Statistical Learning and Data Science

Xinwei Deng

xdeng@vt.edu

Department of Statistics

A Brief Review of Multivariate Normal

- Review of matrix algebra
- Multivariate normal density
- Conditional distribution
- Some inferences and applications

Review of Matrix Algebra

- Matrix trace, determinant, inverse, etc
- Matrix partition
- Kronecker product
- Vector operation
- Matrix derivative
- Spectral Decomposition

Matrix Trace, Inverse…

- Properties of Trace
  - $\mathrm{Tr(A)} = \sum_i^n a_{ii}$
  - Tr(AB) = Tr(BA), Tr(A+B) = Tr(A)+Tr(B)
- Properties of Inverse

If $|\mathcal{A}| \neq 0$ and $\mathcal{A}(p \times p)$, then the inverse $\mathcal{A}^{-1}$ exists:

$$\mathcal{A}\,\mathcal{A}^{-1} = \mathcal{A}^{-1}\,\mathcal{A} = \mathcal{I}_p.$$

For small matrices, the inverse of $\mathcal{A} = (a_{ij})$ can be calculated as

$$\mathcal{A}^{-1} = \frac{\mathcal{C}}{|\mathcal{A}|},$$

where $\mathcal{C} = (c_{ij})$ is the adjoint matrix of $\mathcal{A}$. The elements c_{ji} of $\mathcal{C}^{\top}$ are the co-factors of $\mathcal{A}$:

$$c_{ji} = (-1)^{i+j} \begin{vmatrix} a_{11} & \dots & a_{1(j-1)} & a_{1(j+1)} & \dots & a_{1p} \\ \vdots & & & & & \\ a_{(i-1)1} & \dots & a_{(i-1)(j-1)} & a_{(i-1)(j+1)} & \dots & a_{(i-1)p} \\ a_{(i+1)1} & \dots & a_{(i+1)(j-1)} & a_{(i+1)(j+1)} & \dots & a_{(i+1)p} \\ \vdots & & & & & \\ a_{p1} & \dots & a_{p(j-1)} & a_{p(j+1)} & \dots & a_{pp} \end{vmatrix}.$$

Matrix Trace, Inverse… (Cont.)

- Properties of Inverse
  - A useful equation $(\mathcal{A} - ab^\top)^{-1} = \mathcal{A}^{-1} + \dfrac{\mathcal{A}^{-1}ab^\top\mathcal{A}^{-1}}{1 - b^\top\mathcal{A}^{-1}a}.$
  - A more general form

$$(\mathbf{A}\text{-}\mathbf{BCD})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}_{-1}\mathbf{B}(\mathbf{C}^{-1} - \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}.$$

- From above, one easily get

$$(\mathbf{A}+\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{C}^{-1})\mathbf{A}^{-1},$$

$$(\mathbf{A}^{-1}+\mathbf{C}^{-1})^{-1} = (\mathbf{I} + \mathbf{CA}^{-1})^{-1}\mathbf{C} = \mathbf{C}(\mathbf{A}+\mathbf{C})^{-1}\mathbf{A}.$$

- One Useful Identity from Ridge Regression

$$(\mathbf{X}'\mathbf{X} + h_n\mathbf{I}_p)^{-1}\mathbf{X}' = \mathbf{X}'(\mathbf{X}\mathbf{X}' + h_n\mathbf{I}_n)^{-1}$$

Matrix Partition

- Suppose **A** is an $n \times n$ matrix, we write

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where $\mathbf{A}_{11}$ is $n_1 \times n_1$ matrix and $\mathbf{A}_{22}$ is $n_2 \times n_2$

- Then we can have

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \\ -A_{22}^{-1}A_{21}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}$$

- Moreover,

$$|\mathcal{A}| = |\mathcal{A}_{11}||\mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}|$$

$$|\mathcal{A}| = |\mathcal{A}_{22}||\mathcal{A}_{11} - \mathcal{A}_{12}\mathcal{A}_{22}^{-1}\mathcal{A}_{21}|$$

Matrix Determinant

- Let B be a $p \times n$ matrix, **C** is an $n \times p$ matrix and **A** is a $p \times p$ matrix, then we have

$$|\mathbf{A}+\mathbf{BC}| = |\mathbf{A}| \times |\mathbf{I}_p + \mathbf{A}^{-1}\mathbf{BC}| = |\mathbf{A}| \times |\mathbf{I}_n + \mathbf{CA}^{-1}\mathbf{B}|.$$

- Some special cases

$$|\mathbf{A} + \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}| = |\mathbf{A}| \times (1 + \boldsymbol{x}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{x}),$$

$$|\mathbf{I}_{\mathrm{p}} + \mathbf{BC}| = |\mathbf{I}_{\mathrm{n}} + \mathbf{CB}|.$$

- From the matrix partition, if

$$\mathcal{B} = \begin{pmatrix} 1 & b^\top \\ a & \mathcal{A} \end{pmatrix}$$

Then we have,

$$|\mathcal{B}| = |\mathcal{A} - ab^\top| = |\mathcal{A}||1 - b^\top \mathcal{A}^{-1} a|$$

Spectral Decomposition

Theorem 2.1 (Eigen Decomposition) *Each symmetric matrix $\mathcal{A}(p \times p)$ can be written as*

$$\mathcal{A} = \Gamma \, \Lambda \, \Gamma^\top = \sum_{j=1}^{p} \lambda_j \gamma_j \gamma_j^\top \tag{2.18}$$

where

$$\Lambda = diag(\lambda_1, \ldots, \lambda_p)$$

and where

$$\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)$$

is an orthogonal matrix consisting of the eigenvectors γ_j of $\mathcal{A}$.

Remark: it gives a framework to define the matrix square root $A^{1/2}$, matrix power A^k, matrix exponential *exp*(A) and logarithm *log*(A).

- Example: $\Sigma = \sum_{k=0}^{\infty} \boldsymbol{A}^k / k! \equiv \exp(\boldsymbol{A})$ where *exp*(A) is called the matrix exponential of A.
- The negative log-likelihood $L_n(\Sigma) = -\log|\Sigma^{-1}| + \mathrm{tr}[\Sigma^{-1} S]$. becomes

$$L_n(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}[\exp(-\boldsymbol{A})\boldsymbol{S}].$$

Matrix Rank

- Column-rank of a matrix: is the dimension of the vector space generated by its columns (i.e., the max. number of linearly independent columns).
  - rank of a matrix = the number of non-zero singular values of the matrix.
- Example: $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \end{bmatrix}$ $\Rightarrow$ Column-Rank = 2

List all combinations of columns (linear indept or not: Y/N)

$$\begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 4 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 2 & 6 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 2 & 8 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 4 & 6 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 4 & 8 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{Y} \\ \begin{bmatrix} 1 & 1 \\ 3 & 4 \\ 6 & 8 \end{bmatrix} \end{array}$$

$$\begin{array}{c} \text{N} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{N} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 2 & 4 & 8 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{N} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 4 \\ 2 & 6 & 8 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{N} \\ \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \\ 4 & 6 & 8 \end{bmatrix} \end{array} \qquad \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 6 & 8 \end{bmatrix} \ \text{N}$$

Spectral Decomposition (Con't)

THEOREM 2.2 (Singular Value Decomposition) *Each matrix $\mathcal{A}(n \times p)$ with rank r can be decomposed as*

$$\mathcal{A} = \Gamma\ \Lambda\ \Delta^{\top},$$

where $\Gamma(n \times r)$ and $\Delta(p \times r)$. Both Γ and Δ are column orthonormal, i.e., $\Gamma^{\top}\Gamma = \Delta^{\top}\Delta = \mathcal{I}_r$ and $\Lambda = diag\left(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2}\right)$, $\lambda_j > 0$. The values $\lambda_1, \ldots, \lambda_r$ are the non-zero eigenvalues of the matrices $\mathcal{A}\mathcal{A}^{\top}$ and $\mathcal{A}^{\top}\mathcal{A}$. Γ and Δ consist of the corresponding r eigenvectors of these matrices.

- Extension to sparse SVD with applications in clustering, PCA, CCA.
- Example: consider data matrix $\mathbf{X} = (x_{\mathrm{ij}})_{n\times p}$. Then SVD of data can be written as

$$\mathbf{X} = \mathbf{UDV}^T, \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_n, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_p, \quad d_1 \geqslant d_2 \geqslant \cdots \geqslant d_K > 0.$$

- It is well-known (e.g. Eckart and Young, 1936) that for any r ≤ K,

$$\sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^T = \arg \min_{\hat{\mathbf{X}} \in M(r)} \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2,$$

- The first r components give a best rank-r approximation to the matrix.

Kronecker Product

- **Defintion:** Let **A** be an $n \times p$ matrix and **B** an $m \times q$ matrix. The $mn \times pq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{1,1}\boldsymbol{B} & a_{1,2}\boldsymbol{B} & \cdots & a_{1,p}\boldsymbol{B} \\ a_{2,1}\boldsymbol{B} & a_{2,2}\boldsymbol{B} & \cdots & a_{2,p}\boldsymbol{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1}\boldsymbol{B} & a_{n,2}\boldsymbol{B} & \cdots & a_{n,p}\boldsymbol{B} \end{bmatrix}$$

is called the Kronecker product of **A** and **B**, also call as tensor product or direct product.

- **Properties:**
  - **1.** $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$
  - **2.** $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$
  - **3.** $\mathrm{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})$

Vec Operator

- **Definition:** The ***vec*** operator creates a column vector from a matrix **A** by stacking its column vectors of A = [$\boldsymbol{a}_1$, … , $\boldsymbol{a}_p$]. i.e.,

$$\operatorname{vec}(\boldsymbol{A}) = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}.$$

- **Properties:**
  - **1.** $\operatorname{vec}(\boldsymbol{AXB}) = (\boldsymbol{B}^T \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{X}).$
  - **2.** $\operatorname{vec}(\boldsymbol{AB}) = (\boldsymbol{I} \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B}) = (\boldsymbol{B}^T \otimes \boldsymbol{I}) \operatorname{vec}(\boldsymbol{A})$
  - **3.** $\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{vec}(\boldsymbol{A}^T)^T (\boldsymbol{I} \otimes \boldsymbol{B}) \operatorname{vec}(\boldsymbol{C})$
  - **4.** $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{vec}(\boldsymbol{A}^T)^T \operatorname{vec}(\boldsymbol{B})$

Frobenius Matrix Norm

- For a matrix A, its Frobenius norm is defined as

$$||A||_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\mathrm{tr}(A^H A)}$$

- Such a norm is in a similar spirit to the Euclidean norm in vector.
- Thus the trace operator play a role of inner product in matrix.
- Example

$$\begin{aligned} ||A\text{-}B||_F^2 &= \mathrm{tr}[(A\text{-}B)^T(A\text{-}B)] \\ &= ||A||_F^2 + ||B||_F^2 \text{-} 2\mathrm{tr}[(A^T B)]. \end{aligned}$$

Matrix Derivation

- Derivative from the intuition

$$\boldsymbol{f}(\boldsymbol{x}+d\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}'(\boldsymbol{x})d\boldsymbol{x} + (\text{higher order terms}).$$

- Definition from the following example

$$\frac{\operatorname{tr}(AdX)}{dX} = \frac{\operatorname{tr}\begin{bmatrix} \tilde{a}_1^T dx_1 & & \\ & \ddots & \\ & & \tilde{a}_n^T dx_n \end{bmatrix}}{dX} = \frac{\sum_{i=1}^n \tilde{a}_i^T dx_i}{dX}.$$

Thus, we have

$$\left[\frac{\operatorname{tr}(AdX)}{dX}\right]_{ij} = \left[\frac{\sum_{i=1}^n \tilde{a}_i^T dx_i}{\partial x_{ji}}\right] = a_{ij}$$

so that

$$\frac{\operatorname{tr}(AdX)}{dX} = A^T$$

Based on Definition

$$
\begin{aligned}
\mathrm{tr}AB &= \mathrm{tr}\begin{bmatrix} \longleftarrow & \vec{a_1} & \longrightarrow \\ \longleftarrow & \vec{a_2} & \longrightarrow \\ & \vdots & \\ \longleftarrow & \vec{a_n} & \longrightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{b_1} & \vec{b_2} & \cdots & \vec{b_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \\
&= \mathrm{tr}\begin{bmatrix} \vec{a_1}^T\vec{b_1} & \vec{a_1}^T\vec{b_2} & \cdots & \vec{a_1}^T\vec{b_n} \\ \vec{a_2}^T\vec{b_1} & \vec{a_2}^T\vec{b_2} & \cdots & \vec{a_2}^T\vec{b_n} \\ \vdots & & \ddots & \vdots \\ \vec{a_n}^T\vec{b_1} & \vec{a_n}^T\vec{b_2} & \cdots & \vec{a_n}^T\vec{b_n} \end{bmatrix} \\
&= \sum_{i=1}^{m} a_{1i}b_{i1} + \sum_{i=1}^{m} a_{2i}b_{i2} + \ldots + \sum_{i=1}^{m} a_{ni}b_{in} \\
\Rightarrow \quad \frac{\partial \mathrm{tr}AB}{\partial a_{ij}} &= b_{ji} \\
\Rightarrow \quad \nabla_A \mathrm{tr}AB &= B^T
\end{aligned}
$$

Matrix Derivative: Chain Rule

- Suppose $\mathbf{U} = f(\mathbf{X})$

$$\frac{\partial g(\mathbf{U})}{\partial \mathbf{X}} = \frac{\partial g(f(\mathbf{X}))}{\partial \mathbf{X}}$$

- The chain rule:

$$\frac{\partial g(\mathbf{U})}{\partial x_{ij}} = \sum_{k=1}^{M} \sum_{l=1}^{N} \frac{\partial g(\mathbf{U})}{\partial u_{kl}} \frac{\partial u_{kl}}{\partial x_{ij}}$$

$$\frac{\partial g(\mathbf{U})}{\partial X_{ij}} = \mathrm{Tr}\left[\left(\frac{\partial g(\mathbf{U})}{\partial \mathbf{U}} \right)^T \frac{\partial \mathbf{U}}{\partial X_{ij}} \right]$$

Matrix Derivative of Traces

- Assume $F(\boldsymbol{X})$ is an element-wise differentiable function
  - $f()$ is the scalar derivative of $F()$.

$$\frac{\partial \mathrm{Tr}(F(\mathbf{X}))}{\partial \mathbf{X}} = f(\mathbf{X})^T$$

- **Properties**:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}\mathbf{A}) &= \mathbf{A}^T \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A}\mathbf{X}\mathbf{B}) &= \mathbf{A}^T\mathbf{B}^T \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{A}\mathbf{X}^T\mathbf{B}) &= \mathbf{B}\mathbf{A} \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}^T\mathbf{A}) &= \mathbf{A} \\
\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{X}^2) &= 2\mathbf{X}^T
\end{aligned}$$

Matrix Derivation: More Properties

- Suppose $\mathbf{X}$ is a square and invertible matrix, then

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \det(\mathbf{X})(\mathbf{X}^{-1})^T$$

$$\frac{\partial \det(\mathbf{X}^T\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = 2\det(\mathbf{X}^T\mathbf{A}\mathbf{X})\mathbf{X}^{-T}$$

- Nonlinear forms

$$\frac{\partial \ln|\det(\mathbf{X})|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T = (\mathbf{X}^T)^{-1}$$

$$\frac{\partial \det(\mathbf{X}^k)}{\partial \mathbf{X}} = k\det(\mathbf{X}^k)\mathbf{X}^{-T}$$

- Others

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \det(\mathbf{X})(\mathbf{X}^{-1})^T$$

$$\frac{\partial \mathrm{Tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})}{\partial \mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1})^T$$

Multivariate Normal Distribution: Random Vector

- For a random vector $Y = (y_1,\ldots y_n)$, the mean vector is

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

and the covariance matrix is
$$\begin{aligned} \operatorname{cov}(\mathbf{Y}) &= E\left\{[\mathbf{Y} - E(\mathbf{Y})][\mathbf{Y} - E(\mathbf{Y})]'\right\} \\ &= E\left\{[\mathbf{Y} - \boldsymbol{\mu}][\mathbf{Y} - \boldsymbol{\mu}]'\right\} \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \Sigma \end{aligned}$$

where $\sigma_{ij} = \operatorname{cov}(Y_i, Y_j) = E\left\{[Y_i - \mu_i][Y_j - \mu_j]\right\}$.

Multivariate Random Vector

- **Proposition 1**: For two random vector **X** and **Y**,

$$1.\ \operatorname{cov}\left(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}\right) = \mathbf{a}'\Sigma_{XY}\mathbf{b}$$

$$2.\ \operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \operatorname{cov}(\mathbf{Y}, \mathbf{X})$$

$$3.\ \operatorname{cov}(\mathbf{a} + \mathbf{A}\mathbf{X}, \mathbf{b} + \mathbf{B}\mathbf{Y}) = \mathbf{A}\operatorname{cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}'$$

- **Proposition 2:** A $p \times p$ matrix is a covariance matrix *if and only if* it is positive semi-definite.
- The Mahalanobis distance between **Y** and $\boldsymbol{\mu}$ is defined as

$$D_{\Sigma}(\mathbf{Y}, \mu) = [(\mathbf{Y} - \mu)'\Sigma^{-1}(\mathbf{Y} - \mu)]^{1/2}$$

- The multivariate skewness and kurtosis measures for are

$$\beta_{1,\,p} = E\left\{(\mathbf{Y} - \mu)'\,\Sigma^{-1}\,(\mathbf{X} - \mu)\right\}^{3}$$

$$\beta_{2,\,p} = E\left\{(\mathbf{Y} - \mu)'\,\Sigma^{-1}\,(\mathbf{Y} - \mu)\right\}^{2}$$

where **Y** and **X** are independent, identically distributed (iid).

Multivariate Random Vector (CDF and PDF)

- The joint CDF (=cumulative distribution function) of a multivariate random vector **X** in R^n is

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(\mathbf{X} \leq \mathbf{x}) =$$

$$= P(X_1 \leq x_1,\ldots,X_n \leq x_n)$$

- The joint probability density function (PDF) is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$$

Moment Generating and Characteristic Functions

Definition

Moment generating function of $\mathbf{X}$ *is defined as*

$$\psi_{\mathbf{X}}(\mathbf{t}) \stackrel{\text{def}}{=} Ee^{\mathbf{t}^T\mathbf{X}} = Ee^{t_1X_1+t_2X_2+\cdots+t_nX_n}$$

Definition

Characteristic function of $\mathbf{X}$ *is defined as*

$$\varphi_{\mathbf{X}}(\mathbf{t}) \stackrel{\text{def}}{=} Ee^{i\mathbf{t}^T\mathbf{X}} = Ee^{i(t_1X_1+t_2X_2+\cdots+t_nX_n)}$$

Special cases: take $t_1 = 1, t_2 = t_3 = \ldots = t_n = 0$, then $\varphi_{\mathbf{X}}(\mathbf{t}) = \varphi_{X_1}(t_1)$.

One-dimensional Normal RV

- Suppose $x \sim N(\mu,\sigma^2)$, then the moment generating function is

$$\psi_X(t) = E\left[e^{tX}\right] = e^{t\mu + \frac{1}{2}t^2\sigma^2}$$

- The characteristic function is

$$\varphi_X(t) = E\left[e^{itX}\right] = e^{it\mu - \frac{1}{2}t^2\sigma^2}$$

Multivariate Normal Distribution

- In the univariate case,

$$f_{Y_i}(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(y_i - \mu_i)^2 / 2\sigma^2\} \quad -\infty < y_i < \infty$$

- Let $\mathbf{y} = (y_1, \ldots y_p)$. If each y_i is independent normal with mean μ_i and variance σ^2, then

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \prod_{i=1}^{p} f_{Y_i}(y_i) \\ &= \prod_{i=1}^{p} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(y_i - \mu_i)^2 / 2\sigma^2\} \\ &= (2\pi)^{-p/2} \left(\frac{1}{\sigma^p}\right) \exp\{-\sum_{i=1}^{p} (y_i - \mu_i)^2 / 2\sigma^2\} \\ &= (2\pi)^{-p/2} \left|\left(\sigma^2 \mathbf{I}_p\right)\right|^{-1/2} \exp\{-(\mathbf{y} - \boldsymbol{\mu})' \left(\sigma^2 \mathbf{I}_p\right)^{-1} (\mathbf{y} - \boldsymbol{\mu}) / 2\} \end{aligned}$$

Multivariate Normal Distribution (Con't)

Proposition (transformation): suppose $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$, the inverse transformation is $\boldsymbol{x} = \boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{b})$, then the *p.d.f.* of $\boldsymbol{y}$ is

$$f_Y(y) = \mathrm{abs}(|\mathcal{A}|^{-1}) f_X\{\mathcal{A}^{-1}(y-b)\}.$$

- Let $\mathbf{y} = (y_1,\ldots y_p)$, where y ~N(μ, Σ), then its probability density function is

$$f(\boldsymbol{y}) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\mu)^\top \Sigma^{-1} (\boldsymbol{y}-\mu)\right\}.$$

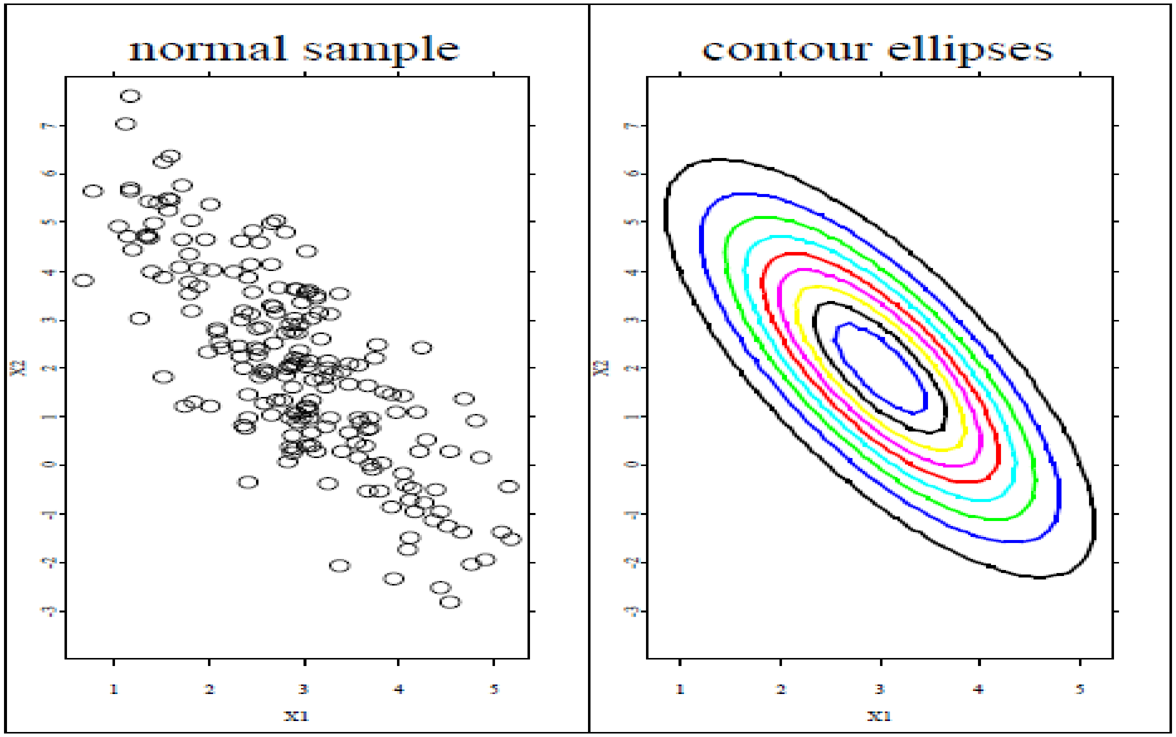
Theorem 2.3 *Let $X \sim N_p(\mu, \Sigma)$ and $\mathcal{A}(p \times p)$, $c \in \mathbb{R}^p$, where $\mathcal{A}$ is nonsingular. Then $Y = \mathcal{A}X + c$ is again a p-variate Normal, i.e.,*

$$Y \sim N_p(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top).$$

Interpretation of Multivariate Normal

- **Geometrical interpretation:** the density of $\boldsymbol{y}$ ~N(μ, Σ) distribution is constant on ellipsoids of the form

$$(y - \mu)^{\top} \Sigma^{-1} (y - \mu) = d^2$$



Interpretation of Multivariate Normal (Con't)

Theorem 2.4 *$X \sim N_p(\mu, \Sigma)$, then the variable $U = (X-\mu)^\top \Sigma^{-1}(X-\mu)$ has a χ^2_p distribution.*

Theorem 2.5 *The characteristic function (cf) of a multinormal $N_p(\mu, \Sigma)$ is given by*

$$\varphi_X(t) = \exp(\mathrm{i}\ t^\top \mu - \frac{1}{2} t^\top \Sigma t).$$

- Proof: start from simple, suppose y ~N(0, I), what is the characteristic function.
- **Proposition**: What is the moment generating function for y ~$N(\mu, \Sigma)$.

Example: Verifying Inversion Formula

- **Theorem** (Inversion Formula) If characteristic function φ_X is integrable, then CDF is absolutely continuous, then the PDF is given by

$$f(x) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} e^{-\mathbf{i}t^\top x} \varphi_X(t)\, dt.$$

- Under the multivariate normal, we can verify this theorem.

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^p} \int \exp\left(-\mathbf{i}t^\top x + \mathbf{i}t^\top \mu - \frac{1}{2} t^\top \Sigma t\right) dt \\ &= \frac{1}{|2\pi\Sigma^{-1}|^{1/2} |2\pi\Sigma|^{1/2}} \int \exp\left[-\frac{1}{2}\{t^\top \Sigma t + 2\mathbf{i}t^\top (x-\mu) - (x-\mu)^\top \Sigma^{-1} (x-\mu)\}\right] \\ &\qquad \cdot \exp\left[-\frac{1}{2}\{(x-\mu)^\top \Sigma^{-1}(x-\mu)\}\right] dt \\ &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\{(x-\mu)^\top \Sigma (x-\mu)\}\right] \end{aligned}$$

Example: Using MGF

- If X ~ $N(\mu, \Lambda)$, then the moment generating function is

$$\psi_{\mathbf{X}}(\mathbf{t}) = Ee^{\mathbf{t}^T\mathbf{X}} = e^{\mathbf{t}^T\mu + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Lambda}\mathbf{t}}.$$

Property 1

An $n \times 1$ random vector $\mathbf{X}$ has a normal distribution iff for **every** *$n \times 1$-vector $\mathbf{a}$ the one-dimensional random vector $\mathbf{a}^T\mathbf{X}$ has a normal distribution.*

- Recall Theorem 2.3:

$\mathbf{X} \in N(\mu, \boldsymbol{\Lambda})$ *and* $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$. *Then*

$$\mathbf{Y} \in N\left(B\mu + \mathbf{b}, B\boldsymbol{\Lambda}B^T\right).$$

Using MGF for Proof of Theorem 2.3

- Proof of Theorem 2.3 $\psi_{\mathbf{Y}}(\mathbf{s}) = E\left[e^{\mathbf{s}^T\mathbf{Y}}\right] = E\left[e^{\mathbf{s}^T(\mathbf{b}+B\mathbf{X})}\right] =$

$$= e^{\mathbf{s}^T\mathbf{b}} E\left[e^{\mathbf{s}^T B\mathbf{X}}\right] = e^{\mathbf{s}^T\mathbf{b}} E\left[e^{(B^T\mathbf{s})^T\mathbf{X}}\right]$$

$$E\left[e^{(B^T\mathbf{s})^T\mathbf{X}}\right] = \psi_{\mathbf{X}}\left(B^T\mathbf{s}\right).$$

- Since, we know X ~ $N(\mu, \Lambda)$

$$\psi_{\mathbf{X}}\left(B^T\mathbf{s}\right) = e^{(B^T\mathbf{s})^T\mu + \frac{1}{2}(B^T\mathbf{s})^T\mathbf{\Lambda}(B^T\mathbf{s})}.$$

$$\left(B^T\mathbf{s}\right)^T\mu = \mathbf{s}^T B\mu,$$

$$\left(B^T\mathbf{s}\right)^T\mathbf{\Lambda}\left(B^T\mathbf{s}\right) = \mathbf{s}^T B\mathbf{\Lambda}B^T\mathbf{s},$$

$$e^{(B^T\mathbf{s})^T\mu + \frac{1}{2}(B^T\mathbf{s})^T\mathbf{\Lambda}(B^T\mathbf{s})} = e^{\mathbf{s}^T B\mu + \frac{1}{2}\mathbf{s}^T B\mathbf{\Lambda}B^T\mathbf{s}}$$

Using MGF for Proof of Theorem 2.3 (Con't)

$$\psi_{\mathbf{X}}\left(B^T\mathbf{s}\right) = e^{\mathbf{s}^T B\mu + \frac{1}{2}\mathbf{s}^T B\Lambda B^T \mathbf{s}}.$$

$$\psi_{\mathbf{Y}}(\mathbf{s}) = e^{\mathbf{s}^T\mathbf{b}}\psi_{\mathbf{X}}\left(B^T\mathbf{s}\right) = e^{\mathbf{s}^T\mathbf{b}}e^{\mathbf{s}^T B\mu + \frac{1}{2}\mathbf{s}^T B\Lambda B^T \mathbf{s}}$$

$$\psi_{\mathbf{Y}}(\mathbf{s}) = e^{\mathbf{s}^T(\mathbf{b}+B\mu) + \frac{1}{2}\mathbf{s}^T B\Lambda B^T \mathbf{s}},$$

which proves the claim as asserted. $\square$

Conditional Distribution

- **Theorem:** suppose $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ follows multivariate normal $\sim N(\mu, \Sigma)$. Then the conditional distribution

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 \sim N_{p_1}\left[\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right]$$

where $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$

- Proof: using the property of matrix decomposition, we can verify the density

$$f(\boldsymbol{y}_1 \mid \boldsymbol{y}_2)\, f(\boldsymbol{y}_2) = f(\boldsymbol{y}_1, \boldsymbol{y}_2)$$

- *Remark: using* $(\boldsymbol{A}\text{-}\boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} - \boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}$

Conditional Distribution (Con't)

- Another angle of proof
  - Define $\boldsymbol{y}' = \boldsymbol{y}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2$.
  - $\text{Var}(\boldsymbol{y}') = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$
  - $\text{Cov}(\boldsymbol{y}', \boldsymbol{y}_2) = 0$
- Then $\boldsymbol{y}_1 = \boldsymbol{y}' + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2$. We have,

$$E(\boldsymbol{y}_1 | \boldsymbol{y}_2) = (\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2) + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2$$

$$\text{Var}(\boldsymbol{y}_1 | \boldsymbol{y}_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

- Therefore,

$$\boldsymbol{y}_1 | \boldsymbol{y}_2 \sim N_{p_1}\left[\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right]$$

A Few Properties

Corollary 1 *Let* $X = \binom{X_1}{X_2} \sim N_p(\mu, \Sigma)$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. $\Sigma_{12} = 0$ *if and only if* X_1 *is independent of* X_2.

Corollary 2 *If* $X \sim N_p(\mu, \Sigma)$ *and given some matrices* $\mathcal{A}$ *and* $\mathcal{B}$, *then* $\mathcal{A}X$ *and* $\mathcal{B}X$ *are independent if and only if* $\mathcal{A}\Sigma\mathcal{B}^\top = 0$.

Corollary 3 *If* $X_1 \sim N_r(\mu_1, \Sigma_{11})$ *and* $(X_2 | X_1 = x_1) \sim N_{p-r}(\mathcal{A}x_1 + b, \Omega)$ *where* Ω *does not depend on* x_1, *then* $X = \binom{X_1}{X_2} \sim N_p(\mu, \Sigma)$, *where*

$$\mu = \begin{pmatrix} \mu_1 \\ \mathcal{A}\mu_1 + b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{11}\mathcal{A}^\top \\ \mathcal{A}\Sigma_{11} & \Omega + \mathcal{A}\Sigma_{11}\mathcal{A}^\top \end{pmatrix}.$$

Inference: Conditional Independency

- Let $\mathbf{y} = (\mathbf{y}^{+}, \mathbf{y}^{*})$, where $\mathbf{y}^{+} = (y_1, y_2)$, then

$$\mathbf{Var}(\mathbf{y}^{+} | \mathbf{y}^{*}) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

- Note that $\mathbf{y}^{+}$ follows the normal distribution.
- If $\mathrm{cov}(y_1, y_2 \,|\, \mathbf{y}^{*}) = 0$, it is called conditional independence.
- **Proposition:** Let $\mathbf{y} = (\mathbf{y}^{+}, \mathbf{y}^{*})$ follows $N(\mu, \Sigma)$. Define $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ and $\boldsymbol{\Omega} = (c_{ij}) = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$. Then

$$\mathrm{cov}([y_1, y_2] \,|\, \mathbf{y}^{*}) = (K_{11})^{-1}.$$

Therefore, $\mathrm{cov}(y_1, y_2 \,|\, \mathbf{y}^{*}) = 0 \leftrightarrow c_{12} = 0$.

Inference: Regression

- Recall the linear regression $y = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ is iid normal.
- The MLE estimator of $\boldsymbol{\beta}$ is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}\mathbf{y}$.
- Define $\tilde{\boldsymbol{y}} = (y, \boldsymbol{x})'$. Assume $\tilde{y}$ is the multivariate normal $\tilde{\boldsymbol{y}} \sim N(\mu, \Sigma)$, with

$$\mu = (\mu_y, \mu_{\boldsymbol{x}})', \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}y} & \boldsymbol{\Sigma}_{\boldsymbol{xx}} \end{pmatrix}.$$

- Applying conditional distribution theorem, we have

$$\begin{aligned} E(y|\boldsymbol{x}) &= u_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}) \\ &= u_y - \boldsymbol{\mu}_x'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} + \boldsymbol{x}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}. \end{aligned}$$

Inference: Regression (Con't)

- If data are centered, i.e., $\mu_y = 0$, $\mu_{\boldsymbol{x}} = 0$, the conditional mean is $E(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$.
  - It can be linked to MLE $\boldsymbol{x}'\hat{\beta} = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- Let $\tilde{\boldsymbol{y}} = (y, \boldsymbol{x})' \sim N(0, \Sigma)$, where $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}y} & \boldsymbol{\Sigma}_{\boldsymbol{xx}} \end{pmatrix}$, and denote $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} K_{yy} & K_{y\boldsymbol{x}} \\ K_{\boldsymbol{x}y} & K_{\boldsymbol{xx}} \end{pmatrix}$, then

$$E(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{yx} = -\boldsymbol{x}'\frac{K_{xy}}{K_{yy}}$$

- Estimating $\boldsymbol{\beta}$ in linear model can be formulated by

$$\max_{\boldsymbol{\Omega}} \log|\boldsymbol{\Sigma}^{-1}| - tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) \quad \text{where } \boldsymbol{S} = \textstyle\sum_{i=1}^{n} \tilde{\boldsymbol{y}}_i'\tilde{\boldsymbol{y}}_i.$$

Inference: Gaussian Process

- A Gaussian distribution is a distribution over vectors.
- Notation: $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, specified by a mean vector and a covariance matrix.
- The position of random variable x_i in vector $\boldsymbol{x}$ plays the role of indexing.

- A Gaussian process is a distribution over functions.
- Notation: $f(x) \sim \mathrm{GP}(m(x), k(x))$, specified by a mean function and a covariance function.
- The argument of x plays the role of indexing.

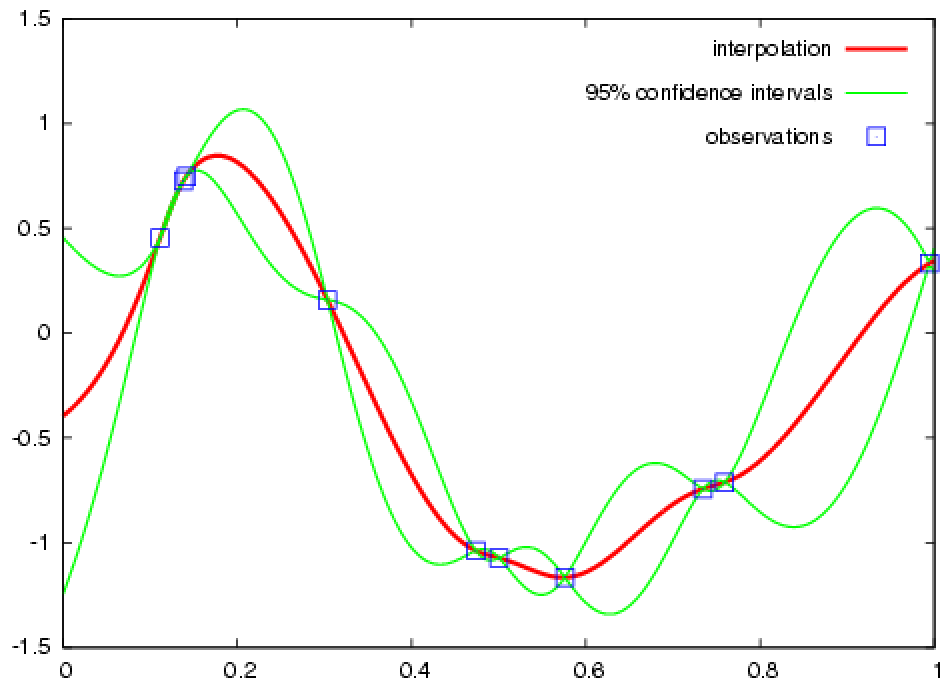
Gaussian Process: Definition

- A Gaussian Process (GP) is an infinite dimensional object.
- **Definition:** *GP is a collection of random variables, any finite number of which have joint Gaussian distributions.*
- Suppose $y = f(\boldsymbol{x})$ is a GP. Let $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$ be an n-dimensional vector of values evaluated at $\boldsymbol{x_i} \in \boldsymbol{X}$. Then,
- Each $f(\boldsymbol{x}_1)$ is a random variable with normal distribution

Proposition: $y = f(\boldsymbol{x})$ *is a Gaussian process if for any finite* subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \boldsymbol{X}$, $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ has a multivariate Gaussian distribution.

Gaussian Process: for Regression (Kriging)

- *Goal:* predict the output value y_* for a new input value $\boldsymbol{x}_*$.
- Given the training data $\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n\}$.



Gaussian Process: for Kriging (Prediction)

- A GP is fully specified by mean function and covariance function: $\boldsymbol{f} \sim GP(\boldsymbol{m}, \boldsymbol{K})$.
  - Parametric model for **m** and **K**. For example, $\mathbf{m} = \boldsymbol{x}'\boldsymbol{\beta}$, and $\boldsymbol{K}$ is specified by $$K_{ij} = k(x_i, x_j) = v_0 \exp\left\{-\frac{1}{2}\sum_{m=1}^{d} \ell_m (x_i^m - x_j^m)^2\right\}$$
- Prediction essentially is to apply conditional distribution, from $$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{m}_* \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right),$$
- Then we have $$\boldsymbol{f}_* | \boldsymbol{f} \sim \mathcal{G}\left(\boldsymbol{m}_* + \mathbf{K}_*^T \mathbf{K}^{-1}(\boldsymbol{f} - \boldsymbol{m}), \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_*\right)$$

Thank you!

- **Questions and Comments?**