Maximal Correlation Functions: Hermite, Laguerre, and Jacobi

Anuran Makur
EECS Department
Massachusetts Institute of Technology
Email: a_makur@mit.edu

I. Introduction

Hermite, Laguerre, and Jacobi form an intriguing trinity that is ubiquitous in mathematics. For example, the eigenvalue densities of Wigner, Wishart, and Manova matrix ensembles correspond to manifestations of Hermite, Laguerre, and Jacobi, respectively, in random matrix theory. Likewise, the symmetric eigenvalue decomposition (or spectral decomposition), the singular value decomposition (SVD), and the generalized singular value decomposition are the Hermite, Laguerre, and Jacobi, respectively, in linear algebra. Finally, undirected graphs, bipartite graphs, and regular graphs constitute the Hermite, Laguerre, and Jacobi, respectively, in graph theory. Several other instances of this pervasive pattern are presented in Chapter 5 of [Edelman, 2016] and the references therein.

In this report, we elucidate yet another instance of Hermite, Laguerre, and Jacobi in the context of maximal correlation functions (although our association is not perfect). Maximal correlation was introduced as a measure of statistical dependence between two random variables in [Rényi, 1959]. Since then, it has received considerable attention in the context of statistics [Breiman and Friedman, 1985], [Anantharam et al., 2013], [Calmon et al., 2013]. Computing the maximal correlation between two random variables X and Y involves finding two functions $f(X)$ and $g(Y)$ that are maximally correlated with each other in the Pearson correlation sense. We will illustrate that such maximal correlation functions are precisely the Hermite, Laguerre, or Jacobi polynomials when the joint distribution of X and Y has the form of a natural exponential family with quadratic variance function likelihood (articulated in [Morris, 1982]) along with its conjugate prior. Such joint distributions are considered a very elegant class of distributions for both theoretical analysis purposes and more applied inference scenarios. We introduce and formalize these concepts in Sections II and III, and delineate the Hermite, Laguerre, and Jacobi cases in Sections IV, V, and VI, respectively.

II. Maximal Correlation Functions

We commence by recalling that the classical *Pearson correlation coefficient* between two jointly distributed random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ is defined as:

$$\rho\left(X;Y\right) \triangleq \frac{\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)\left(Y-\mathbb{E}\left[Y\right]\right)\right]}{\sqrt{\mathbb{VAR}(X)\mathbb{VAR}(Y)}} \tag{1}$$

where we assume that X and Y have strictly positive and finite variance. Although the Pearson correlation is analytically simple to evaluate in theory and computationally tractable to implement in practice, it only measures the linear relationship between X and Y rather than capturing true statistical dependence. Indeed, $|\rho\left(X;Y\right)| = 1$ if and only if Y is almost surely a linear function of X, and $\rho\left(X;Y\right) = 0$ does not necessarily imply that X and Y are independent. In [Rényi, 1959], Rényi provided an elegant generalization of Pearson correlation that captures dependence between random variables. For any two jointly distributed random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, he proposed that a "reasonable" measure of statistical dependence, $\Delta(\cdot;\cdot)$, must satisfy the following axioms [Rényi, 1959]:

1) (Non-Degeneracy) $\Delta(X;Y)$ is well-defined as long as X and Y are not constant almost surely.
2) (Symmetry) $\Delta(X;Y) = \Delta(Y;X)$.
3) (Normalization) $0 \leq \Delta(X;Y) \leq 1$.
4) (Vanishing for Independence) $\Delta(X;Y) = 0$ if and only if X and Y are independent.
5) (Maximium for Strict Dependence) $\Delta(X;Y) = 1$ if there exists a Borel measurable function $f : \mathbb{R} \to \mathbb{R}$ (or $g : \mathbb{R} \to \mathbb{R}$) such that $f(X) = Y$ almost surely (or $g(Y) = X$ almost surely).

6) (Bijection Invariance) If $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are bijective Borel measurable functions, then $\Delta(X;Y) = \Delta(f(X);g(Y))$.
7) (Simplification for Gaussian) If X and Y are jointly Gaussian, then $\Delta(X;Y) = |\rho(X;Y)|$. (This is because Gaussian dependence structure is completely characterized by second order statistics.)

Rényi then proved that *maximal correlation*, which is often referred to as the *Hirschfeld-Gebelein-Rényi maximal correlation*, satisfies these axioms in [Rényi, 1959]. We present maximal correlation in the ensuing definition.

Definition 1 (Maximal Correlation)**.** For any two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, the maximal correlation between X and Y is given by:

$$\rho_{\mathsf{max}}(X;Y) \triangleq \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\ g:\mathcal{Y}\to\mathbb{R}\ :\\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0\\ \mathbb{E}\left[f^2(X)\right]=\mathbb{E}\left[g^2(Y)\right]=1}} \mathbb{E}\left[f(X)g(Y)\right]$$

where the supremum is taken over all Borel measurable functions. If X or Y is a constant almost surely, there exist no functions f and g which satisfy the constraints, and we define $\rho_{\mathsf{max}}(X;Y) = 0$.

Letting $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ recovers Rényi's original definition, but statisticians often prefer Definition 1 as it allows for categorical random variables or high dimensional vector random variables. To unravel this definition further, we provide some illustrations of data and the corresponding values of Pearson and maximal correlation in Figure 1. The Julia code to generate these plots is provided in Appendix A. It uses the "sample versions" of Pearson and maximal correlation presented in (1) and Definition 1, respectively (which are "population versions"). We see that in each case in Figure 1, there is an explicit formula that relates X and Y, and so $\rho_{\mathsf{max}}(X;Y) = 1$. On the other hand, $\rho(X;Y) = 0$ when the data is quadratic or circular. Hence, Pearson correlation completely fails to capture dependence in these cases.

We can also understand maximal correlation better from a linear algebraic standpoint. Indeed, Definition 1 appears to have the flavor of a Courant-Fischer-Weyl variational characterization of a singular value. It turns out that maximal correlation is in fact the second largest singular value of a *conditional expectation operator*. We formalize this in the next subsection.

A. Spectral Characterization of Maximal Correlation Functions

We fix a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, and define the random variable $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}$ with probability density P_X with respect to a σ-finite measure λ on the standard Borel measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra on $\mathcal{X}$. We also define the random variable $Y : \Omega \to \mathcal{Y} \subseteq \mathbb{R}$, whose law is determined by the conditional probability densities $\left\{P_{Y|X=x} : x \in \mathcal{X}\right\}$ with respect to a σ-finite measure μ on the standard Borel measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. This specifies a joint probability density $P_{X,Y}$ on the product measure space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}), \lambda \times \mu)$ such that $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We let $\mathbb{P}_X$ (with support $\mathcal{X}$) and $\mathbb{P}_Y$ (with support $\mathcal{Y}$) denote the marginal probability laws of X and Y, respectively. Corresponding to $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$, we define the separable Hilbert space $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ over the field $\mathbb{R}$:

$$\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \triangleq \left\{f : \mathcal{X} \to \mathbb{R} \,|\, \mathbb{E}\left[f^2(X)\right] < +\infty\right\} \tag{2}$$

which is the space of all Borel measurable and $\mathbb{P}_X$-square integrable functions with inner product defined as:

$$\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X),\ \langle f_1, f_2\rangle_{\mathbb{P}_X} \triangleq \mathbb{E}\left[f_1(X)f_2(X)\right]. \tag{3}$$

This inner product is precisely the correlation between $f_1(X)$ and $f_2(X)$. The corresponding induced norm is:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X),\ \|f\|_{\mathbb{P}_X} \triangleq \sqrt{\mathbb{E}\left[f^2(X)\right]}. \tag{4}$$

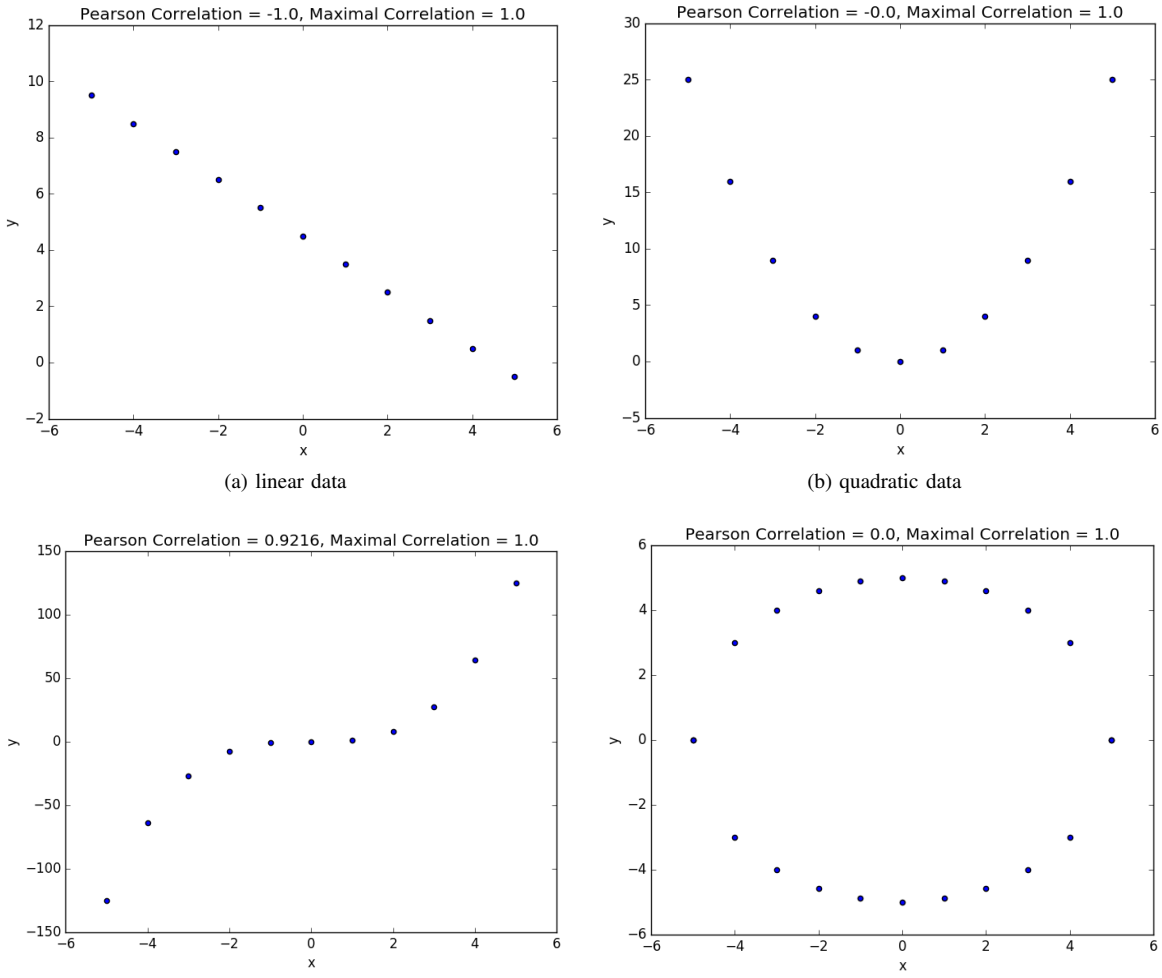
Similarly, we also define the separable Hilbert space $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ corresponding to $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathbb{P}_Y)$.

The *conditional expectation operators* are bounded linear maps that are defined between these Hilbert spaces. The "forward" conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is defined as:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X),\ \ (C(f))(y) \triangleq \mathbb{E}\left[f(X)|Y = y\right]. \tag{5}$$

Observe that for any $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$:

$$\|C(f)\|^2_{\mathbb{P}_Y} = \mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]^2\right] \le \mathbb{E}\left[\mathbb{E}\left[f^2(X)|Y\right]\right] = \|f\|^2_{\mathbb{P}_X}$$



(a) linear data

(b) quadratic data

(c) cubic data

(d) circular data

Fig. 1: Plots of noiseless bivariate data with corresponding values of Pearson correlation coefficient and maximal correlation.

using conditional Jensen's inequality and the tower property, and equality can be achieved if f is the everywhere unity function. Hence, C has operator norm:

$$\|C\|_{\mathsf{op}} \triangleq \sup_{f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)} \frac{\|C(f)\|_{\mathbb{P}_Y}}{\|f\|_{\mathbb{P}_X}} = 1. \tag{6}$$

We may define the "reverse" conditional expectation operator $C^* : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ as:

$$\forall g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y), \quad (C^*(g))(x) \triangleq \mathbb{E}[g(Y)|X = x] \tag{7}$$

which is the unique adjoint operator of C with operator norm $\|C^*\|_{\mathsf{op}} = 1$.[1] Indeed, for every $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and every $g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, we have:

$$\langle C(f), g \rangle_{\mathbb{P}_Y} = \mathbb{E}[\mathbb{E}[f(X)|Y]\, g(Y)] = \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)\mathbb{E}[g(Y)|X]] = \langle f, C^*(g) \rangle_{\mathbb{P}_X}$$

using the tower property. The next result characterizes maximal correlation as the second largest singular value of C.

[1]By the Riesz representation theorem, every bounded linear operator between separable Hilbert spaces has a unique adjoint operator with equal operator norm [Stein and Shakarchi, 2005].

Theorem 1 (Maximal Correlation as a Singular Value [Rényi, 1959]). *Given jointly distributed random variables X and Y as defined earlier, we have:*

$$\rho_{max}(X;Y) = \sup_{\substack{f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X) \\ \mathbb{E}[f(X)]=0}} \frac{\|C(f)\|_{\mathbb{P}_Y}}{\|f\|_{\mathbb{P}_X}}$$

where the supremum is achieved by some $f^\star \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X)$ if C is a compact operator.[2]

Proof. We follow the proof in [Rényi, 1959]. Observe that for $f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X)$ and $g \in \mathcal{L}^2(\mathcal{Y},\mathbb{P}_Y)$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}\left[f^2(X)\right] = \mathbb{E}\left[g^2(Y)\right] = 1$, we have:

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[g(Y)\mathbb{E}[f(X)|Y]] \leq \sqrt{\mathbb{E}\left[\mathbb{E}[f(X)|Y]^2\right]} = \|C(f)\|_{\mathbb{P}_Y}$$

using the tower property and the Cauchy-Schwarz inequality. Hence, we have:

$$\rho_{\mathsf{max}}(X;Y) \leq \sup_{\substack{f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X) \\ \mathbb{E}[f(X)]=0 \\ \mathbb{E}\left[f^2(X)\right]=1}} \|C(f)\|_{\mathbb{P}_Y} = \sup_{\substack{f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X) \\ \mathbb{E}[f(X)]=0}} \frac{\|C(f)\|_{\mathbb{P}_Y}}{\|f\|_{\mathbb{P}_X}}.$$

On the other hand, we have:

$$\|C(f)\|_{\mathbb{P}_Y}^2 = \mathbb{E}\left[\mathbb{E}[f(X)|Y]^2\right] = \mathbb{E}[\mathbb{E}[f(X)|Y]\, f(X)] \leq \|\mathbb{E}[f(X)|Y]\|_{\mathbb{P}_Y}\, \rho_{\mathsf{max}}(X;Y)$$

using the tower property, Definition 1, and the fact that $h(Y) = \mathbb{E}[f(X)|Y] \,/\, \|\mathbb{E}[f(X)|Y]\|_{\mathbb{P}_Y}$ satisfies $\mathbb{E}[h(Y)] = 0$ and $\mathbb{E}\left[h^2(Y)\right] = 1$. This implies that:

$$\sup_{\substack{f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X) \\ \mathbb{E}[f(X)]=0 \\ \mathbb{E}\left[f^2(X)\right]=1}} \|C(f)\|_{\mathbb{P}_Y} = \sup_{\substack{f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X) \\ \mathbb{E}[f(X)]=0}} \frac{\|C(f)\|_{\mathbb{P}_Y}}{\|f\|_{\mathbb{P}_X}} \leq \rho_{\mathsf{max}}(X;Y)$$

which completes the proof. Note that if C is compact, then the supremum is achievable (see [Stein and Shakarchi, 2005]). ■

Let us assume for the remainder of our discussion that C is compact. Then, it can be inferred from the preceding proof that the functions $f^\star(X)$ and $g^\star(Y)$ that maximize correlation in Definition 1 satisfy:

$$\rho_{\mathsf{max}}(X;Y) f^\star(X) = \mathbb{E}[g^\star(Y)|X] \;\; a.s. \tag{8}$$

$$\rho_{\mathsf{max}}(X;Y) g^\star(Y) = \mathbb{E}[f^\star(X)|Y] \;\; a.s. \tag{9}$$

which means that they are minimum mean-squared error estimators of each other. Furthermore, notice that the everywhere unity function $\mathbf{1}_\mathcal{X} \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X)$ (defined as $\mathbf{1}_\mathcal{X}(x) = 1$ for every $x \in \mathcal{X}$) is a right singular vector of C with corresponding left singular vector $\mathbf{1}_\mathcal{Y}$:

$$C(\mathbf{1}_\mathcal{X}) = \mathbf{1}_\mathcal{Y} \tag{10}$$

$$C^*(\mathbf{1}_\mathcal{Y}) = \mathbf{1}_\mathcal{X} \tag{11}$$

where the singular value is $\|C\|_{\mathsf{op}} = 1$. The orthogonal complement of $\mathsf{span}(\mathbf{1}_\mathcal{X})$ is the sub-Hilbert space $\left\{f \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X) : \mathbb{E}[f(X)] = 0\right\} \subset \mathcal{L}^2(\mathcal{X},\mathbb{P}_X)$. Maximizing the Rayleigh quotient $\|C(f)\|_{\mathbb{P}_Y}^2 \,/\, \|f\|_{\mathbb{P}_X}^2$ over every function f in this sub-Hilbert space produces the second largest squared singular value of C with $f^\star$ being the corresponding right singular vector. Therefore, maximal correlation is indeed the second largest singular value of C, and the functions that maximize correlation are the corresponding singular vectors. We remark that the "fixed point" equations, (8) and (9), and the linear algebraic interpretation of maximal correlation have allowed

[2]The Banach space of compact operators is the closure of the subspace of finite rank operators with respect to the operator norm [Stein and Shakarchi, 2005]. So, compact operators have "nice" spectral structure that parallels the spectral structure of matrices.

statisticians to develop algorithms that compute maximal correlation and its optimizing functions in infinite dimensional settings [Breiman and Friedman, 1985].[3]

There is no inherent reason to restrict our attention to the pair of singular vectors corresponding to the second largest singular value of the conditional expectation operator C. Inspired by the linear algebraic view of maximal correlation as a singular value, we may impart other pairs of singular vectors of C with similar operational interpretations. The pair of singular vectors corresponding to maximal correlation are the functions that are regularized (zero mean and unit variance) and maximally correlated in the Pearson correlation sense. Likewise, the pair singular vectors corresponding to the kth largest largest singular value (for $k \geq 2$) are the regularized functions that are maximally correlated subject to being orthogonal to all the previous pairs of singular vectors. Under regularity conditions, we may decompose any pair of regularized functions $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ into components corresponding to the right and left singular vectors of C, respectively. This provides a rather intriguing decomposition of the dependence structure of X and Y that is potentially useful in statistical learning applications, but we do not delve into it here for the sake of brevity. In the remainder of this discourse, we will refer to the pairs of singular vectors corresponding to the second largest and smaller singular values of C as *maximal correlation functions*.

It is a compelling theoretical question to analyze for what choices of joint distributions $P_{X,Y}$ are the associated pairs of maximal correlation functions orthonormal polynomials with respect to $\mathbb{P}_X$ and $\mathbb{P}_Y$. In the next section, we introduce natural exponential families with quadratic variance functions and their conjugate prior families. It will turn out that for joint distributions defined using these families, the maximal correlation functions are indeed orthonormal polynomials.

III. Natural Exponential Families and Conjugate Priors

We now introduce the notion of exponential families of distributions, which are popular in statistics because they have a very analytically tractable form and are intimately tied to several theoretical phenomena. For example, they form the "correct" model for efficient estimation,[4] and they can be used to understand large deviation exponents.[5] Such relationships and much of our ensuing discussion in this section can be found in [Keener, 2010] and [Wornell, 2015], and we will use these resources without tediously referring back to them every time. The next definition presents a subclass of exponential families known as natural exponential families, which will be pertinent to our discussion.

Definition 2 (Natural Exponential Family)**.** Given $(\mathcal{Y} \subseteq \mathbb{R}, \mathcal{B}(\mathcal{Y}), \mu)$, the parametrized family of probability densities $\{P_Y(\cdot; x) : x \in \mathcal{X}\}$ with respect to μ is called a *natural exponential family* when each density has the form:[6]

$$\forall y \in \mathcal{Y}, \;\; P_Y(y; x) = \exp\left(xy - \alpha(x) + \beta(y)\right)$$

where the probability density $P_Y(y; 0) = \exp\left(\beta(y)\right)$ is called the *base distribution*, and:

$$\forall x \in \mathcal{X}, \;\; \alpha(x) = \log\left(\int_{\mathcal{Y}} \exp\left(xy + \beta(y)\right) \, d\mu(y)\right)$$

is known as the *log-partition function* with $\alpha(0) = 0$ without loss of generality. The *natural parameter space* $\mathcal{X} \subseteq \mathbb{R}$ is defined as:

$$\mathcal{X} \triangleq \{x \in \mathbb{R} : |\alpha(x)| < +\infty\}$$

which is the largest interval where the log-partition function is finite.

[3] Such algorithms start with some arbitrary function and alternatively compute the conditional expectation given X and the conditional expectation given Y in a manner analogous to the *power iteration method* in numerical linear algebra. The challenge is to simultaneously achieve consistency (produce the true optimizing functions in the asymptotic limit of infinite data) and convergence (produce some solution after a reasonable number of steps) in the infinite dimensional setting. Since conditional expectations are difficult to estimate given real data, Breiman and Friedman employ various *data smoothers* in [Breiman and Friedman, 1985].

[4] Estimation is called *efficient* when the *Cramér-Rao bound* is met with equality.

[5] Exponential families provide a geometric interpretation for the equality between the classical *Chernoff exponent* and its dual representation as the infimum of *relative entropy*.

[6] The support of the probability densities in the family does not depend on the parameter.

We remark that the log-partition function $\alpha : \mathcal{X} \to \mathbb{R}$ is infinitely differentiable on $\mathcal{X}^\circ$ (the interior of $\mathcal{X}$), and satisfies the following properties:

$$\forall x \in \mathcal{X}, \;\; \alpha(x) = \log\left(\mathbb{E}_{P_Y(\cdot;0)}\left[\exp\left(xY\right)\right]\right) \tag{12}$$

$$\forall x \in \mathcal{X}^\circ, \;\; \alpha'(x) = \mathbb{E}_{P_Y(\cdot;x)}\left[Y\right] \tag{13}$$

$$\forall x \in \mathcal{X}^\circ, \;\; \alpha''(x) = \mathbb{VAR}_{P_Y(\cdot;x)}\left(Y\right) \tag{14}$$

where (12) illustrates that the log-partition function is in fact a cumulant generating function.[7] Morris specialized Definition 2 further and introduced the notion of a *natural exponential family with quadratic variance function* in an effort to justify why distributions like the Gaussian, Poisson, binomial, gamma, and negative binomial enjoy "many useful mathematical properties" [Morris, 1982]. Following the exposition in [Morris, 1982], we let the expected value of Y with respect to $P_Y(\cdot;x)$ be $\gamma : \mathcal{X} \to \mathcal{M} \subseteq \mathbb{R}$:[8]

$$\forall x \in \mathcal{X}, \;\; \gamma(x) \triangleq \mathbb{E}_{P_Y(\cdot;x)}\left[Y\right] = \int_{\mathcal{Y}} y P_Y(y;x)\, d\mu(y) = \alpha'(x) \tag{15}$$

where the final equality follows from (13). The variance of Y with respect to $P_Y(\cdot;x)$ as a function of γ is known as the *variance function*, $V : \mathcal{M} \to \mathbb{R}^+$, and is defined as:

$$\forall \gamma \in \mathcal{M}, \;\; V\left(\gamma\right) \triangleq \mathbb{VAR}_{P_Y(\cdot;x)}\left(Y\right) = \int_{\mathcal{Y}} \left(y-\gamma\right)^2 P_Y(y;x)\, d\mu(y) = \alpha''(x)$$

where V is a well-defined function of γ, because $\gamma : \mathcal{X} \to \mathcal{M}$ is an injective function.[9] *A natural exponential family has quadratic variance function if V is a polynomial in γ with degree at most* 2. Morris showed that there are only six such families of distributions [Morris, 1982]:[10]

1) Gaussian pdfs with expectation as the parameter and fixed variance,
2) Poisson pmfs with rate as the parameter,
3) binomial pmfs with success probability as the parameter and fixed number of Bernoulli trials,
4) gamma pdfs with rate as the parameter and fixed "shape,"
5) negative binomial pmfs with success probability as the parameter and fixed "number of failures,"
6) generalized hyperbolic secant pdfs (see [Morris, 1982] for details regarding this family).

These families share several properties, such as the existence and finiteness of all moments, closure under convolutions, and infinite divisibility (for all the families except binomial pmfs which are only divisible as they have bounded support) [Morris, 1982]. For any such natural exponential family, we may define a conjugate prior family as shown next.

Definition 3 (Conjugate Prior)**.** Given the natural exponential family in Definition 2, and assuming that $\mathcal{X} \subseteq \mathbb{R}$ is a non-empty open interval defining the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \lambda)$, the corresponding *conjugate prior* family, $\{P_X(\cdot;y',n) : (y',n) \in \Xi\}$, is the parametrized family of probability densities with respect to λ with *hyper-parameters* y' and n that have the form:

$$\forall x \in \mathcal{X}, \;\; P_X(x;y',n) = \exp\left(y'x - n\alpha(x) - \tau(y',n)\right)$$

where $\tau : \Xi \to \mathbb{R}$ is the *log-partition function*:

$$\forall (y',n) \in \Xi, \;\; \tau(y',n) = \log\left(\int_{\mathcal{X}} \exp\left(y'x - n\alpha(x)\right)\, d\lambda(x)\right)$$

and the *hyper-parameter space* $\Xi \subseteq \mathbb{R} \times \mathbb{R}$ is defined as:

$$\Xi \triangleq \left\{(y',n) \in \mathbb{R} \times \mathbb{R} : |\tau(y',n)| < +\infty\right\}.$$

[7] It is worth noting that the variance in (14) is known as the *Fisher information* in statistics.

[8] Here, $\mathcal{M}$ is the set of all possible expected values of Y as x varies over $\mathcal{X}$.

[9] The function $\gamma : \mathcal{X} \to \mathcal{M}$ is injective because we will assume that $\alpha''(x) > 0$ for every $x \in \mathcal{X}^\circ$. This assumption is reasonable since $\alpha''(x)$ is the variance of Y with respect to $P_Y(\cdot;x)$ as indicated in (14).

[10] The canonical parametrizations of these families (for example, Poisson pmf with its rate) do not necessarily coincide with their "natural" parametrizations as natural exponential families. However, we will not require the natural parametrizations in our discourse.

Conjugate priors can be construed as the "eigenfunctions" of the operation of computing the posterior distribution from the prior distribution when the likelihoods are given by an exponential family. Suppose our conditional probability densities with respect to μ are defined by a natural exponential family:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; P_{Y|X}(y|x) = \exp\left(xy - \alpha(x) + \beta(y)\right) = P_Y(y;x) \tag{16}$$

and the prior probability density with respect to λ belongs to the corresponding conjugate prior family:

$$\forall x \in \mathcal{X}, \;\; P_X(x) = \exp\left(y'x - n\alpha(x) - \tau(y', n)\right) = P_X(x; y', n). \tag{17}$$

Then, the posterior probability density with respect to λ is:

$$\begin{aligned} \forall y \in \mathcal{Y}, \forall x \in \mathcal{X}, \;\; P_{X|Y}(x|y) &= \exp\left((y'+y)x - (n+1)\alpha(x) - \tau(y'+y, n+1)\right) \\ &= P_X(x; y'+y, n+1) \end{aligned} \tag{18}$$

which also belongs to the conjugate prior family. This structure is extremely useful in sequential Bayesian inference problems, where statisticians use an exponential family likelihood model and a conjugate prior to allow for efficient updating of beliefs (or posterior distributions) as samples of a Markov chain (with the likelihoods defining its kernel) are observed in sequence.

In the ensuing sections, we will illustrate that letting the conditional probability distributions $P_{Y|X}$ be the Gaussian, Poisson, or binomial natural exponential families with quadratic variance functions and the marginal distribution P_X be the corresponding conjugate priors produces maximal correlation functions that are orthonormal polynomials. The roles of Hermite, Laguerre, and Jacobi will be played by the Gaussian, Poisson, and binomial conditional distributions, respectively. It turns out that when the remaining three natural exponential families with quadratic variance functions (recall that there are only six) are used to define $P_{Y|X}$ with P_X belonging to the associated conjugate prior family, some of the moments are either infinite or do not exist. For example, let $\mathcal{X} = (0,1)$ with λ as the Lebesgue measure, $\mathcal{Y} = \mathbb{N}$ with μ as the counting measure, and define:

$$\forall x \in (0,1), \forall y \in \mathbb{N}, \;\; P_{Y|X}(y|x) = x(1-x)^y \tag{19}$$

which is a geometric (special case of negative binomial) natural exponential family with quadratic variance function. The conjugate prior for this family is the beta distribution, and we set the marginal distribution of X as the uniform distribution (which is a beta distribution):

$$\forall x \in (0,1), \;\; P_X(x) = 1. \tag{20}$$

The marginal distribution of Y is:

$$\forall y \in \mathbb{N}, \;\; P_Y(y) = \int_{(0,1)} x(1-x)^y \, d\lambda(x) = \int_{(0,1)} \frac{(1-x)^{y+1}}{y+1} \, d\lambda(x) = \frac{1}{(y+1)(y+2)} \tag{21}$$

where the second equality follows using integration by parts. Evidently, the first and higher order moments of Y are infinite as the harmonic series diverges. So, the Hilbert space $\mathcal{L}^2\left(\mathbb{N}, \mathbb{P}_Y\right)$ does not contain polynomials with degree strictly greater than zero, and we cannot hope for maximal correlation functions that are orthonormal polynomials. Therefore, there are only three natural exponential families with quadratic variance functions that induce joint distributions with finite moments, and they can be classified as Hermite, Laguerre, and Jacobi.

IV. The Hermite Case

In this subsection, we let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$, and let λ and μ be the Lebesgue measure. We then set the conditional pdfs $\left\{P_{Y|X=x} = \mathcal{N}(x, \nu) : x \in \mathbb{R}\right\}$ to be the natural exponential family with quadratic variance function of Gaussian pdfs:

$$\forall x, y \in \mathbb{R}, \;\; P_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(y-x)^2}{2\nu}\right) \tag{22}$$

where $x \in \mathbb{R}$ is the expectation parameter and $\nu \in (0, \infty)$ is some fixed variance. The corresponding conjugate prior family is also Gaussian:

$$\forall x \in \mathbb{R}, \;\; P_X(x; a, p) = \frac{1}{\sqrt{2\pi p}} \exp\left(-\frac{(y-a)^2}{2p}\right) \tag{23}$$

with expectation hyper-parameter $a \in \mathbb{R}$ and variance hyper-parameter $p \in (0, \infty)$. We select the marginal pdf of X to have expectation $a = 0$:

$$\forall x \in \mathbb{R}, \;\; P_X(x) = P_X(x; 0, p). \tag{24}$$

The marginal pdf of Y is then $P_Y = \mathcal{N}(0, p + \nu)$. The ensuing theorem presents the SVD of the conditional expectation operator $C = \mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X) \to \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$ defined by this joint distribution.

Theorem 2 (Hermite SVD)**.** *For the Gaussian likelihood* (22) *with Gaussian prior* (24)*, the conditional expectation operator C has SVD:*

$$\forall k \in \mathbb{N}, \;\; C\left(H_k^{(p)}\right) = \sigma_k H_k^{(p+\nu)}$$

where $\{\sigma_k \in (0, 1] : k \in \mathbb{N}\}$ are the singular values such that $\sigma_0 = 1$ and $\lim_{k\to\infty} \sigma_k = 0$, and for $r \in (0, \infty)$, $\{H_k^{(r)}$ with degree $k : k \in \mathbb{N}\}$ denote the Hermite polynomials *that are orthonormal with respect to the Gaussian distribution $\mathcal{N}(0, r)$.*

The pairs of singular vectors of C (excluding the first pair of singular vectors) delineated in Theorem 2 are precisely the pairs of maximal correlation functions of $P_{X,Y}$. Thus, maximal correlation functions are Hermite polynomials[11] when the conditional distributions $P_{Y|X}$ belong to the Gaussian natural exponential family with quadratic variance function. We note that Theorem 2 itself is well-known outside the context of exponential families and maximal correlation functions (see [Abbe and Zheng, 2012] for example). To prove it, we will require the following result which provides simple necessary and sufficient conditions for a conditional expectation operator to have orthonormal polynomial singular vectors.

Theorem 3 (Conditional Moment Conditions)**.** *Suppose we are given the infinite dimensional separable Hilbert spaces $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ that have unique (up to arbitrary sign changes) orthonormal polynomial bases $\{p_k$ with degree $k : k \in \mathbb{N}\}$ and $\{q_k$ with degree $k : k \in \mathbb{N}\}$, respectively.[12] Suppose further that $C = \mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is compact. Then, for every $n \in \mathbb{N}$, $\mathbb{E}[X^n|Y]$ is a polynomial in Y with degree n, and $\mathbb{E}[Y^n|X]$ is a polynomial in X with degree n if and only if C has SVD:*

$$\forall k \in \mathbb{N}, \;\; C(p_k) = \beta_k q_k$$

where $\{\beta_k \in (0, 1] : k \in \mathbb{N}\}$ are the singular values such that $\beta_0 = 1$ and $\lim_{k\to 0} \beta_k = 0$.

Proof. Recall that the adjoint operator of C is $C^* = \mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$. Suppose for every $n \in \mathbb{N}$, $\mathbb{E}[X^n|Y]$ is a polynomial in Y with degree n, and $\mathbb{E}[Y^n|X]$ is a polynomial in X with degree n. This implies that C and C^* are invariant over polynomials, and preserve the degree of their input polynomial. Let us construct the Gramian operator $C^*C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, which is compact, self-adjoint, and positive.[13] Moreover, C^*C is also invariant over polynomials and preserves the degree of its input polynomial. By the spectral theorem for compact self-adjoint operators [Stein and Shakarchi, 2005], C^*C has a countable orthonormal eigenbasis $\{r_i : i \in \mathbb{N}\}$:

$$\forall i \in \mathbb{N}, \; C^*C(r_i) = \alpha_i r_i$$

where α_i are real eigenvalues such that $\alpha_i \to 0$ as $i \to \infty$. We will prove by induction that these eigenfunctions are orthonormal polynomials.

The first eigenfunction of C^*C must be $r_0 = p_0$ since C^*C preserves degrees of input polynomials. Assume that $r_i = p_i$ for $i \in \{0, \ldots, k\}$. Then, since p_{k+1} is orthogonal to $\mathsf{span}(r_0, \ldots, r_k)$, we have:

$$p_{k+1} = \sum_{j=k+1}^{\infty} \langle p_{k+1}, r_j \rangle_{\mathbb{P}_X} r_j$$

[11] The Hermite polynomials and several other orthogonal polynomial families we will use in our discourse are expounded in [Andrews and Askey, 1985].

[12] Note that $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ is infinite dimensional when $\mathcal{X}$ is an infinite set, and separable when it has a countable orthonormal Schauder basis [Stein and Shakarchi, 2005].

[13] "Positive" operators are the analog of positive semidefinite matrices.

where all such equalities hold in the $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$-norm sense. Applying C^*C to both sides and using the continuity of C^*C (which is equivalent to the boundedness of C^*C [Stein and Shakarchi, 2005]), we get:

$$C^*C(p_{k+1}) = \sum_{j=k+1}^{\infty} \alpha_j \langle p_{k+1}, r_j \rangle_{\mathbb{P}_X} r_j.$$

Hence, $C^*C(p_{k+1})$ is orthogonal to $\mathsf{span}(p_0, \ldots, p_k)$ using the continuity of the inner product. Since $C^*C(p_{k+1})$ is a polynomial with degree $k+1$ (as C^*C preserves degrees), we must have $C^*C(p_{k+1}) = \alpha_{k+1} p_{k+1}$, where $\alpha_{k+1} > 0$ as C^*C is a positive operator. So, $r_{k+1} = p_{k+1}$, and induction gives us that $\{p_k \text{ with degree } k : k \in \mathbb{N}\}$ are the eigenfunctions of C^*C:

$$\forall k \in \mathbb{N}, \;\; C^*C(p_k) = \alpha_k p_k$$

where the eigenvalues are $\alpha_k > 0$ for every $k \in \mathbb{N}$.

Now observe that $\{C(p_k) : k \in \mathbb{N}\}$ are also a set of orthogonal polynomials in $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, where $C(p_k)$ is a polynomial with degree k because C preserves degrees, and the polynomials satisfy the orthogonality relation:

$$\forall j, k \in \mathbb{N}, \;\; \langle C(p_j), C(p_k) \rangle_{\mathbb{P}_Y} = \langle p_j, C^*C(p_k) \rangle_{\mathbb{P}_X} = \alpha_k \langle p_j, p_k \rangle_{\mathbb{P}_X} = \alpha_k \delta_{jk}$$

where δ_{jk} is the Kronecker delta function. Hence, we must have:

$$\forall k \in \mathbb{N}, \;\; C(p_k) = \beta_k q_k$$

which is the SVD of C, with singular values $\beta_k = \sqrt{\alpha_k} > 0$ for every $k \in \mathbb{N}$. Finally, recall from (6) that $\|C\|_{\mathsf{op}} = 1$, and from (10) that the corresponding right and left singular vectors for the largest singular value of unity are $p_0 = \mathbf{1}_{\mathcal{X}}$ and $q_0 = \mathbf{1}_{\mathcal{Y}}$, respectively. This shows that $\beta_k \in (0, 1]$ for every $k \in \mathbb{N}$, $\beta_0 = 1$, and $\beta_k \to 0$ as $k \to \infty$ because $\alpha_k \to 0$. This completes the proof of the forward direction.

To prove the converse direction, notice that C having SVD:

$$\forall k \in \mathbb{N}, \;\; C(p_k) = \beta_k q_k$$

implies that C^* has SVD:

$$\forall k \in \mathbb{N}, \;\; C^*(q_k) = \beta_k p_k.$$

This is an exercise in functional analysis. Since monomials can be decomposed into weighted sums of orthonormal polynomials, we have that for every $n \in \mathbb{N}$, $\mathbb{E}[X^n|Y]$ is a polynomial in Y with degree n, and $\mathbb{E}[Y^n|X]$ is a polynomial in X with degree n. This completes the proof. ■

Theorem 3 can be extended appropriately to include scenarios where $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ or $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ are finite dimensional, but we omit this generalization for brevity. We will see one such case in Section VI. While Theorem 3 requires C to be a compact operator, we will omit explicit verifications of this in our examples. However, we remark that a common approach is to verify that C is a *Hilbert-Schmidt operator* (see [Stein and Shakarchi, 2005]). We close this section by proving Theorem 2 using Theorem 3.

Proof of Theorem 2. First observe that both C and C^* are *convolution* operators when $P_{X,Y}$ is defined by (22) and (24). Indeed we have for any $f \in \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X)$ and any $g \in \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$:

$$(C(f))(y) = \int_{\mathbb{R}} f(x) P_{X|Y}(x|y)\, d\lambda(x) = \int_{\mathbb{R}} f(x) \frac{1}{\sqrt{2\pi\left(\frac{p\nu}{p+\nu}\right)}} \exp\left(-\frac{\left(x - \frac{p}{p+\nu} y\right)^2}{2\left(\frac{p\nu}{p+\nu}\right)}\right) d\lambda(x) \qquad \mathbb{P}_Y\text{-a.e.}$$

$$(C^*(g))(x) = \int_{\mathbb{R}} g(y) P_{Y|X}(y|x)\, d\mu(y) = \int_{\mathbb{R}} g(y) \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(y-x)^2}{2\nu}\right) d\mu(y) \qquad \mathbb{P}_X\text{-a.e.}$$

where the conditional distribution $P_{X|Y}$ can be readily computed from (22) and (24). Letting $g(y) = y^n$ for any $n \in \mathbb{N}$, we have:

$$(C^*(g))(x) = \frac{1}{\sqrt{2\pi\nu}} \int_{\mathbb{R}} y^n \exp\left(-\frac{(y-x)^2}{2\nu}\right) d\mu(y) = \frac{1}{\sqrt{2\pi\nu}} \int_{\mathbb{R}} (x-y)^n \exp\left(-\frac{y^2}{2\nu}\right) d\mu(y) \qquad \mathbb{P}_X\text{-a.e.}$$

using the commutativity of convolution. Hence, $C^*(g)$ is a polynomial with degree n. Likewise, we can show that $C(f)$ is a polynomial with degree n if $f(x) = x^n$ for any $n \in \mathbb{N}$. Employing Theorem 3 then completes the proof. ■

V. The Laguerre Case

Here, we let $\mathcal{X} = (0, \infty)$ and $\mathcal{Y} = \mathbb{N}$, and let λ be the Lebesgue measure and μ be the counting measure. We then set the conditional pmfs $\left\{P_{Y|X=x} = \mathsf{Poisson}(x) : x \in (0,\infty)\right\}$ to be the natural exponential family with quadratic variance function of Poisson pmfs:

$$\forall x \in (0,\infty), \forall y \in \mathbb{N}, \quad P_{Y|X}(y|x) = \frac{x^y e^{-x}}{y!} \tag{25}$$

where $x \in (0,\infty)$ is the rate parameter. The corresponding conjugate prior family is consists of gamma pdfs, and we assume that $P_X = \mathsf{gamma}(\alpha, \beta)$:

$$\forall x \in (0,\infty), \quad P_X(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{26}$$

where $\alpha \in (0,\infty)$ is the shape hyper-parameter, $\beta \in (0,\infty)$ is the rate hyper-parameter, and $\Gamma(\cdot)$ denotes the gamma function. The marginal pmf of Y is then given by a negative binomial pmf $P_Y = \mathsf{negative\text{-}binomial}(\alpha, p = 1/(\beta+1))$:

$$\forall y \in \mathbb{N}, \quad P_Y(y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{1}{\beta+1}\right)^y \left(\frac{\beta}{\beta+1}\right)^\alpha \tag{27}$$

where $\alpha \in (0,\infty)$ is the number of failures and $p = \frac{1}{\beta+1} \in (0,1)$ is the success probability parameter. The next result presents the SVD of the conditional expectation operator $C = \mathbb{E}[\cdot|Y] : \mathcal{L}^2((0,\infty), \mathbb{P}_X) \to \mathcal{L}^2(\mathbb{N}, \mathbb{P}_Y)$ associated with this joint distribution.

Theorem 4 (Laguerre SVD)**.** *For the Poisson likelihood* (25) *with gamma prior* (26)*, the conditional expectation operator C has SVD:*

$$\forall k \in \mathbb{N}, \quad C\left(L_k^{(\alpha,\beta)}\right) = \sigma_k M_k^{\left(\alpha, \frac{1}{\beta+1}\right)}$$

where $\{\sigma_k \in (0,1] : k \in \mathbb{N}\}$ are the singular values such that $\sigma_0 = 1$ and $\lim_{k\to\infty} \sigma_k = 0$, $\{L_k^{(\alpha,\beta)}$ with degree $k : k \in \mathbb{N}\}$ are the Laguerre polynomials *that are orthonormal with respect to the gamma distribution $\mathbb{P}_X$* (26)*, and $\{M_k^{(\alpha,1/(\beta+1))}$ with degree $k : k \in \mathbb{N}\}$ are the* Meixner polynomials *that are orthonormal with respect to the negative binomial distribution $\mathbb{P}_Y$* (27)*.*

Proof. First notice that given $X = x > 0$, Y is Poisson distributed with rate x as shown in (25). This means that the cumulants of $P_{Y|X=x}$ are all equal to x. Since the nth moment $\mathbb{E}[Y^n|X=x]$ for $n \in \mathbb{N}$ is a polynomial in the first n cumulants with degree n (see [Kendall, 1945]), $\mathbb{E}[Y^n|X]$ is a polynomial in X with degree n for every $n \in \mathbb{N}$.

Next, observe that the posterior pdfs $\left\{P_{X|Y=y} = \mathsf{gamma}(\alpha+y, \beta+1) : y \in \mathbb{N}\right\}$ are also gamma pdfs with updated parameters. We omit this straightforward calculation, but remark that the result is unsurprising since the key property of conjugate priors is that the posterior and the prior lie in the same family. We now compute for any fixed $Y = y \in \mathbb{N}$ and any $n \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}[X^n|Y=y] &= \int_{(0,\infty)} x^n \frac{(\beta+1)^{\alpha+y} x^{\alpha+y-1} e^{-(\beta+1)x}}{\Gamma(\alpha+y)}\, d\lambda(x) \\
&= \frac{(\beta+1)^{\alpha+y}}{\Gamma(\alpha+y)} \int_{(0,\infty)} x^{\alpha+y+n-1} e^{-(\beta+1)x}\, d\lambda(x) \\
&= \frac{(\beta+1)^{\alpha+y-1}}{\Gamma(\alpha+y)} \int_{(0,\infty)} \left(\frac{t}{\beta+1}\right)^{\alpha+y+n-1} e^{-t}\, d\lambda(t) \\
&= \frac{\Gamma(\alpha+y+n)}{\Gamma(\alpha+y)(\beta+1)^n} \\
&= \frac{1}{(\beta+1)^n} \prod_{k=0}^{n-1} (\alpha+y+k)
\end{aligned}$$

where the first equality uses (26) with the updated parameters of the posterior pdfs, and the third equality follows from the substitution $t = (\beta + 1)x$. Hence, $\mathbb{E}[X^n|Y]$ is a polynomial in Y with degree n for every $n \in \mathbb{N}$. As before, employing Theorem 3 completes the proof. ■

Thus, Theorem 4 illustrates that maximal correlation functions (or the pairs of singular vectors of C excluding the first pair) are Laguerre and Meixner polynomials when the conditional distributions $P_{Y|X}$ belong to the Poisson natural exponential family with quadratic variance function. We refer readers to [Andrews and Askey, 1985] for more details on Laguerre and (the lesser known) Meixner polynomials. Although the left singular vectors of C are Meixner polynomials, we refer to this result as the "Laguerre case." This is because Meixner polynomials behave like discrete Laguerre polynomials. Indeed, observe that the geometric distribution is the discrete analog of the exponential distribution, which means that a sum of α i.i.d. geometric distributions is the discrete analog of a sum of α i.i.d. exponential distributions (or an Erlang distribution), where α is a positive integer. If we generalize α to be any positive real number, then we get that the negative binomial distribution is the discrete analog of the gamma distribution. Therefore, the Meixner polynomials (which are orthonormal with respect to the negative binomial distribution) are the discrete analog of the Laguerre polynomials (which are orthonormal with respect to the gamma distribution).

VI. The Jacobi Case

Finally, we present the Jacobi case, which turns out to be the poorest fit in this Hermite, Laguerre, and Jacobi pattern. We let $\mathcal{X} = (0,1)$ and $\mathcal{Y} = [n] \triangleq \{0, \dots, n\}$ for $n \in \mathbb{N}\backslash\{0\}$, and let λ be the Lebesgue measure and μ be the counting measure. We then set the conditional pmfs $\left\{P_{Y|X=x} = \mathsf{binomial}(n, x) : x \in (0,1)\right\}$ to be the natural exponential family with quadratic vairance function of binomial pmfs:

$$\forall x \in (0,1), \forall y \in [n],\ P_{Y|X}(y|x) = \binom{n}{y} x^y (1-x)^{n-y} \tag{28}$$

where $x \in (0,1)$ is the success probability parameter and n is the fixed number of Bernoulli trials. The corresponding conjugate prior family consists of beta pdfs, and we assume that $P_X = \mathsf{beta}(\alpha, \beta)$:

$$\forall x \in (0,1),\ \ P_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \tag{29}$$

where $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$ are the shape hyper-parameters, and $\mathrm{B}(\cdot,\cdot)$ denotes the beta function. The marginal pmf of Y is then given by a beta-binomial pmf $P_Y = \mathsf{beta\text{-}binomial}(n, \alpha, \beta)$:

$$\forall y \in [n],\ \ P_Y(y) = \binom{n}{y} \frac{\mathrm{B}(\alpha + y, \beta + n - y)}{\mathrm{B}(\alpha,\beta)}. \tag{30}$$

The ensuing theorem presents the SVD of the corresponding conditional expectation operator $C = \mathbb{E}[\cdot|Y] : \mathcal{L}^2((0,1), \mathbb{P}_X) \to \mathcal{L}^2([n], \mathbb{P}_Y)$.

Theorem 5 (Jacobi SVD)**.** *For the binomial likelihood* (28) *with beta prior* (29)*, the conditional expectation operator C has SVD:*

$$\forall k \in [n],\ \ C\left(J_k^{(\alpha,\beta)}\right) = \sigma_k Q_k^{(\alpha,\beta)}$$
$$\forall k \in \mathbb{N}\backslash[n],\ \ C\left(J_k^{(\alpha,\beta)}\right) = 0$$

where $\{\sigma_k \in (0,1] : k \in [n]\}$ are the singular values such that $\sigma_0 = 1$, $\{J_k^{(\alpha,\beta)}$ with degree $k : k \in \mathbb{N}\}$ are the Jacobi polynomials *that are orthonormal with respect to the beta distribution $\mathbb{P}_X$* (29)*, and $\{Q_k^{(\alpha,\beta)}$ with degree $k : k \in [n]\}$ are the* Hahn polynomials *that are orthonormal with respect to the beta-binomial distribution $\mathbb{P}_Y$* (30)*.*

Proof. As in Theorems 2 and 4, we will prove Theorem 5 using Theorem 3. However, since $\mathcal{L}^2([n], \mathbb{P}_Y)$ has dimension $n+1$, it only uniquely identifies polynomials with degree at most n. It turns out that in this scenario, the conditions of Theorem 3 can be modified as follows: for every $m \in [n]$, $\mathbb{E}[X^m|Y]$ is a polynomial in Y with degree m, for every $m \in \mathbb{N}\backslash[n]$, $\mathbb{E}[X^m|Y]$ is a polynomial in Y with degree at most n, and for every $m \in [n]$, $\mathbb{E}[Y^m|X]$ is a polynomial in X with degree m. We now check these conditions.

First observe that given $X = x \in (0,1)$, $P_{Y|X=x} = \mathsf{binomial}(n, x)$, which means that $Y = Z_1 + \cdots + Z_n$ where $Z_1, \ldots, Z_n$ are conditionally i.i.d. Bernoulli random variables with success probability parameter x (i.e. $\mathbb{P}(Z_i = 1) = x$ and $\mathbb{P}(Z_i = 0) = 1 - x$ for $i = 1, \ldots, n$). Hence, we have for any $m \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}\left[Y^m | X = x\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i\right)^m \middle| X = x\right] \\
&= \sum_{\substack{0 \le k_1, \ldots, k_n \le m \\ k_1 + \cdots + k_n = m}} \frac{m!}{k_1! \cdots k_n!} \prod_{i=1}^{n} \mathbb{E}\left[Z_i^{k_i} \middle| X = x\right] \\
&= \sum_{\substack{0 \le k_1, \ldots, k_n \le m \\ k_1 + \cdots + k_n = m}} \frac{m!}{k_1! \cdots k_n!} x^{N(k_1, \ldots, k_n)}
\end{aligned}$$

where the second equality follows from the multinomial theorem, the third equality follows from the fact that the moments of the Bernoulli random variables are $\mathbb{E}\left[Z_i^0 | X = x\right] = 1$ and $\forall m \in \mathbb{N} \backslash \{0\}$, $\mathbb{E}\left[Z_i^m | X = x\right] = x$, and we let $N(k_1, \ldots, k_n)$ denote the number of non-zero k_i. Since $N(k_1, \ldots, k_n) \le \min\{m, n\}$ and $N(k_1, \ldots, k_n) = \min\{m, n\}$ for at least one of the terms, we have that for every $m \in [n]$, $\mathbb{E}\left[Y^m | X\right]$ is a polynomial in X with degree m.

Finally, we note that the posterior pdfs $\left\{P_{X|Y=y} = \mathsf{beta}(\alpha + y, \beta + n - y) : y \in [n]\right\}$ are also beta pdfs with updated parameters. We once again omit this straightforward calculation, but remark that the result follows from the property of conjugate priors where the posterior and the prior belong to the same family. For any fixed $Y = y \in [n]$ and any $m \in \mathbb{N}$, we have:

$$\begin{aligned}
\mathbb{E}\left[X^m | Y = y\right] &= \int_{(0,1)} x^m \frac{x^{\alpha+y-1}(1-x)^{\beta+n-y-1}}{\mathrm{B}(\alpha + y, \beta + n - y)} \, d\lambda(x) \\
&= \frac{\mathrm{B}(\alpha + y + m, \beta + n - y)}{\mathrm{B}(\alpha + y, \beta + n - y)} \\
&= \frac{\Gamma(\alpha + y + m)\,\Gamma(\beta + n - y)\,\Gamma(\alpha + y + \beta + n - y)}{\Gamma(\alpha + y + m + \beta + n - y)\,\Gamma(\alpha + y)\,\Gamma(\beta + n - y)} \\
&= \frac{\Gamma(\alpha + y + m)\,\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + y)\,\Gamma(\alpha + \beta + n + m)} \\
&= \frac{\prod_{k=0}^{m-1} (\alpha + y + k)}{\prod_{k=0}^{m-1} (\alpha + \beta + n + k)}
\end{aligned}$$

where the first equality uses (29) with the updated parameters of the posterior pdfs, and the remaining equalities are standard manipulations using the pertinent special functions. Therefore, for every $m \in \mathbb{N}$, $\mathbb{E}\left[X^m | Y\right]$ is a polynomial in Y with the required degree. This completes the proof. ■

Theorem 5 illustrates that maximal correlation functions (or the pairs of singular vectors of C excluding the first pair) are Jacobi and Hahn polynomials when the conditional distributions $P_{Y|X}$ belong to the binomial natural exponential family with quadratic variance function. As before, we refer readers to [Andrews and Askey, 1985] for further information regarding Jacobi and (the lesser known) Hahn polynomials. In particular, the Jacobi polynomials can be obtained as a limit of Hahn polynomials (see [Andrews and Askey, 1985]). Despite this limiting relation, the Hahn polynomials can only loosely be construed as the discrete analog of Jacobi polynomials. Regardless, we dub this case the "Jacobi case" for the sake of completeness in the Hermite, Laguerre, and Jacobi pattern. We note that in the special case where $\alpha = \beta = 1$, there is a strong association between the corresponding Jacobi and Hahn polynomials, because P_X is a uniform pdf over $(0, 1)$ and P_Y is a uniform pmf on $[n]$.[14]

[14]The orthonormal polynomials corresponding to the uniform pdf over $(0, 1)$ are known as *Legendre polynomials*, and the orthonormal polynomials corresponding to the uniform pmf on $[n]$ are known as *discrete Chebyshev or Gram polynomials*.

VII. Conclusion

In closing, we review the main ideas of our discussion. We first introduced a notion of statistical dependence between random variables known as maximal correlation. Then, we unveiled how the spectral structure of the conditional expectation operator $C = \mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ underlies this dependence measure, and defined the maximal correlation functions as the pairs of singular vectors of C excluding the first pair. To understand when maximal correlation functions are orthonormal polynomials, we considered well-known joint distributions with elegant statistical properties. Namely, the conditional distribution $P_{Y|X}$ was defined by a natural exponential family with quadratic variance function, and the marginal distribution P_X was defined by the corresponding conjugate prior. There turned out to be only three significant cases of joint distributions with this structure. The *Hermite case* corresponded to Gaussian $P_{Y|X}$ with Gaussian P_X, and Hermite polynomial maximal correlation functions. The *Laguerre case* corresponded to Poisson $P_{Y|X}$ with gamma P_X, and Laguerre and Meixner polynomial maximal correlation functions. Finally, the *Jacobi case* corresponded to binomial $P_{Y|X}$ with beta P_X, and Jacobi and Hahn polynomial maximal correlation functions.

Appendix A
Julia Code to Compare Pearson and Maximal Correlation

```
using PyPlot

function corrCoeff(data) # input = data with x in column 1 and y in column 2
    n = length(data[:,1])
    x = data[:,1] - (sum(data[:,1])/n)
    y = data[:,2] - (sum(data[:,2])/n)
    x = x/sqrt(sum(x.^2))
    y = y/sqrt(sum(y.^2))
    cor = sum(x.*y) # output = Pearson correlation coefficient
end

function maxCorr(data,X,Y) # inputs = data with x in col 1 and y in col 2, alphabets X and Y
    lx = length(X)
    ly = length(Y)
    Pxy = zeros(lx,ly) # x indexes rows and y indexes columns
    n = length(data[:,1])
    for i = 1:n
        indx = find(X .== data[i,1])
        indy = find(Y .== data[i,2])
        Pxy[indx,indy] = Pxy[indx,indy] + 1
    end
    Pxy = Pxy/sum(Pxy) # empirical joint distribution of data
    Px = Pxy*ones(ly,1) # empirical mariginal distribution of data
    Py = ones(1,lx)*Pxy # empirical mariginal distribution of data
    B = Pxy.*((1./sqrt(Px))*(1./sqrt(Py)))
    for r = 1:lx
        for s = 1:ly
            if isnan(B[r,s])|isinf(B[r,s])
                B[r,s] = 0 # change all NaNs or infinities to 0
            end
        end
    end
    U,S,V = svd(B)
    S[2] # output = maximal correlation
end

# Example 1 of Pearson Correlation versus Maximal Correlation
datax = -5:1:5
X = datax
datay = -datax+4.5 # linear data
Y = datay
data = [datax datay]
corr = corrCoeff(data)
mcorr = maxCorr(data,X,Y)

# Plot of Linear Data
plt[:scatter](datax,datay)
xlabel("x")
```

```
ylabel("y")
title("Pearson Correlation = $(round(corr,4)), Maximal Correlation = $(round(mcorr,4))")

# Example 2 of Pearson Correlation versus Maximal Correlation
datax = -5:1:5
X = datax
datay = datax.^2 # quadratic data
Y = (0:5).^2
data = [datax datay]
corr = corrCoeff(data)
mcorr = maxCorr(data,X,Y)

# Plot of Quadratic Data
plt[:scatter](datax,datay)
xlabel("x")
ylabel("y")
title("Pearson Correlation = $(round(corr,4)), Maximal Correlation = $(round(mcorr,4))")

# Example 3 of Pearson Correlation versus Maximal Correlation
datax = -5:1:5
X = datax
datay = datax.^3 # cubic data
Y = datay
data = [datax datay]
corr = corrCoeff(data)
mcorr = maxCorr(data,X,Y)

# Plot of Cubic Data
plt[:scatter](datax,datay)
xlabel("x")
ylabel("y")
title("Pearson Correlation = $(round(corr,4)), Maximal Correlation = $(round(mcorr,4))")

# Example 4 of Pearson Correlation versus Maximal Correlation
datax = [-5:1:5; -5:1:5]
X = -5:1:5
datay = [sqrt(25 - (-5:1:5).^2); -sqrt(25 - (-5:1:5).^2)] # circular data
Y = [sqrt(25 - (0:1:5).^2); -sqrt(25 - (0:1:4).^2)]
data = [datax datay]
corr = corrCoeff(data)
mcorr = maxCorr(data,X,Y)

# Plot of Circular Data
plt[:scatter](datax,datay)
xlabel("x")
ylabel("y")
title("Pearson Correlation = $(round(corr,4)), Maximal Correlation = $(round(mcorr,4))")
```

REFERENCES

[Abbe and Zheng, 2012] Abbe, E. and Zheng, L. (2012). A coordinate system for Gaussian networks. *IEEE Transactions on Information Theory*, 58(2):721–733.

[Anantharam et al., 2013] Anantharam, V., Gohari, A., Kamath, S., and Nair, C. (2013). On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. arXiv:1304.6133 [cs.IT].

[Andrews and Askey, 1985] Andrews, G. E. and Askey, R. (1985). *Polynômes Orthogonaux et Applications*, volume 1171 of *Lecture Notes in Mathematics*, chapter Classical Orthogonal Polynomials, pages 36–62. Springer.

[Breiman and Friedman, 1985] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598.

[Calmon et al., 2013] Calmon, F. P., Varia, M., Médard, M., Christiansen, M. M., Duffy, K. R., and Tessaro, S. (2013). Bounds on inference. In *Proceedings of the 51st Annual Allerton Conference on Communication, Control and Computing*, pages 567–574, Allerton House, UIUC, Illinois, USA.

[Edelman, 2016] Edelman, A. (2016). Random matrix theory. MIT Mathematics 18.338 Course Notes.

[Keener, 2010] Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer, New York.

[Kendall, 1945] Kendall, M. G. (1945). *The Advanced Theory of Statistics*, volume 1. Charles Griffin and Co. Ltd., London, second edition.

[Morris, 1982] Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80.

[Rényi, 1959] Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451.

[Stein and Shakarchi, 2005] Stein, E. M. and Shakarchi, R. (2005). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, volume 3 of *Princeton Lectures in Analysis*. Princeton University Press, New Jersey.

[Wornell, 2015] Wornell, G. W. (2015). Inference and information. MIT Electrical Engineering and Computer Science 6.437 Course Notes.