionariedad, holgura complementaria, factibilidad primal y factibilidad dual respectivamente.

Teorema 8. *Si x^*, u^* y v^* satisfacen las condiciones KKT entonces x^* y u^*, v^* son soluciones del primal y el dual respectivamente.*

En resumen, las condiciones KKT son un criterio suficiente para garantizar que la brecha dual es igual a 0.

Referencias

[1] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[2] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[3] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Math. Program.*, 155(1-2, Ser. A):57–79, 2016.

[4] K. Deng and D. Huang. Model reduction of markov chains via low-rank approximation. In *2012 American Control Conference (ACC)*, pages 2651–2656. IEEE, 2012.

[5] M. Fortin and R. Glowinski. *Augmented Lagrangian methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983. Applications to the numerical solution of boundary value problems, Translated from the French by B. Hunt and D. C. Spicer.

[6] R. Glowinski. *Numerical methods for nonlinear variational problems*. Scientific Computation. Springer-Verlag, Berlin, 2008. Reprint of the 1984 original.

[7] M. R. Hestenes. Multiplier and gradient methods. *J. Optim. Theory Appl.*, 4:303–320, 1969.

[8] K. Jiang, D. Sun, and K.-C. Toh. A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. *Math. Program. Comput.*, 6(3):281–325, 2014.

[9] X. Li, D. Sun, and K.-C. Toh. A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. *Math. Program.*, 155(1-2, Ser. A):333–373, 2016.

[10] D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, Inc., New York-London-Sydney, 1969.

[11] J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

[12] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In *Optimization (Sympos., Univ. Keele, Keele, 1968)*, pages 283–298. Academic Press, London, 1969.

[13] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[14] Z. Zhu, X. Li, M. Wang, and A. Zhang. Learning markov models via low-rank optimization. *Operations Research*, 2021.