Analysis of Embedding Space of Speech Synthesis System as High-Dimension, Low-Sample-Size Data

Yuuki Tachioka
Denso IT Laboratory Tokyo, Japan
tachioka.yuki@core.d-itlab.co.jp

***Abstract*—Principal component analysis (PCA) is often used to visualize and cluster embedding vectors of deep-learning models; however large noise can degrade the results of PCA when the dimension of the embedding vectors is much larger than the data size. These types of data are referred to as high-dimension, low-sample-size (HDSS) data, where the number of data dimension is much larger than the sample size, and they must be treated differently than typical data, because the eigenvalues of a HDSS data covariance matrix are impacted by large noise. Previous studies on HDSS data proposed the noise reduction methodology (NRM) and cross-data matrix methodology (CDM) to reduce such effects. In this study, we apply NRM and CDM to visualize the embedding vectors of an end-to-end speech synthesis system. Experimental results show that the NRM can estimate reduced eigenvalues by eliminating the large noise and that the CDM can provide a more reasonable visualization, although the results are dependent on the initial CDM clusters. In addition, we propose a method for estimating the power exponent of a general spiked model to assess whether the estimated eigenvalues are consistent.**

***Index Terms*—noise reduction methodology, cross-data matrix methodology, general spiked model, embedding vector**

I. Introduction

Embedding vectors for deep-learning-based methods include important information which needs to be visualized to understand a model's behavior. Because embedding vectors are high-dimensional and it is difficult to visualize their characteristics, they are typically converted into two- or three-dimensional vectors by principal component analysis (PCA) [1]. For speech synthesis systems that can control speaker and style, it is important to visualize embedding vectors and capture the difference between embedding vectors in terms of speakers or styles. For example, the dimension of embedding vectors d is around 2^9 and the number of speakers or styles n is around 2^2–2^4. In this case, d is much greater than the sample size n. In the field of statistics, these types of data ($d \gg n$) are referred to as high-dimension, low-sample-size (HDSS) data, which cannot be treated as typical data [2] for PCA because the eigenvalues of sample covariance matrices are impacted by large noise [3]–[5].

To reduce this effect and estimate consistent eigenvalues of HDSS data, previous studies have proposed the noise reduction methodology (NRM) [2] and the cross-data-matrix methodology (CDM) [6]. In this paper, we compare PCA with these two methodologies for visualizing embedding vectors of speech synthesis systems. On the other hand, the papers [1], [7] have shown that the PCA scores are consistent for HDSS data under certain conditions, which means that for certain types of HDSS data, large noise does not impact the results. A general spiked model can model the eigenvalues of sample covariance matrices of HDSS data, and whether or not the eigenvalues are consistent can be known from a parameter of the general spiked model [2]. For given data, we propose a method to estimate this parameter to determine whether the estimated eigenvalues are consistent.

II. How to deal with HDSS data

A. Dual space

Data matrix $\boldsymbol{X}(\in \mathbb{R}^{d\times n})$ is composed of d-dimensional vectors with n observations. To estimate the eigenvalues of HDSS data, instead of a large-size sample covariance matrix

$$\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^\top \ (\in \mathbb{R}^{d\times d}), \tag{1}$$

a small-size dual sample covariance matrix

$$\boldsymbol{S}_D = \boldsymbol{X}^\top\boldsymbol{X} \ (\in \mathbb{R}^{n\times n}), \tag{2}$$

is used, where $\top$ is a transpose. In this case, the first n-th eigenvalues of $\boldsymbol{S}$ are equal to the eigenvalues of $\boldsymbol{S}_D$. In dual space, if the contribution of the maximum eigenvalue of $\boldsymbol{S}_D$ converges to 0 at $d \to \infty$, the eigenvalues are consistent. If consistent, $\boldsymbol{S}_D$ converges to the surface of a sphere and the eigenvalues are determined. If not, $\boldsymbol{S}_D$ converges to the axis and the eigenvalues and eigenvectors are not fixed.

B. Assumption of mutual independence of $\boldsymbol{Z}$

After the matrix

$$\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{H}^\top\boldsymbol{X}, \tag{3}$$

is calculated, if d row vectors of $\boldsymbol{Z}$ are mutually independent, NRM in II-C is used; otherwise, CDM in II-D is used. Here, $\boldsymbol{\Lambda}$ and $\boldsymbol{H}$ are the results of the eigenvalue decomposition, which is applied to $\bar{\boldsymbol{S}}$ as $\bar{\boldsymbol{S}} = \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^\top$. Before the eigenvalue decomposition, the centralization where the average over the column is subtracted from the data matrix is applied to $\boldsymbol{X}$ as

$$\bar{\boldsymbol{X}} = \boldsymbol{X} - \frac{1}{n}\sum_{\nu}\boldsymbol{X}[:,\nu] \ (\in \mathbb{R}^d). \tag{4}$$

979-8-3503-4017-4/23/$31.00 ©2023 IEEE

C. NRM [2]

If the assumption in II-B is satisfied, the noise is consistent and converges to the surface of a sphere. The objective of the NRM is to eliminate this noise. In the HDSS settings, sample eigenvalues $\lambda = [\lambda_1, \lambda_2, ...]$ are overly estimated due to the effect of noise. The NRM reduces $s(= 1, ..., n-2)$-th eigenvalues, λ_s, as

$$\tilde{\lambda}_s = \lambda_s - \frac{\mathrm{tr}\,(\boldsymbol{S}_D) - \sum_{s'=1}^{s} \lambda_{s'}}{n-1-s}. \tag{5}$$

D. CDM [6]

The CDM is used to estimate consistent eigenvalues when the assumption in II-B is not satisfied. The CDM divides a column index set of the data matrix $\boldsymbol{X}$, $\mathcal{N} = \{1, ..., n\}$, into disjoint two sets $\mathcal{N}_1$ and $\mathcal{N}_2$ ($\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2$, $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, and $|\mathcal{N}_1| - |\mathcal{N}_2| \in \{0, 1\}$). This gives two distinct data matrices, $\boldsymbol{X}_1 = \boldsymbol{X}[:, \nu \in \mathcal{N}_1]$ and $\boldsymbol{X}_2 = \boldsymbol{X}[:, \nu \in \mathcal{N}_2]$. After the centralization of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, which subtracts the mean from both matrices, cross-data matrices

$$\begin{aligned} \boldsymbol{S}_D^1 &= \frac{\boldsymbol{X}_1^\top \boldsymbol{X}_2}{\sqrt{(|\mathcal{N}_1| - 1)(|\mathcal{N}_2| - 1)}}, \\ \boldsymbol{S}_D^2 &= (\boldsymbol{S}_D^1)^\top, \end{aligned} \tag{6}$$

are obtained. Because $\boldsymbol{S}_D^1$ is a symmetric matrix, singular value decomposition of $\boldsymbol{S}_D^1$ yields eigenvalues $\acute{\lambda}_s (s = 1, ..., |\mathcal{N}_2| - 1)$.

III. General spiked model and its parameter estimation method

A general spiked model [2] can be applied to HDSS data in which a few eigenvalues among the eigenvalues of a sample covariance matrix are spiked, i.e., when a few eigenvalues are much greater than the others [2], [5], [8]. For eigenvalues of $\boldsymbol{S}$, $\lambda_s (s = 1, ..., d)$, a general spiked model assumes that the first m eigenvalues decay exponentially.

$$\lambda_s = \begin{cases} c_s d^{\alpha_s} & (s = 1, ..., m), \\ c_s & (s = m+1, ..., d), \end{cases} \tag{7}$$

where $c_s (> 0)$ and α_s are unknown constants preserving the order $1 > \alpha_{s'} > \alpha_{s''} > 0$ $(s' < s'')$ and α_s plays a key role in judging whether sample eigenvalues are consistent or not. For PCA, if this model can be applied, sample eigenvalues λ are consistent when

$$\gamma \geq 1 - \alpha_s, \tag{8}$$

where γ is $\log_d n$. For the NRM, if

$$\gamma \geq 1 - 2\alpha_s, \tag{9}$$

is satisfied, the eigenvalues $\tilde{\lambda}$ are consistent. The consistent region is larger than that of PCA. For the CDM, eigenvalues $\acute{\lambda}$ are consistent when

$$\alpha_s > \frac{1}{2} \vee \gamma \geq 1 - \alpha_s, \tag{10}$$

is satisfied.

We estimate a power exponent α_s in Eq. (7) as follows. Normalized data matrices $\sqrt{d/\mathrm{tr}\,(\boldsymbol{S}_n)}\boldsymbol{X}$ are used where the sum of eigenvalues are d and tr is a trace. We take a subset $\mathcal{T}'$ from a row index set $\mathcal{T} = \{1, ..., d\}$ of $\boldsymbol{X}$, where $\mathcal{T}' \subset \mathcal{T}$ and $|\mathcal{T}'| = d' < |\mathcal{T}| = d$. In this case, the division of both hand sides of Eq. (7) can remove the common constants c_s, giving

$$\frac{\lambda'_s}{\lambda_s} = \left(\frac{d'}{d}\right)^{\alpha_s}, \tag{11}$$

where λ'_s is an eigenvalue of the sample covariance of $\boldsymbol{X}[\tau \in \mathcal{T}', :]$. Because d' can be arbitrarily set, α_s can be estimated by the power approximation of multiple sampling of $\mathcal{T}'$ to calculate λ'_s/λ_s.

IV. Experiments

A. Experimental condition

We used Tacotron2 as the speech synthesis system, which receives five characters for encoders, obtains $512(= d)$-dimensional embedding vectors, and estimates Mel-spectrograms using an auto-regressive decoder [9]. Then the Waveglow model [10] generates waveforms from the estimated Mel-spectrogram. The sampling frequency was 22.5 kHz. To control Tacotron2, we added speaker and style tags [11] before the characters of the utterance. For training, four professional narrators (two male and two female) read out scripted 2,456 utterances and ten narrators (five male and five female) read out 600 utterances in four styles. For evaluation, we used 30 utterances for each style, which were different from training data. Style tags were single alphabet characters corresponding to each style (neutral (h), sad (k), joyful (t), or angry (z)). Speaker tags consisted of four letters, where the initial character indicated sex (female (f) and male (m)) and the other three characters were the narrator's name. Speech was synthesized for 20 speakers. In addition to the above 14 training speakers, speech by six additional speakers (three male and three female) were synthesized by using randomly generated speaker tags.

Embedded vectors were converted to two-dimensional vectors by PCA, NRM, and CDM. The figures show the eigenvalues and their cumulative ones. Publicly available codes[1] were used for the NRM and CDM. We prepared three $\mathcal{N}$ divisions for the CDM. In addition, for the first six eigenvalues $(s = 1, ..., 6)$, the proposed method presented in III was used to estimate α_s by sampling $\mathcal{T}'$ 500 times.

B. Results and discussion

1) Style: Fig. 1 shows the PCA results of embedding vectors of four styles and two female speakers (ftak and fhar) $(n = 4 \times 2)$, which indicate that embedding space was divided by speakers and that both orders of the styles {h,t,z,k} were the same. Fig. 2 shows the results of the NRM. The first eigenvalue was reduced but the trends were similar to that of PCA. Fig. 3 shows the results of the CDM, which contained three $\mathcal{N}$ divisions. The elements in $\mathcal{N}_1$ were indicated by △. In the top figure, the $\mathcal{N}$ division by speaker can cluster the

[1] http://www.math.tsukuba.ac.jp/~aoshima-lab/jp/Rcode.html

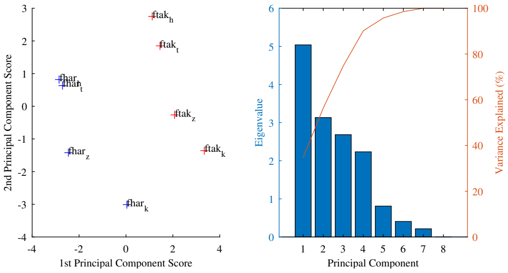

Fig. 1. Embedding vectors estimated by PCA in terms of style.

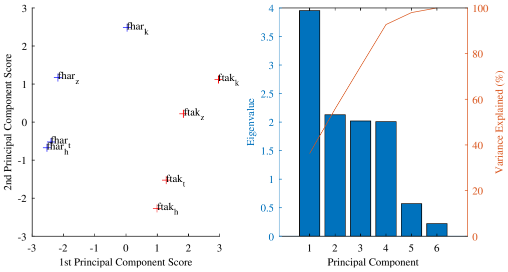

Fig. 2. Embedding vectors estimated by NRM in terms of style.

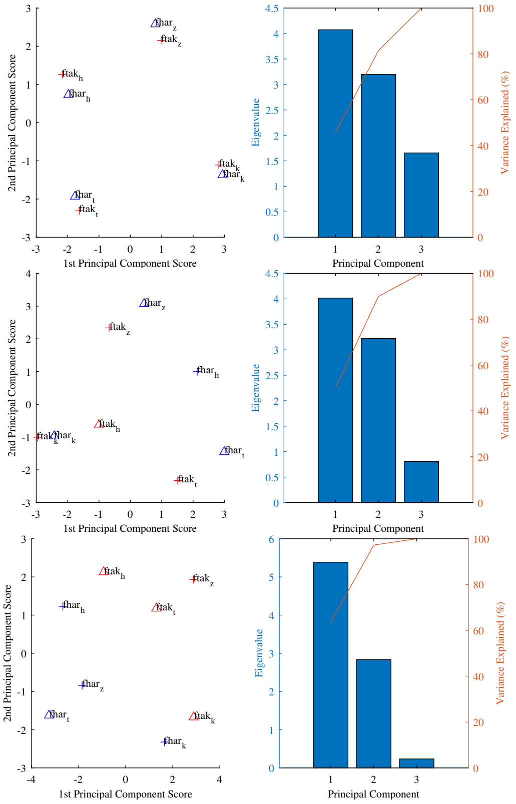

Fig. 3. Embedding vectors estimated by CDM in terms of style when three $\mathcal{N}$ divisions were used.

embedding vectors of styles. However, in the middle figure where one element is changed between $\mathcal{N}_1$ and $\mathcal{N}_2$ (‘fhar$_\mathrm{h}$’ $\leftrightarrow$‘ftak$_\mathrm{h}$’), the ‘h’ cluster disappeared. This indicates that the results of the CDM are dependent on the initial division of $\mathcal{N}$. In the bottom figure, $\mathcal{N}_1$ included the style ‘t’ twice and did not include style ‘z’. Clusters of ‘t’ and ‘z’ were mixed, which indicates that the division which included the elements with different properties degraded the results of the CDM.

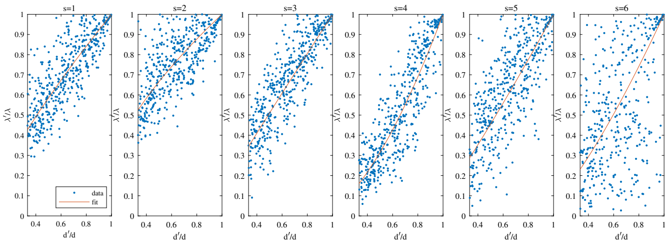

Fig. 4. Relation between the ratio of λ and that of the dimension.

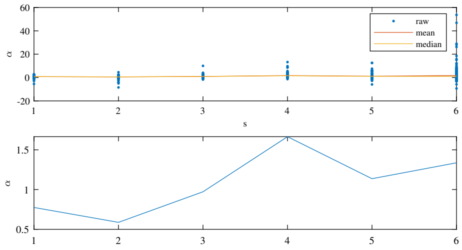

Fig. 5. Estimated α_s. Top: point estimation results; bottom: power approximation results.

To estimate the power exponent α_s in Eq. (7), we first plot $y = \lambda'/\lambda$ and $x = d'/d$ as shown in Fig. 4. As the figure shows, the eigenvalues increased with data dimension. The red lines are the power approximation ($y = x^{\alpha_s}$) for each s. The top figure in Fig. 5 shows that the mean and median values of α_s from a point estimation of $\alpha_s = \log(y)/\log(x)$ have outliers but converged. The bottom figure shows α_s estimated by power approximation. Except for $s = 2$, α_s was greater than 0.8. In Eq. (8), if $\gamma \geq 0.2$, sample eigenvalues are consistent. In this case, γ is $3/9 \simeq 0.33$, which satisfies the condition that the eigenvalues are consistent.

2) Speaker: The PCA results of embedding vectors from changing the speaker tags and keeping style tag ‘h’ are shown in Fig. 6. In addition to the 14 training speakers (seven male and seven female) plotted with ‘+’, six additional speakers (three male and three female) are plotted with ‘x’ ($n = 20$). The results show that the sign of the first principal component scores clearly correspond to the male and female speakers, where the first character indicates sex. The results of the NRM exhibit similar trends as shown in Fig. 7. The results of the CDM are shown in Fig. 8, where the elements of $\mathcal{N}_1$ are indicated by $\triangle$ in the case of ‘+’ and $\bigtriangledown$ in the case of ‘x’. In the top figure, where $\mathcal{N}_1$ only includes female speakers, the sign did not distinguish sex. In contrast, the random divisions shown in the middle and bottom figures made it possible to distinguish sex by the sign of the first principal component scores, demonstrating that random divisions could distinguish speakers more precisely than sex-dependent division.

We plot $y = \lambda'/\lambda$ and $x = d'/d$ in Fig. 9. The estimation result of the power exponent α_s in Eq. (7) is shown in Fig. 10. In this case, $\alpha_s > 0.6$ and parameter $\gamma \simeq 0.48$, so the sample eigenvalues are consistent.

V. Conclusion

Analyzing the embedding vectors of speech synthesis system poses a challenge due to the impact of large noise on HDSS data. To address this, we applied the NRM and CDM

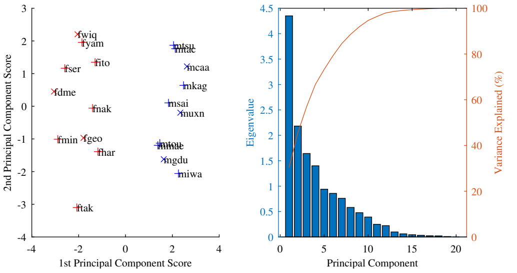

Fig. 6. Embedding vectors estimated by PCA in terms of speaker. ‘+’: speakers in the training set; ‘x’: speakers who do not exist in the training set (such as fwiq).

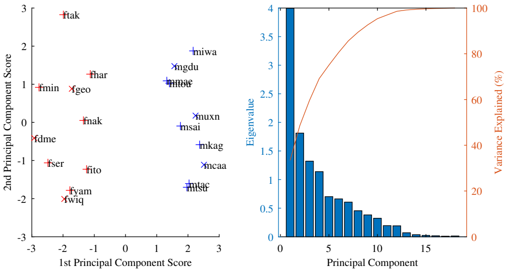

Fig. 7. Embedding vectors estimated by NRM in terms of speaker.

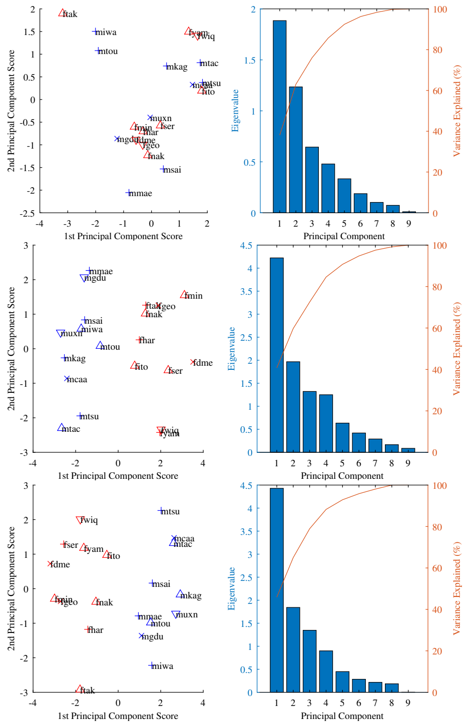

Fig. 8. Embedding vectors estimated by CDM in terms of speaker.

and proposed a methodology to estimate the power exponent of a general spiked model, which can determine whether the estimated eigenvalues are consistent. Experimental results show that the NRM reduced the eigenvalues of PCA but the results of the NRM and PCA were similar. Meanwhile , the CDM yielded more reasonable results but they were dependent on the division of initial sets. The proposed method can estimate the power exponent. In future work, we aim to improve the settings of the CDM for more accurate representations of embedding vectors.

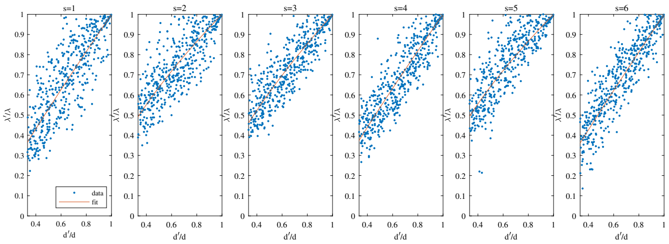

Fig. 9. Relation between the ratio of λ and that of the dimension.

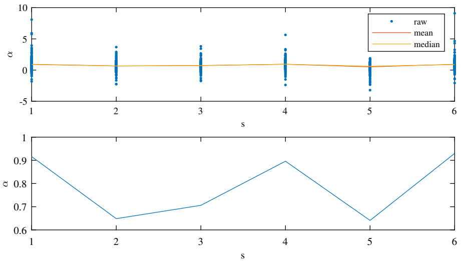

Fig. 10. Estimated α_s.

References

[1] K. H. Hellton and M. Thoresen, “When and why are principal component scores a good tool for visualizing high-dimensional data?” *Scandinavian Journal of Statistics*, vol. 44, no. 3, pp. 581–597, 2017. [Online]. Available: http://www.jstor.org/stable/26428583

[2] K. Yata and M. Aoshima, “PCA consistency for the power spiked model in high-dimensional settings,” *Journal of Multivariate Analysis*, vol. 122, pp. 334–354, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X13001644

[3] I. M. Johnstone, “On the distribution of the largest eigenvalue in principal components analysis,” *The Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001. [Online]. Available: https://doi.org/10.1214/aos/1009210544

[4] J. Baik and J. W. Silverstein, “Eigenvalues of large sample covariance matrices of spiked population models,” *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X0500134X

[5] S. Jung and J. S. Marron, “PCA consistency in high dimension, low sample size context,” *The Annals of Statistics*, vol. 37, no. 6B, pp. 4104–4130, 2009. [Online]. Available: https://doi.org/10.1214/09-AOS709

[6] K. Yata and M. Aoshima, “High-dimensional inference on covariance structures via the extended cross-data-matrix methodology,” *Journal of Multivariate Analysis*, vol. 151, pp. 151–166, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X16300550

[7] ——, “Geometric consistency of principal component scores for high-dimensional mixture models and its application,” *Scandinavian Journal of Statistics*, vol. 47, no. 3, pp. 899–921, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12432

[8] D. Shen, H. Shen, H. Zhu, and J. Marron, “The statistics and mathematics of high dimension low sample size asymptotics,” *Statistica Sinica*, vol. 26, pp. 1747–1770, 10 2016.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions,” in *Procceings of ICASSP*, 2018, pp. 4779–4783.

[10] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *Procceings of ICASSP*, 2019, pp. 3617–3621.

[11] K. Kurihara, N. Seiyama, and T. Kumano, “Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS,” *IEICE Transactions on Information and Systems*, vol. E104-D, no. 2, pp. 302–311, 2 2021.