Maximal Correlation Functions: Hermite, Laguerre, and Jacobi

Anuran Makur

EECS Department, Massachusetts Institute of Technology

18.338 Project Presentation
9 May 2016

Outline

1. What are maximal correlation functions?
2. The Hermite, Laguerre, and Jacobi cases
3. Why are these joint distributions special?

Pearson Correlation Coefficient

Pearson Correlation Coefficient:
For two jointly distributed random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ with finite positive variance, the Pearson correlation coefficient is defined as:

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

Pearson Correlation Coefficient

Pearson Correlation Coefficient:
For two jointly distributed random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ with finite positive variance, the Pearson correlation coefficient is defined as:

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

Properties:

- $|\rho(X;Y)| = 1$ if and only if Y is almost surely a linear function of X.
- X and Y are independent implies that $\rho(X;Y) = 0$, but the converse is not true.

Maximal Correlation

Definition (Maximal Correlation [Rényi, 1959])

For two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with positive variance, the Hirschfeld-Gebelein-Rényi maximal correlation is defined as:

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

Maximal Correlation

Definition (Maximal Correlation [Rényi, 1959])

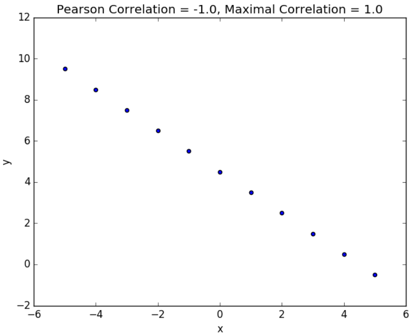
For two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with positive variance, the Hirschfeld-Gebelein-Rényi maximal correlation is defined as:

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

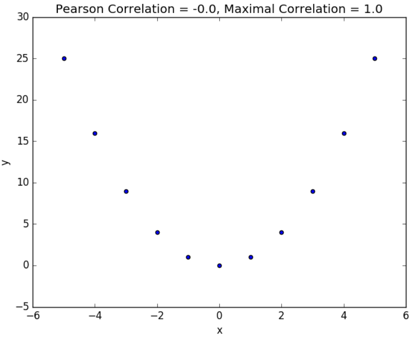
Properties:

- $0 \le \rho_{\max}(X;Y) \le 1$.
- $\rho_{\max}(X;Y) = 0$ if and only if X and Y are independent.
- $\rho_{\max}(X;Y) = 1$ if there exist functions such that $f(X) = g(Y)$ *a.s.*
- $\rho_{\max}(X;Y) = \rho_{\max}(f(X);g(Y))$ for bijective $f:\mathcal{X}\to\mathbb{R}$, $g:\mathcal{Y}\to\mathbb{R}$.
- If X and Y are jointly Gaussian, then $\rho_{\max}(X;Y) = |\rho(X;Y)|$.

Examples of Pearson versus Maximal Correlation

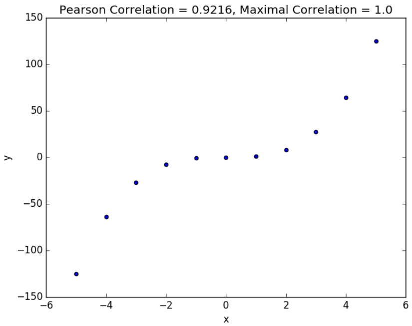


linear data

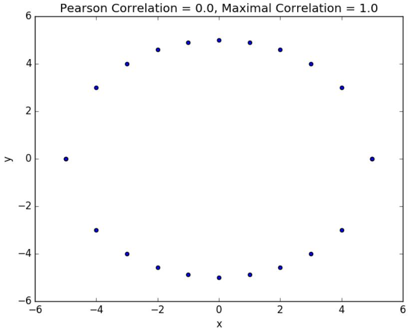


quadratic data

Examples of Pearson versus Maximal Correlation



cubic data



circular data

Maximal Correlation as a Singular Value

- Fix a joint distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R} \times \mathbb{R}$.

Maximal Correlation as a Singular Value

- Fix a joint distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R} \times \mathbb{R}$.
- Define Hilbert spaces:

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

with inner products $\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, $\langle f_1, f_2 \rangle_{\mathbb{P}_X} \triangleq \mathbb{E}[f_1(X) f_2(X)]$, and $\forall g_1, g_2 \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, $\langle g_1, g_2 \rangle_{\mathbb{P}_Y} \triangleq \mathbb{E}[g_1(Y) g_2(Y)]$, respectively.

Maximal Correlation as a Singular Value

- Fix a joint distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R} \times \mathbb{R}$.
- Define Hilbert spaces:

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

with inner products $\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, $\langle f_1, f_2 \rangle_{\mathbb{P}_X} \triangleq \mathbb{E}[f_1(X) f_2(X)]$, and $\forall g_1, g_2 \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, $\langle g_1, g_2 \rangle_{\mathbb{P}_Y} \triangleq \mathbb{E}[g_1(Y) g_2(Y)]$, respectively.
- Define conditional expectation operators, $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ and $C^* : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$:

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

with operator norms $\|C\|_{\text{op}} = \|C^*\|_{\text{op}} = 1$.

Maximal Correlation as a Singular Value

Theorem (Spectral Characterization [Rényi, 1959])

For random variables X and Y as defined earlier, we have:

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

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

Maximal Correlation as a Singular Value

Theorem (Spectral Characterization [Rényi, 1959])

For random variables X and Y as defined earlier, we have:

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

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

Interpretation:

- C has largest singular value $\|C\|_{\text{op}} = 1$ with singular vectors the constant functions $\mathbf{1}_{\mathcal{X}}$ and $\mathbf{1}_{\mathcal{Y}}$: $C\left(\mathbf{1}_{\mathcal{X}}\right) = 1.\mathbf{1}_{\mathcal{Y}}$ and $C^{*}\left(\mathbf{1}_{\mathcal{Y}}\right) = 1.\mathbf{1}_{\mathcal{X}}$.

Maximal Correlation as a Singular Value

Theorem (Spectral Characterization [Rényi, 1959])

For random variables X and Y as defined earlier, we have:

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

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

Interpretation:

- C has largest singular value $\|C\|_{\text{op}} = 1$ with singular vectors the constant functions $\mathbf{1}_\mathcal{X}$ and $\mathbf{1}_\mathcal{Y}$: $C(\mathbf{1}_\mathcal{X}) = 1.\mathbf{1}_\mathcal{Y}$ and $C^*(\mathbf{1}_\mathcal{Y}) = 1.\mathbf{1}_\mathcal{X}$.
- $f^\star \in \text{span}(\mathbf{1}_\mathcal{X})^\perp$ and $g^\star = C(f^\star)/\rho_{\max}(X;Y)$ are both functions which maximize correlation and singular vectors corresponding to $\rho_{\max}(X;Y) =$ second largest singular value of C.

Maximal Correlation Functions

Definition (Maximal Correlation Functions)

If C is compact, we refer to pairs of singular vectors of C excluding the first pair of constant functions as maximal correlation functions.

Maximal Correlation Functions

Definition (Maximal Correlation Functions)

If C is compact, we refer to pairs of singular vectors of C excluding the first pair of constant functions as maximal correlation functions.

For which joint distributions $P_{X,Y}$ are maximal correlation functions orthonormal polynomials?

Outline

The Hermite Case

Gaussian Conditional Distribution: $P_{Y|X=x} = \mathcal{N}(x, \nu)$ with expectation parameter $x \in \mathbb{R}$ and fixed variance $\nu \in (0, \infty)$

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

Gaussian Marginal Distribution of X: $P_X = \mathcal{N}(0, p)$ with fixed variance $p \in (0, \infty)$

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

The Hermite Case

Gaussian Conditional Distribution: $P_{Y|X=x} = \mathcal{N}(x, \nu)$ with expectation parameter $x \in \mathbb{R}$ and fixed variance $\nu \in (0, \infty)$

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

Gaussian Marginal Distribution of X: $P_X = \mathcal{N}(0, p)$ with fixed variance $p \in (0, \infty)$

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

Gaussian Marginal Distribution of Y: $P_Y = \mathcal{N}(0, p+\nu)$

$$\forall y \in \mathbb{R}, \quad P_Y(y) = \frac{1}{\sqrt{2\pi(p+\nu)}} \exp\left(-\frac{y^2}{2(p+\nu)}\right)$$

Theorem (Hermite SVD)

For Gaussian $P_{Y|X}$ and Gaussian P_X as defined earlier, the conditional expectation operator $C : \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X) \to \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$ 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 $\lim_{k \to \infty} \sigma_k = 0$.

The Hermite Case

Theorem (Hermite SVD)

For Gaussian $P_{Y|X}$ and Gaussian P_X as defined earlier, the conditional expectation operator $C : \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X) \rightarrow \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$ 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 $\lim_{k \rightarrow \infty} \sigma_k = 0$.

Maximal Correlation Functions:

- $\{H_k^{(p)}$ with degree $k : k \in \mathbb{N}\}$ - Hermite polynomials that are orthonormal with respect to $\mathbb{P}_X$.
- $\{H_k^{(p+\nu)}$ with degree $k : k \in \mathbb{N}\}$ - Hermite polynomials that are orthonormal with respect to $\mathbb{P}_Y$.

The Laguerre Case

Poisson Conditional Distribution: $P_{Y|X=x} = \text{Poisson}(x)$ with rate parameter $x \in (0, \infty)$

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

Gamma Marginal Distribution of X: $P_X = \text{gamma}(\alpha, \beta)$ with shape parameter $\alpha \in (0, \infty)$ and rate parameter $\beta \in (0, \infty)$

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

Poisson Conditional Distribution: $P_{Y|X=x} = \text{Poisson}(x)$ with rate parameter $x \in (0, \infty)$

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

Gamma Marginal Distribution of X: $P_X = \text{gamma}(\alpha, \beta)$ with shape parameter $\alpha \in (0, \infty)$ and rate parameter $\beta \in (0, \infty)$

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

Negative Binomial Marginal Distribution of Y:
$P_Y = \text{negative-binomial}\left(p = \frac{1}{\beta+1}, \alpha\right)$ with success probability parameter $p \in (0, 1)$ and number of failures parameter $\alpha \in (0, \infty)$

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

The Laguerre Case

Theorem (Laguerre SVD)

For Poisson $P_{Y|X}$ and gamma P_X as defined earlier, the conditional expectation operator $C : \mathcal{L}^2\left((0,\infty), \mathbb{P}_X\right) \rightarrow \mathcal{L}^2\left(\mathbb{N}, \mathbb{P}_Y\right)$ has SVD:

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

where $\{\sigma_k \in (0,1] : k \in \mathbb{N}\}$ are the singular values such that $\sigma_0 = 1$ and $\lim_{k\rightarrow\infty} \sigma_k = 0$.

The Laguerre Case

Theorem (Laguerre SVD)

For Poisson $P_{Y|X}$ and gamma P_X as defined earlier, the conditional expectation operator $C : \mathcal{L}^2\left((0,\infty), \mathbb{P}_X\right) \to \mathcal{L}^2\left(\mathbb{N}, \mathbb{P}_Y\right)$ has SVD:

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

where $\{\sigma_k \in (0,1] : k \in \mathbb{N}\}$ are the singular values such that $\sigma_0 = 1$ and $\lim_{k\to\infty} \sigma_k = 0$.

Maximal Correlation Functions:

- $\{L_k^{(\alpha,\beta)}$ with degree $k : k \in \mathbb{N}\}$ - Laguerre polynomials that are orthonormal with respect to $\mathbb{P}_X$.
- $\{M_k^{(\alpha,1/(\beta+1))}$ with degree $k : k \in \mathbb{N}\}$ - Meixner polynomials that are orthonormal with respect to $\mathbb{P}_Y$.

The Jacobi Case

Binomial Conditional Distribution: $P_{Y|X=x} = \text{binomial}(n, x)$ with number of trials parameter $n \in \mathbb{N}\backslash\{0\}$ and success probability parameter $x \in (0,1)$

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

Beta Marginal Distribution of X: $P_X = \text{beta}(\alpha, \beta)$ with shape parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$

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

The Jacobi Case

Binomial Conditional Distribution: $P_{Y|X=x} = \text{binomial}(n, x)$ with number of trials parameter $n \in \mathbb{N}\backslash\{0\}$ and success probability parameter $x \in (0, 1)$

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

Beta Marginal Distribution of X: $P_X = \text{beta}(\alpha, \beta)$ with shape parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$

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

Beta-Binomial Marginal Distribution of Y: $P_Y = \text{beta-binomial}(n, \alpha, \beta)$

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

The Jacobi Case

Theorem (Jacobi SVD)

For binomial $P_{Y|X}$ and beta P_X as defined earlier, the conditional expectation operator $C : \mathcal{L}^2\left((0,1), \mathbb{P}_X\right) \rightarrow \mathcal{L}^2\left([n], \mathbb{P}_Y\right)$ 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 $\sigma_0 = 1$.

The Jacobi Case

Theorem (Jacobi SVD)

For binomial $P_{Y|X}$ and beta P_X as defined earlier, the conditional expectation operator $C : \mathcal{L}^2\left((0,1), \mathbb{P}_X\right) \rightarrow \mathcal{L}^2\left([n], \mathbb{P}_Y\right)$ 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 $\sigma_0 = 1$.

Maximal Correlation Functions:

- $\{J_k^{(\alpha,\beta)}$ with degree $k : k \in \mathbb{N}\}$ - Jacobi polynomials that are orthonormal with respect to $\mathbb{P}_X$.
- $\{Q_k^{(\alpha,\beta)}$ with degree $k : k \in [n]\}$ - Hahn polynomials that are orthonormal with respect to $\mathbb{P}_Y$.

Outline

Why are these joint distributions special?

- $P_{Y|X}$ is a natural exponential family with quadratic variance function (introduced in [Morris, 1982]):

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

Why are these joint distributions special?

- $P_{Y|X}$ is a natural exponential family with quadratic variance function (introduced in [Morris, 1982]):

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

where $P_{Y|X}(y|0) = \exp\left(\beta(y)\right)$ is the *base distribution*, $\alpha(x)$ is the *log-partition function* with $\alpha(0) = 0$, and $\mathbb{VAR}(Y|X = x)$ is a *quadratic* function of $\mathbb{E}\left[Y|X = x\right]$.

  - theoretical importance: efficient estimation, large deviation exponents
  - useful properties: (infinite) divisibility, closure under convolutions

Why are these joint distributions special?

- $P_{Y|X}$ is a natural exponential family with quadratic variance function (introduced in [Morris, 1982]):

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

- P_X belongs to the corresponding conjugate prior family:

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

where $\tau(y', n)$ is the *log-partition function*.

  - "Eigen"-Property: useful in Bayesian inference since the posterior $P_{X|Y}(x|y) = P_X(x; y' + y, n + 1)$ is in the same family as the prior

Why are these joint distributions special?

- $P_{Y|X}$ is a natural exponential family with quadratic variance function (introduced in [Morris, 1982]):

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

- P_X belongs to the corresponding conjugate prior family:

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

- There are only three such joint distribution families where all moments exist and are finite:
  - Gaussian likelihood with Gaussian prior,
  - Poisson likelihood with gamma prior,
  - binomial likelihood with beta prior.


That's all Folks!