Gaussian determinantal processes: A new model for directionality in data

Subhroshekhar Ghosh[a,1] and Philippe Rigollet[b]

[a]Department of Mathematics, National University of Singapore, Singapore 119076, Singapore; and [b]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Steven N. Evans, University of California, Berkeley, CA, and approved April 24, 2020 (received for review October 2, 2019)

Determinantal point processes (DPPs) have recently become popular tools for modeling the phenomenon of negative dependence, or repulsion, in data. However, our understanding of an analogue of a classical parametric statistical theory is rather limited for this class of models. In this work, we investigate a parametric family of Gaussian DPPs with a clearly interpretable effect of parametric modulation on the observed points. We show that parameter modulation impacts the observed points by introducing directionality in their repulsion structure, and the principal directions correspond to the directions of maximal (i.e., the most long-ranged) dependency. This model readily yields a viable alternative to principal component analysis (PCA) as a dimension reduction tool that favors directions along which the data are most spread out. This methodological contribution is complemented by a statistical analysis of a spiked model similar to that employed for covariance matrices as a framework to study PCA. These theoretical investigations unveil intriguing questions for further examination in random matrix theory, stochastic geometry, and related topics.

determinantal point processes | spiked model | dimension reduction

The easy access to vast computational, data collection, and storage resources has fueled a paradigm shift in our approach to data-driven decision making, requiring us to rethink many modeling assumptions that are at the core of standard statistical techniques. In this paper, we investigate a certain class of determinantal point processes (DPPs) as an effective paradigm in understanding and analyzing directional effects in large datasets.

Determinantal processes have emerged as an important class of models in recent years in probability, statistical physics, and applied mathematics. Originally introduced to model Fermionic particle systems in physics, DPPs are powerful tools to capture repulsive interactions between the components of a system. A DPP is a random set of points characterized by the fact that its k-point correlation functions $\rho_k(x_1, \ldots, x_k)$ (roughly, the probability densities of having random points at $x_1, \ldots, x_k$) are given by determinants $\det[(K(x_i, x_j))_{1\le i,j\le k}]$ for some kernel K. A more detailed account of these processes is given in *Gaussian Determinantal Processes*.

In recent years, DPPs have been examined as an effective tool to model certain phenomena like data diversity in statistics (1) and machine learning (2–6). DPPs are also closely associated with random matrix theory (RMT) models and Coulomb systems, which turn out to be effective models in many practical scenarios that exhibit repulsive behavior between agents. Real-world applications of RMT and Coulomb models are as varied as the study of transportation systems (in particular, the famous bus system of Cuernavaca), parking of cars, study of pedestrians, perched birds, geography, genetics, interacting particle systems, numerical computations, nuclear spectra, and the theory of dynamical games—for detailed discussions we refer the reader to refs. 7–21 and the references therein. However, the application of sophisticated models as DPPs to address many standard statistical questions, such as dimension reduction, remains a challenge and very much a work in progress. For instance, the assumption of independent observations has been essential in the development and analysis of many crucial statistical tools based on the concentration of measure phenomenon (22), as well as the establishment of fundamental limitations of statistical methods using the minimax paradigm (23). Our understanding of such techniques and their applications for models like DPPs are much more limited in comparison.

In this paper, we propose a family of determinantal probability measures, called Gaussian determinantal processes, on $\mathbb{R}^d$ that is parameterized by a so-called scattering matrix of size $d \times d$. In particular, this positive-definite matrix captures the directionality of the DPP in the following sense: Its eigenvectors associated to its largest eigenvalues span low-dimensional spaces along which the DPP exhibits the most repulsion. We propose a consistent estimator of the scattering matrix and demonstrate that these eigenspaces produce low-dimensional representation of the original data that is, in some sense, better than principal component analysis (PCA) on two benchmark datasets: Fisher's Iris data and the Wisconsin breast cancer dataset. In this context, we employ the DPP model merely as an ansatz to serve as a basis for the development of a method. This is similar in spirit to the celebrated Wigner surmise (17, 24, 25), wherein the model of Gaussian random matrices was introduced to approximate a specific aspect of heavy nuclei (namely, the statistics of their spectral gaps), rather than purporting to model the actual nuclear Hamiltonians themselves. Another analogy would be classical PCA itself, wherein the fundamental concepts are motivated and inspired by normal random variables, but are effectively employed in settings far beyond its Gaussian origins.

Significance

The increasingly complex nature of data has led statisticians to rethinking even the most basic of modeling assumptions. In this context, a determinantal point process (DPP) modeling paradigm promotes diversity in the sample at hand. In this work, we introduce a simple and flexible Gaussian DPP model to capture directionality in the data. Using the Gaussian DPP as an ansatz, we obtain an approach for dimensionality reduction that produces a better and more readable representation of the original data than standard principal component analysis (PCA). These findings are supported by a finite sample analysis of the performance of our estimator, in particular in a spiked model similar to the one employed to analyze PCA.

Author contributions: S.G. and P.R. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

[1]To whom correspondence may be addressed. Email: subhrowork@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1917151117/-/DCSupplemental.

First published June 1, 2020.

APPLIED MATHEMATICS

Downloaded at Palestinian Territory, occupied on November 28, 2021

www.manaraa.com

These findings are supported by a statistical analysis of Gaussian determinantal processes in a spiked model that is similar to one employed to study PCA and that has been the theater for important developments in random matrix theory such as the celebrated Baik–Ben Arous–Peche phase transition (26). In particular, despite the lack of independence in the data generated by a DPP, we show that, akin to the spiked covariance model for traditional PCA, the spike in this DPP model may be detected at a signal strength proportional to $1/\sqrt{n}$—up to logarithmic factors—where n is the number of observations in the DPP model.

Gaussian Determinantal Processes

In this work, we are concerned with DPPs on $\mathbb{R}^d$. Recall that, when it exists, the family of k-point correlation functions $\rho_k(x_1, \dots, x_k)$, $k \geq 1$, characterizes the distribution of point process X as follows: For any integer $k \geq 1$ and any set of k disjoint Borel subsets $A_1, \dots, A_k$ of $\mathbb{R}^d$, let $X(A_i)$ denote the number of points of X that are included in A_i. Then

$$\mathbb{E}[X(A_1) \cdots X(A_k)] = \int_{A_1 \times \cdots \times A_k} \rho_k(x_1, \dots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k.$$

A point process X on $\mathbb{R}^d$ is called a DPP if there exists a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ such that X admits k-point correlations ρ_k given by

$$\rho_k(x_1, \dots, x_k) = \det \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_k) \\ \vdots & \ddots & \vdots \\ K(x_k, x_1) & \cdots & K(x_k, x_k) \end{bmatrix}.$$

For more on the theory of point processes in general, and DPPs in particular, we direct the interested reader to refs. 27–29.

We are now in a position to define Gaussian determinantal processes. To that end, let Φ denote the density of a multivariate Gaussian distribution with mean 0 and covariance matrix Σ:

$$\Phi(u) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} u^\top \Sigma^{-1} u\right), \quad u \in \mathbb{R}^d.$$

We say that the point process X over $\mathbb{R}^d$ is a Gaussian determinantal process (GDP) if it is a DPP with kernel $K(x, y) = \Phi(x - y)$ for some positive definite matrix Σ, called the scattering matrix of X.

Note that the density, i.e., the average number of points per unit volume, is given by $K(x, x) = \Phi(0)$. Throughout this paper, we assume the normalization

$$K(x, x) = \Phi(0) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma}} = 1, \qquad [1]$$

to focus on the spatial dependence structure of the points—as opposed to the spatial scale that is captured by the density. This corresponds to a simple rescaling of the coordinate axes by the same constant factor and, importantly, does not affect the eigenvectors of Σ, which are used for dimension reduction. Note that other normalizations may be considered by simply multiplying Φ with an appropriate constant. This was done, for example, in ref. 1 who introduced GDPs in the isotropic case where $\Sigma = I_d$ and therefore did not focus on directionality as is the case here, but rather on density as a variable parameter. Using Fourier analytic techniques to understand the spectrum of K (viewed as an integral operator) together with the celebrated Macchi–Soshnikov theorem (ref. 27, theorem 4.5.5), it can be shown GDPs exist for any positive-definite scattering matrix (*SI Appendix, Theorem 1*). By construction, GDPs are stationary point processes; that is, their distribution is invariant under translations.

The goal of the Gaussian DPP is to capture the spatial dependency structure of data via the scattering matrix Σ. To see how this manifests itself, observe first that when $\Sigma = \sigma^2 I_d$ for some $\sigma > 0$, we have $K(Ux, Uy) = K(x, y)$ for every orthogonal matrix U, and hence the distribution of this DPP is also invariant under rotations. To understand the effect of Σ on directionality at a qualitative level, let us consider the simplest measure of dependence of the data: the two-point correlation function ρ_2. As we already noted, this is roughly the probability density of having a point at x and a point at y. It is customary in statistical physics to pass from ρ_2 to the truncated pair correlation $\bar{\rho}_2$, given by $\bar{\rho}_2(x, y) = \rho_2(x, y) - \rho_1(x)\rho_1(y)$. This amounts to subtracting off the independent part of the pair correlation, which is roughly the probability of having points at x and y if there was no spatial dependence, and this contribution is simply $\rho_1(x)\rho_1(y)$. Thus $\bar{\rho}_2$ captures the pure contribution of the dependence structure between the points. For a determinantal process with kernel K, it is an immediate observation that $\bar{\rho}_2 = -|K(x, y)|^2$. This in particular means that for a DPP, it holds $\bar{\rho}_2 \leq 0$, which captures the spatial repulsion that is characteristic of such processes. The smaller $\bar{\rho}_2(x, y)$ is in magnitude, the more decorrelated (and hence Poissonian) the point field at x and y is. Conversely, the bigger $\bar{\rho}_2(x, y)$ is in magnitude, the more strongly correlated is the process.

For GDPs (with the normalization Eq. **1**), the truncated pair correlation is given by

$$\bar{\rho}_2(x, y) = -\exp\left(-(x - y)^T \Sigma^{-1} (x - y)\right).$$

Therefore, the magnitude of $\bar{\rho}_2(x, y)$ is proportional to $\phi_{\frac{1}{2}\Sigma}(x - y)$, where $\phi_{\frac{1}{2}\Sigma}$ is the density of a centered multivariate Gaussian distribution with covariance matrix $\frac{1}{2} \cdot \Sigma$. In particular, if $x - y$ is well aligned with an eigenvector of Σ associated to a relatively large eigenvalue, then the magnitude of $\bar{\rho}_2(x, y)$ is large even for large values of $\|x - y\|$. In other words, for such directions, the spatial correlation has a longer range.

This motivates us to formulate a spiked model to study the estimation of the spectrum of the scattering matrix Σ. It follows from the above discussion that the eigenvectors of the scattering matrix Σ that are associated to large eigenvalues capture long-range spatial correlations—more specifically repulsion—between the points generated by the GDP. This phenomenon is the basis of a spiked scattering model by analogy to the popular spiked covariance model (30–32) where the covariance matrix S is assumed to be a rank-one perturbation of the identity: $S = I_d + \lambda u u^\top$, $\|u\| = 1, \lambda \geq 0$. In this case, all of the directionality of the model is carried by the rank-one perturbation $uu^\top$ called the spike and its strength is carried by the parameter $\lambda > 0$.

In the context of GDPs, it is natural to work in the situation where the presence of the spike leaves the mean density of points unchanged compared to the isotropic case, where the scattering matrix is the identity. If that is not the case, then the presence of the spike can be detected simply by estimating the density. The mean density remaining unchanged amounts to $\det(\Sigma)$ being equal for the null and alternatives. This leads to the model

$$(2\pi)\Sigma = (1 + \lambda)^{-\frac{1}{d-1}} (I_d - uu^\top) + (1 + \lambda) uu^\top, \quad \|u\| = 1, \qquad [2]$$

where $\lambda \geq 0$ is the strength parameter and $uu^\top$ is the spike. Note that when $\lambda = 0$, we recover the isotropic case $\Sigma = I_d$. The constant 2π in front of Σ is inconsequential and simply ensures the normalization adopted Eq. **1**.

Downloaded at Palestinian Territory, occupied on November 28, 2021

www.manaraa.com

Under the spiked model Eq. **2**, we ask two natural statistical questions:

1) The detection question consists in studying the signal strength $\lambda > 0$ sufficient to detect the presence of the spike from data (the precise data scheme is described in detail in the next section); this is a simple vs. composite hypothesis-testing problem of the form

$$H_0 : (2\pi)\Sigma = I_d \quad \text{vs.}$$
$$H_1 : (2\pi)\Sigma = (1+\lambda)^{-\frac{1}{d-1}}(I_d - uu^\top) + (1+\lambda)uu^\top, \quad \|u\| = 1.$$

To define this problem more precisely, let $\mathbb{P}_0$ (resp. $\mathbb{P}_u$) denote the probability distribution associated to the GDP with scattering matrix I_d (resp. $\lambda^{-\frac{1}{d-1}}(I_d - uu^\top) + \lambda uu^\top$). We say that we can detect the spike at strength $\lambda > 0$ if, for any $\delta \in (0,1)$, there exists a Borel function of the data $\psi \in \{0,1\}$, called a test, such that

$$\mathbb{P}_0(\psi = 1) \vee \sup_{u:\|u\|=1} \mathbb{P}_u(\psi = 0) \le \delta. \qquad [3]$$

2) The estimation question consists in estimating the direction u at a constant signal strength λ. Such results follow from a good estimator of Σ together with matrix perturbation arguments.

Both questions tie in naturally with the spiked covariance matrix models that are nowadays ubiquitous in high-dimensional statistics, particularly in the context of PCA (33–36).

Statistical Estimation

Despite the clarity of the underlying principle behind DPPs, our understanding of fundamental statistical procedures such as the method of moments and maximum-likelihood estimation in this context is largely limited to the discrete settings (37, 38) when multiple independent realizations of the process are available to the statistician. While this setup is particularly relevant to machine learning, it leaves behind the fundamental framework of continuous DPPs where a single realization of the process is observed (1). This framework poses a significant challenge due to the complete lack of independence between the sample points.

Note that by stationarity, a DPP realization has almost surely an infinite number of points over $\mathbb{R}^d$. In this work, we assume that we observe the realization of this DPP in a Euclidean ball $B(R) \subset \mathbb{R}^d$ centered at the origin and with radius $R > 0$.

Our estimation strategy is driven by the fact that we can examine many distinct local neighborhoods of our observation window. By translation invariance, we may average various statistics over such local neighborhoods to recover global properties of the DPP.

Let X be a GDP with scattering matrix Σ and let $X_1, \dots, X_N \in B(R)$ denote the set of points of X that are included in $B(R)$. Note that N is a random number but it concentrates sharply (*SI Appendix*, *Theorem 2*) around its expectation

$$n := \mathbb{E}[N] = |B(R)| = |B(1)|R^d,$$

where $|B(t)|$ denotes the volume of a Euclidean ball with radius $t \ge 0$.

We are now in a position to define an estimator $\hat{\Sigma}$ for the scattering matrix Σ. It is parameterized by a positive cutoff threshold $r < R$ that is chosen theoretically by realizing a bias–variance tradeoff. However, empirical investigations indicate that our estimator shows little sensitivity to the choice of r.

For each observation X_i, $i = 1, \dots, N$, let $\mathcal{N}_i$ denote the set of other observations that are at distance at most r from it:

$$\mathcal{N}_i = \{j : j \ne i,\ \|X_i - X_j\| < r\}.$$

To avoid boundary effects, we also define $\mathcal{N}_0 = \{j : \|X_j\| < R - r\}$. Define the estimator $\hat{\Sigma}$ of Σ to be the $d \times d$ matrix given by

$$\hat{\Sigma} = |B(1)|\frac{r^{d+2}}{d+2} I_d - \frac{1}{|B(R-r)|}\sum_{i\in\mathcal{N}_0}\sum_{j\in\mathcal{N}_i}(X_i - X_j)(X_i - X_j)^\top.$$

Using the Fourier analytic properties of the Gaussian kernel, we can control the accuracy of the estimator $\hat{\Sigma}$ in a Frobenius norm. Throughout our considerations, we assume that $\|\Sigma\|_{\mathrm{op}}$ is bounded above by a universal constant. This amounts to assuming that all of the eigenvalues of Σ are of order one or, in other words, that the scattering matrix is not degenerate.

This estimator is negatively biased in the sense that for any unit vector $v \in \mathbb{R}^d$ we have $\mathbb{E}[v^\top \hat{\Sigma} v] \le v^\top \Sigma v^\top$. Moreover, it can be shown that, for r large enough, we have

$$\|\mathbb{E}\hat{\Sigma} - \Sigma\|_{\mathrm{F}}^2 \lesssim d\exp\left(C(2d - r^2)\right),$$

for some positive constant C. In particular, the bias vanishes exponentially fast as $r \to \infty$. For details on these results, we refer the reader *SI Appendix*, *Lemma 4*.

Moreover, the variance of $\hat{\Sigma}$ can be controlled as follows (*SI Appendix*, *Lemma 5*): For r large enough,

$$\mathbb{E}\|\hat{\Sigma} - \mathbb{E}\hat{\Sigma}\|_{\mathrm{F}}^2 \le d^2\left(\frac{C}{d}\right)^d \frac{r^{2d+4}}{n}.$$

In particular, the bound on the variance is inversely proportional to the sample size as is the case in the independent setting.

If we choose the $r = C\sqrt{d\log n}$ for a universal constant $C > 0$ so as to realize the bias–variance tradeoff (*SI Appendix*, *Theorem 3*), we get

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}} \le \mathfrak{r}_{n,d} := \frac{d^2(c\sqrt{\log n})^{d+1}}{\sqrt{n}}, \quad c > 0, \qquad [4]$$

for n large enough. As a result, we actually get a nonasymptotic bound $\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$ that is of order $1/\sqrt{n}$ up to logarithmic terms as in the traditional independent and identically distributed setting, despite the dependence between points. However, the (exponential) dependence on the dimension of the bound in this result leaves scope for improvement in further investigations.

The estimator $\hat{\Sigma}$ might fail to be positive definite, albeit with a small probability that vanishes as the data size tends to infinity, because of the concentration of $\hat{\Sigma}$ around Σ. If, for some application, it is necessary to have a positive-definite estimator, one can naturally consider the projection of $\hat{\Sigma}$ onto the positive-definite cone.

The Spiked Model

We now turn to the spiked model Eq. **2**.

Detection. It is useful to note that in the hypothesis testing associated to detection, it holds that $\|\Sigma\|_{\mathrm{op}} = 1/2\pi$ under H_0 and $\|\Sigma\|_{\mathrm{op}} = (1+\lambda)/2\pi$ under H_1. Of course, the above result readily yields that $\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \le C(d)/\sqrt{n}$. This leads us to consider the test statistic $\|\hat{\Sigma}\|_{\mathrm{op}}$ and the test

$$\psi_t = \mathbf{1}\left\{(2\pi)\|\hat{\Sigma}\|_{\mathrm{op}} > 1 + t\mathfrak{r}_{n,d}\right\}, \quad t > 0,$$

Downloaded at Palestinian Territory, occupied on November 28, 2021
www.manaraa.com

where $\mathbf{1}\{\cdot\}$ is the indicator function. Using the Markov and triangle inequalities, we get that if $t = 1/\delta$ and

$$\lambda > \frac{\mathfrak{r}_{n,d}}{\pi},$$

then the text $\psi_{1/\delta}$ satisfies Eq. **3**. In particular, this implies that we can detect the presence of the spike at strength $\bar{\lambda} = 2\mathfrak{r}_{n,d}$.

Our technique, based on bounding the operator norm by the Frobenius norm, is too blunt to characterize the sharp dependence in the dimension d of the optimal strength at which the spike may be detected. An answer to this question would involve a sophisticated understanding of the large deviations of $\hat{\Sigma}$, requiring detailed investigations in its own right. We believe that this would lead to challenging problems in random matrix theory, involving the understanding of the strong spatial dependence and the stochastic geometry of GDPs.

Estimation. When λ is large enough, the estimation of u in the model Eq. **2** follows from matrix perturbation analysis. Recall that u is the eigenvector associated to the largest eigenvalue $1+\lambda$ of Σ and that Σ exhibits a spectral gap of $1+\lambda-(1+\lambda)^{-\frac{1}{d+1}} \geq \lambda$. Moreover, let $\hat{u}$ denote any eigenvector associated to the largest eigenvalue of $\hat{\Sigma}$. It follows from the Davis–Kahan theorem that

$$\mathbb{E}|\sin(\measuredangle(\hat{u}, u))| \leq \frac{\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\lambda} \leq \frac{\mathfrak{r}_{n,d}}{\lambda},$$

where $\measuredangle(\hat{u}, u)$ denotes the angle between $\hat{u}$ and u. In light of Eq. **4**, the above result indicates that we can estimate the spike at the parametric rate $1/\sqrt{n}$. As before, the optimal dependence in the dimension d is left for future research. Nevertheless, the above result gives us a theoretical basis for the dimension reduction technique presented in the next section.

Dimension Reduction

To illustrate the ability of the GDP model to find directionality in data, we apply our techniques to two benchmark datasets: Fisher's Iris data and the Wisconsin breast cancer dataset. We compare our results with the representations obtained by standard PCA.

More precisely, given the data, we estimate the matrix Σ using the estimator $\hat{\Sigma}$ defined above. Note that for the purpose of computing eigenvectors of $\hat{\Sigma}$, the choice of R does not matter. Moreover, to avoid any parameter tuning, we take r large enough so that all of the points are in the neighborhood $\mathcal{N}_i$ of all other points i. While this choice minimizes the bias term, it potentially allows for a large variance. However, this choice gave the best result in both numerical examples below. It appears from the singular-value decomposition of our estimator that the nondegeneracy assumption that we make in our analysis is violated, leading to a large operator norm for Σ (Fig. 1). In particular, this contributes to the upper bound on the bias and can be compensated by a large value of r, thus explaining the good empirical behavior of a large r.

Then we compute the singular-value decomposition of $\hat{\Sigma}$ and extract the eigenvectors associated to the largest eigenvalues. In both case, it is worth noting that the leading eigenvalue far exceeds the subsequent ones, which indicates that for these two datasets, most of the repulsion is captured by one dimension (Fig. 1).

For both datasets, when displaying a two-dimensional (2D) scatterplot, we simply project the original points on the space spanned by the first two eigenvectors of $\hat{\Sigma}$ and compare with the representation given by PCA (with centering and scaling options activated).

Fisher's Iris. While this dataset contains little mystery, it is perhaps the most standard dataset to apply PCA to (39, 40). It contains $N = 150$ observations in dimension $d = 3$ split into three clusters, each corresponding to a different type of iris. In Fig. 2 the points are colored according the cluster they belong to. While the main structure between the clusters is preserved between the two representations, two phenomena emerge. On the one hand, the DPP approach leads to a point cloud that displays better spatial organization, especially locally, due to the repulsion between the points in these directions. On the other hand, the DPP approach appears to provide a better clustering between the red circles (Versicolor) and the green triangles (Virginica).

Wisconsin Breast Cancer. The Wisconsin breast cancer dataset (41, 42) comprises 569 observations in dimension $d = 30$. Each observation corresponds to either a benign or a malignant tumor. This dataset was originally obtained from the University of Wisconsin Hospitals, Madison, WI, from Dr. William H. Wolberg.

In this dataset, the difference between the two methods is stark as illustrated in Fig. 3. While it appears that the DPP-based approach presents less repulsion between the points, especially within the cluster of black squares, this should be mitigated with

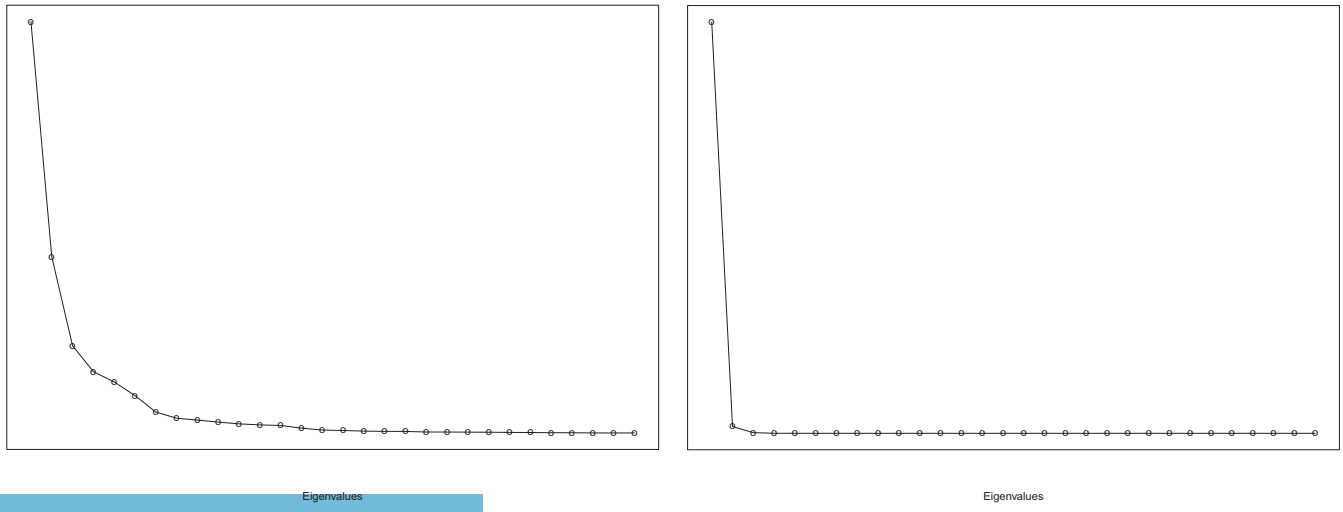

Fig. 1. Ordered eigenvalues of the sample correlation matrix (*Left*) and of $\hat{\Sigma}$ (*Right*).

Downloaded at Palestinian Territory, occupied on November 28, 2021

www.manaraa.com

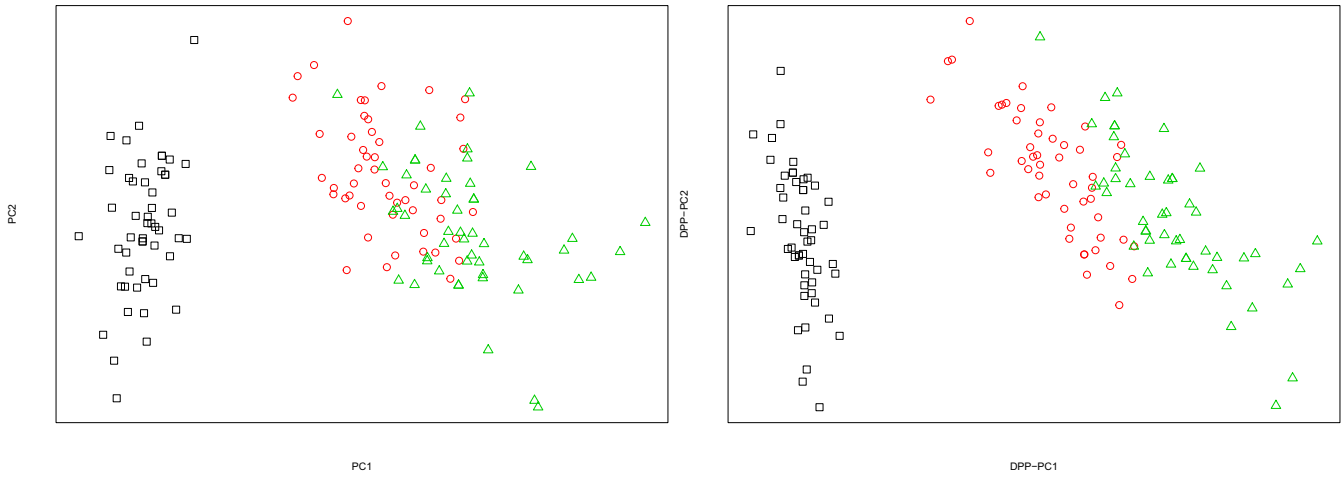

Fig. 2. Projection of the first two principal components (*Left*) and the first two eigenvectors of $\hat{\Sigma}$ (*Right*). Different symbols correspond to different clusters.

the fact that this cluster corresponds, in fact, to benign tumors. As a result, the DPP approach tends to cluster the benign tumors into one tight cluster whereas it maintains significant diversity within the cluster of malignant tumors. This phenomenon is not without resemblance to the opening line of Tolstoy's *Anna Karenina* (ref. 43, p. 1): "Happy families are all alike; every unhappy family is unhappy in its own way."

It turns out that the DPP approach also does a better job at separating benign from malignant tumors in the following sense. Consider the risk score associated to an observation as the negative value of its coordinate when projected onto either the first principal component or the leading eigenvector of $\hat{\Sigma}$. The larger the (negative) value for this coordinate, the higher the risk of the tumor being malignant. Receiver operating characteristic (ROC) curves are a well-established tool to compare risk scores. In Fig. 4, we compare the ROC curves obtained using the coordinates from the first principal component (PCA approach) and the coordinates from the leading eigenvector of $\hat{\Sigma}$ (DPP approach). These are comparable and both quite good: The PCA approach gives an area under the ROC curve of 0.970 whereas the DPP approach gives one of 0.963. However, when zooming toward high scores (that is, greater risk of being malignant), we see in Fig. 4 (*Right*) that the DPP approach dominates the PCA, even displaying more than 50% of true positives while having no false positives. This performance is all the more impressive in that it is a fully unsupervised method.

For this dataset, it is also interesting to look at the associated scree plots which indicate the eigenvalues from largest to smallest. While in the PCA case, we observe a graceful decay, the DPP approach outputs a leading eigenvalue that largely dominates all others. This indicates that for this dataset, the spiked model 2 appears to hold approximately.

Conclusion

In this work, we investigated the Gaussian DPP model to capture directionality in the dependence structure of data. This parametric model is driven by a kernel that is parameterized by a $d \times d$ positive-definite scattering matrix, whose spectral properties govern the directionality inherent in the point set. We propose a consistent estimator of the scattering matrix and employ the leading eigenvectors of this estimating matrix to obtain a low-dimensional representation of the data. Applying this methodology to two benchmark datasets, namely the Fisher's Iris data and the Wisconsin breast cancer data, we exhibit that the subspaces obtained from GDP show arguably better performance than the PCA as a dimension reduction technique. Our empirical investigations are complemented by a theoretical study of an accompanying spiked model for the scattering

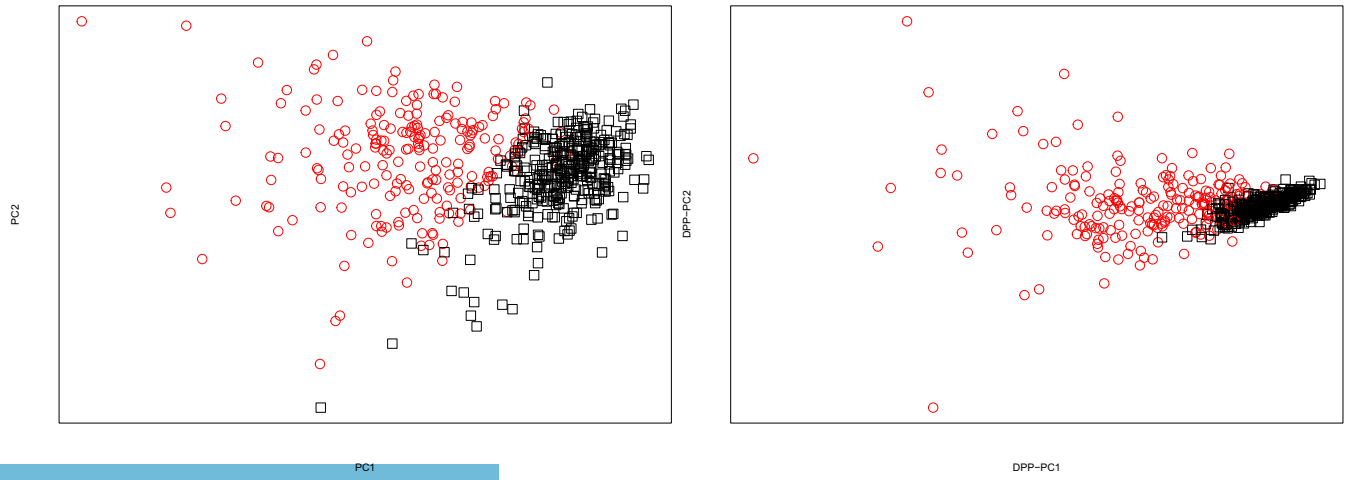

Fig. 3. Projection of the first two principal components (*Left*) and the first two eigenvectors of $\hat{\Sigma}$ (*Right*). Different symbols correspond to different clusters.

Downloaded at Palestinian Territory, occupied on November 28, 2021

www.manaraa.com

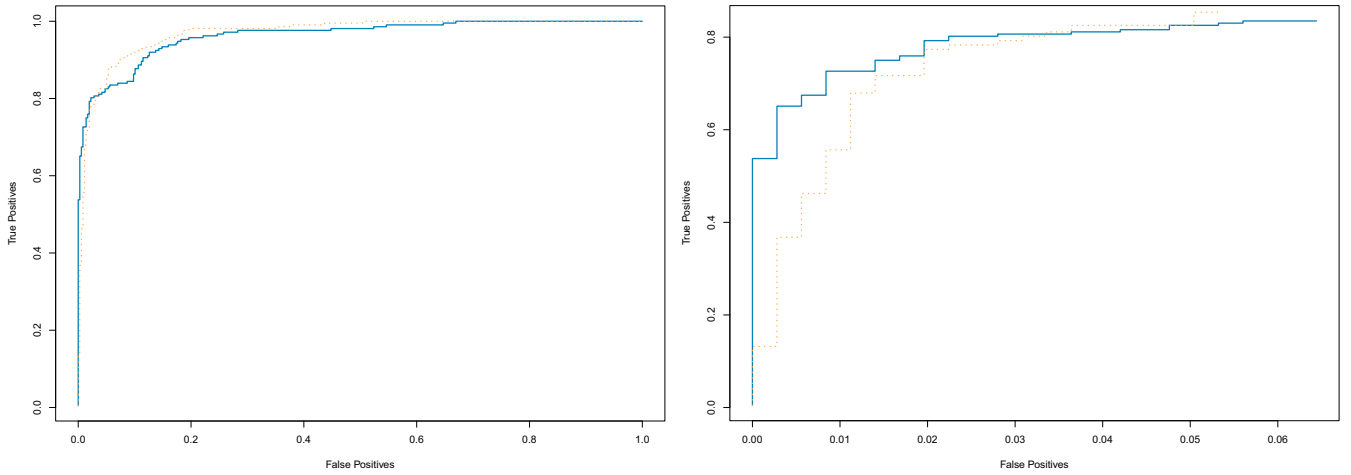

Fig. 4. (*Left*) ROC curves given by the first principal component (dashed curve) and the first eigenvector of $\hat{\Sigma}$ from the DPP approach (solid curve). (*Right*) Only the 200 observations with the highest risk score are represented.

matrix, wherein we provide detection guarantees that hold with high probability (with increasing sample size n) at a detection threshold of $1/\sqrt{n}$ (up to logarithmic factors). Despite the significant challenges posed by the complete absence of any kind of independence in the DPP model, this is comparable to the analogous threshold for spiked covariance matrices.

The GDP model and its application as a tool for dimension reduction throw up natural questions and exciting challenges for further investigation. A natural problem is to extend the variance estimates for our estimator to central limit theorems and tight nonasymptotic bounds, so that a finer theoretical analysis of the estimation and testing problems can be carried out. From the modeling perspective, it would be natural to examine more complex alternatives for the spiked model compared to the fundamental one-parameter case studied in this paper, including a richer interpolating family between the identity and the spike and the possibility of having multiple spike directions. Another direction to study would be the extension of the ideas contained in the present paper to more general kernel classes beyond the Gaussian. The investigation of the estimator $\hat{\Sigma}$ as a random matrix, in particular its spectral properties and their large deviations, would be of great theoretical interest in its own right, as well as have important implications for the statistical problems discussed in this work. These can be related to the general objective of improving the dimensional dependence in the rates obtained in the present paper. Finally, in the context of the spiked model, establishing a sharp threshold for detection poses a natural question to investigate. We envisage that these investigations will lead to intriguing challenges in random matrix theory, stochastic geometry, and the analysis of directionality in data, ultimately involving the tension between the lack of independence on one hand and the concrete determinantal structure of the GDP on the other.

Materials and Methods

The data analyses discussed in this paper have been implemented in R. The relevant codes are publicly available in ref. 44. The two benchmark datasets employed in this paper are available in refs. 39 and 41 for Fisher's Iris data and the Wisconsin breast cancer data, respectively.

ACKNOWLEDGMENTS. S.G. is supported in part by the Singapore Ministry of Education Grant R-146-000-250-133. P.R. is supported by NSF Awards IIS-BIGDATA-1838071, DMS-1712596, and CCF-TRIPODS-1740751; and Oce of Naval Research Grant N00014-17-1-2147. We thank Victor-Emmanuel Brunel for illuminating discussions. We thank the referees for their careful reading of this paper and for their insightful comments and suggestions.

1. F. Lavancier, J. Møller, E. Rubak, Determinantal point process models and statistical inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **77**, 853–877 (2015).
2. A. Kulesza, B. Taskar, *Determinantal Point Processes for Machine Learning* (Foundations and Trends in Machine Learning, Now Publishers Inc., 2012), vol. 5.
3. R. H. Affandi, E. Fox, R. Adams, B. Taskar, "Learning the parameters of determinantal point process kernels" in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds. (Proceedings of Machine Learning Research, 2014), vol. 32, pp. 1224–1232.
4. J. A. Gillenwater, "Approximate inference for determinantal point processes," PhD thesis, University of Pennsylvania, Philadelphia, PA (2014).
5. Z. Mariet, S. Sra, "Fixed-point algorithms for learning determinantal point processes" in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach, D. Blei, Eds. (Proceedings of Machine Learning Research, 2015), vol. 37, pp. 2389–2397.
6. R. H. Affandi, E. Fox, B. Taskar, "Approximate inference in continuous determinantal processes" in *Advances in Neural Information Processing Systems* C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2013), pp. 1430–1438.
7. M. Krbálek, P. Seba, The statistical properties of the city transport in Cuernavaca (Mexico) and random matrix ensembles. *J. Phys. A Math. Gen.* **33**, L229 (2000).
8. J. Baik, A. Borodin, P. Deift, T. Suidan, A model for the bus system in Cuernavaca (Mexico). *J. Phys. A Math. Gen.* **39**, 8965–8975 (2006).
9. A. Jagannath, T. Trogdon, Random matrices and the New York City subway system. *Phys. Rev. E* **96**, 030101 (2017).
10. A. Y. Abul-Magd, Modelling gap-size distribution of parked cars using random-matrix theory. *Phys. Stat. Mech. Appl.* **368**, 536–540 (2006).
11. D. Jezbera, D. Kordek, J. Kříž, P. Šeba, P. Šroll, Walkers on the circle. *J. Stat. Mech. Theor. Exp.* **2010**, L01001 (2010).
12. P. Šeba, Parking and the visual perception of space. *J. Stat. Mech. Theor. Exp.* **2009**, L10002 (2009).
13. G. Le Caër, R. Delannay, The administrative divisions of mainland France as 2d random cellular structures. *J. Phys.* **3**, 1777–1800 (1993).
14. F. Luo, J. Zhong, Y. Yang, J. Zhou, Application of random matrix theory to microarray data for discovering functional gene modules. *Phys. Rev. E* **73**, 031924 (2006).
15. P. Deift, Some open problems in random matrix theory and the theory of integrable systems. *Contemp. Math.* **458**, 419–430 (2008).
16. P. Deift, Some open problems in random matrix theory and the theory of integrable systems. II. arXiv:1703.04931 (15 March 2017).
17. E. P. Wigner, Random matrices in physics. *SIAM Rev.* **9**, 1–23 (1967).
18. A. Edelman, Y. Wang, "Random matrix theory and its innovative applications" in *Advances in Applied Mathematics, Modeling, and Computational Science*, R. Melnikllias, I. S. Kotsireas, Eds. (Springer, 2013), pp. 91–116.
19. J. Von Neumann, H. H. Goldstine, Numerical inverting of matrices of high order. *Bull. Am. Math. Soc.* **53**, 1021–1099 (1947).
20. P. A. Deift, G. Menon, S. Olver, T. Trogdon, Universality in numerical computations with random data. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14973–14978 (2014).
21. R. Carmona, M. Cerenzia, A. Z. Palmer, The Dyson game. arXiv:1808.02464 (1 October 2018).
22. M. Ledoux, *The Concentration of Measure Phenomenon* (Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 2001).

Downloaded at Palestinian Territory, occupied on November 28, 2021
www.manaraa.com

23. A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics, Springer, New York, NY, 2009).
24. E. P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.* **62**, 548–564 (1955).
25. M. L. Mehta, *Random Matrices* (Elsevier, 2004), vol. 142.
26. J. Baik, G. Ben Arous, S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697 (2005).
27. J. Ben Hough, M. Krishnapur, Y. Peres, B. Virág, *Zeros of Gaussian Analytic Functions and Determinantal Point Processes* (University Lecture Series, American Mathematical Society, Providence, RI, 2009), vol. 51.
28. O. Kallenberg, *Foundations of Modern Probability* (Springer Science & Business Media, 2006).
29. A. Borodin, "Determinantal point processes" in *The Oxford Handbook of Random Matrix Theory*, G. Akemann, J. Baik, P. Di Francesco, Eds. (Oxford University Press, Oxford, UK, 2011), pp. 231–249.
30. I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327 (2001).
31. D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sin.* **17**, 1617–1642 (2007).
32. D. Féral, S. Péché, The largest eigenvalues of sample covariance matrices for a spiked population: Diagonal case. *J. Math. Phys.* **50**, 073302 (2009).
33. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, NY, ed. 2, 1984).
34. Q. Berthet, P. Rigollet, Optimal detection of sparse principal components in high dimension. *Ann. Stat.* **41**, 1780–1815 (2013).
35. Q. Berthet, P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection" in *Proceedings of the 26th Annual Conference on Learning Theory*, S. Shalev-Shwartz, I. Steinwart, Eds. (Proceedings of Machine Learning Research, 2013), vol. 30, pp. 1046–1066.
36. I. T. Jolliffe, J. Cadima, Principal component analysis: A review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
37. V.-E. Brunel, A. Moitra, P. Rigollet, J. Urschel, "Rates of estimation for determinantal point processes" in *Proceedings of the 2017 Conference on Learning Theory* S. Kale, O. Shamir, Eds. (Proceedings of Machine Learning Research, 2017), vol. 65, pp. 343–345.
38. J. Urschel, V.-E. Brunel, A. Moitra, P. Rigollet, "Learning determinantal point processes with moments and cycles" in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (Proceedings of Machine Learning Research, 2017), vol. 7, pp. 3511–3520.
39. UCI Machine Learning Repository, Fisher's iris data set.https://archive.ics.uci.edu/ml/datasets/Iris. Accessed 2 October 2019.
40. R. A. Fisher, The use of multiple measurements in taxonomic problems. *Ann. Hum. Genet.* **7**, 179–188 (1936).
41. UCI Machine Learning Repository, Wisconsin breast cancer data set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original). Accessed 2 October 2019.
42. W. H. Wolberg, O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9193–9196 (1990).
43. L. Tolstoy, *Anna Karenina* (Courier Corporation, 2012).
44. P. Rigollet, R codes for data analyses. http://www-math.mit.edu/∼rigollet/Code/DPP/DPPcode.R. Accessed 2 October 2019.

Downloaded at Palestinian Territory, occupied on November 28, 2021

www.manaraa.com