CS189: Intro to Machine Learning Summer 2018

Lecture 5: Bias-variance tradeoff and multivariate gaussians

Josh Tobin
UC Berkeley EECS

Announcements

- HW2 posted

Outline for today

- Bias-variance decomposition
- Multivariate gaussians

Outline for today

- **Bias-variance decomposition**
- Multivariate gaussians

Bias-variance decomposition

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

View x_i as sampled from r.v. X

Good model: one with low *expected* error

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$\underbrace{(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2}$$

“Best” model: choice of theta
with the lowest error on D

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$\underline{(f_\theta^*(\vec{x}; \mathcal{D})} - y)^2$$

↓

"Best" model's estimate for y at the datapoint x

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$\underline{(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2}$$

Squared error between estimate for y and one choice for the observed y at datapoint x

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$\underline{\mathbb{E}_{y\sim p(y|\vec{x})}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]} \qquad p(y \mid \vec{x}) = \mathcal{N}(f(x), \sigma^2)$$

Average squared error between estimate for y and true y

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$\mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}\mathbb{E}_{y\sim p(y|\vec{x})}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] \quad p(\mathcal{D}) = \prod_{i=1}^{n} p(x_i)p(y_i \mid x_i)$$

Average over all datasets of mean squared error on that dataset

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$\mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}\mathbb{E}_{y\sim p(y|\vec{x})}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right] = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

Let's decompose this into parts

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}_{y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

Let's decompose this into parts

$$= \underbrace{(\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right] - \mathbb{E}\left[y\right])^2}_{\text{bias}^2} + \underbrace{\text{Var}(f_\theta^*(\vec{x}; \mathcal{D}))}_{\text{Variance}} + \underbrace{\text{Var}(z)}_{\text{Irreducible error}}$$

Bias-variance tradeoff

$$= \underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D}))}_{\text{Variance}} + \underbrace{\mathrm{Var}(z)}_{\text{Irreducible error}}$$

$$\underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right]}_{\text{Average over all datasets of that dataset's best model's estimate of f(x)}} - \mathbb{E}\left[y\right]\right)^2$$

Bias-variance tradeoff

$$= \underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D}))}_{\text{Variance}} + \underbrace{\mathrm{Var}(z)}_{\text{Irreducible error}}$$

$$\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \underbrace{\mathbb{E}\left[y\right]}_{\text{Average of noisy y's for a given x}}\right)^2$$

Bias-variance tradeoff

$$= \underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D}))}_{\text{Variance}} + \underbrace{\mathrm{Var}(z)}_{\text{Irreducible error}}$$

$$\underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}$$

Average error in estimate of y over all datasets

Bias-variance tradeoff

$$= \underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D}))}_{\text{Variance}} + \underbrace{\mathrm{Var}(z)}_{\text{Irreducible error}}$$

$$\underbrace{\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D}))}$$

Variance of estimate of y over all possible datasets

Bias-variance tradeoff

$$= \underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D}))}_{\text{Variance}} + \underbrace{\mathrm{Var}(z)}_{\text{Irreducible error}}$$

$$\underbrace{\mathrm{Var}(z)}_{\text{Variance of the noise term in the model for y}}$$

Derive bias-variance decomposition

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

Start by analyzing mean & var of y

$$\mathbb{E}[Y] = \mathbb{E}[f(\vec{x}) + z] = f(\vec{x}) + \mathbb{E}[z] = f(\vec{x})$$

$$\mathrm{Var}(Y) = \mathrm{Var}(f(\vec{x}) + z) = \mathrm{Var}(Z)$$

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

Add one more lemma (for any RV)

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$$

$$\implies \mathbb{E}\left[X^2\right] = \mathrm{Var}(X) + \mathbb{E}\left[X\right]^2$$

Bias-variance tradeoff

$$\epsilon(\vec{x}; f) = \mathbb{E}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

$$= \mathbb{E}(f_\theta^*(\vec{x}; \mathcal{D})^2) + \mathbb{E}(y^2) - 2\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D}) \cdot y\right]$$

$$= \left(\mathrm{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right]^2\right)$$

$$+ \left(\mathrm{Var}(Y) + \mathbb{E}\left[y\right]^2\right) - 2\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right] \cdot \mathbb{E}\left[y\right]$$

$$= \left(\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right]^2 - 2\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right] \cdot \mathbb{E}\left[y\right] + \mathbb{E}\left[y\right]^2\right)$$

$$+ \mathrm{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \mathrm{Var}(z)$$

$$= \left(\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2 + \mathrm{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \mathrm{Var}(z)$$

Facts

$$\mathbb{E}\left[X^2\right] = \mathrm{Var}(X) + \mathbb{E}\left[X\right]^2$$

$$\mathrm{Var}(Y) = \mathrm{Var}(z)$$

$$\mathbb{E}(Y) = f(\vec{x})$$

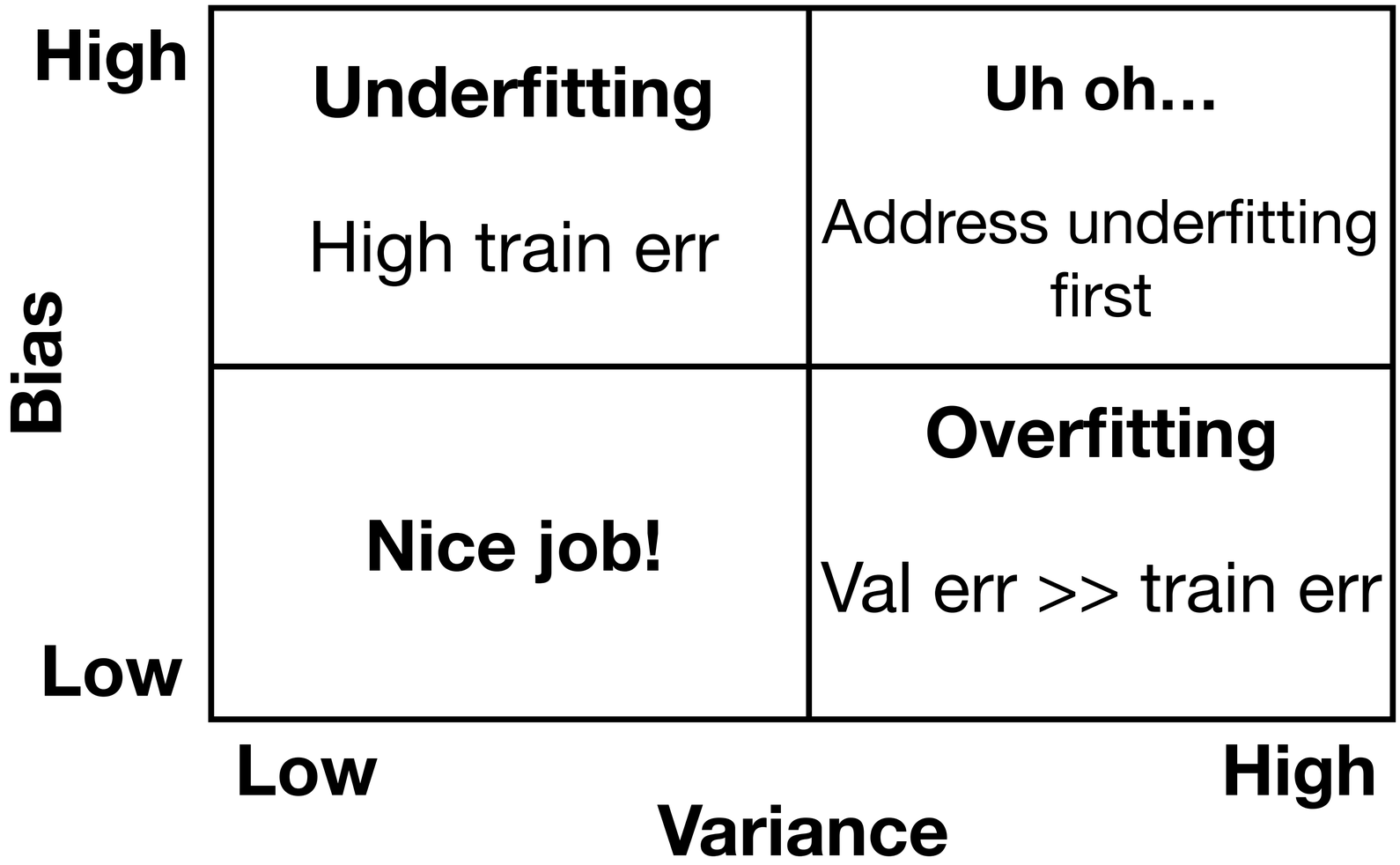
Bias-variance tradeoff

$$\epsilon(\vec{x}; f) = \mathbb{E}_{\vec{x},y,\mathcal{D}}\left[(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2\right]$$

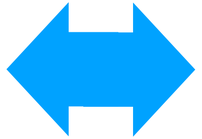
$$= \underbrace{\left(\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(f_\theta^*(\vec{x}; \mathcal{D}))}_{\text{Variance}} + \underbrace{\mathrm{Var}(z)}_{\text{Irreducible error}}$$

Bias-variance cheat sheet

$$\epsilon(\vec{x}; f) = \left(\mathbb{E}\left[f_\theta^*(\vec{x}; \mathcal{D})\right] - \mathbb{E}\left[y\right]\right)^2 + \mathrm{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \mathrm{Var}(z)$$



How to trade off between bias and variance?

Category	Intervention	Bias	Variance
Data	• Add data		
Features	• Add good features		
	• Add bad features		
	• Remove features		
Regularization	• Increase regularization		
	• Decrease regularization		

Outline for today

- Bias-variance decomposition
- **Review of multivariate gaussians**

Breaking down the iid assumptions

Recall:

$$y_i = f(\vec{x}_i) + z_i \qquad\qquad z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Where does iid fail?

- Time series
- “Sliding windows”
- Etc

Multivariate gaussian distribution

Definition

A joint distribution over $\{x_1, \cdots, x_k\}$ with the following pdf:

$$p(x_1, \cdots, x_k) =$$

Multivariate gaussian distribution

Definition

A joint distribution over $\{x_1, \cdots, x_k\}$ with the following pdf:

$$p(x_1, \cdots, x_k) = \frac{\exp(\cdot)}{(\cdot)}$$

Multivariate gaussian distribution

Definition

A joint distribution over $\{x_1, \cdots, x_k\}$ with the following pdf:

$$p(x_1, \cdots, x_k) = \frac{\exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x}. - \vec{\mu})\right)}{(\cdot)}$$

Multivariate gaussian distribution

Definition

A joint distribution over $\{x_1, \cdots, x_k\}$ with the following pdf:

$$p(x_1, \cdots, x_k) = \frac{\exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x}. - \vec{\mu})\right)}{\sqrt{(2\pi)^k |\Sigma|}}$$

Equivalent definition

Definition

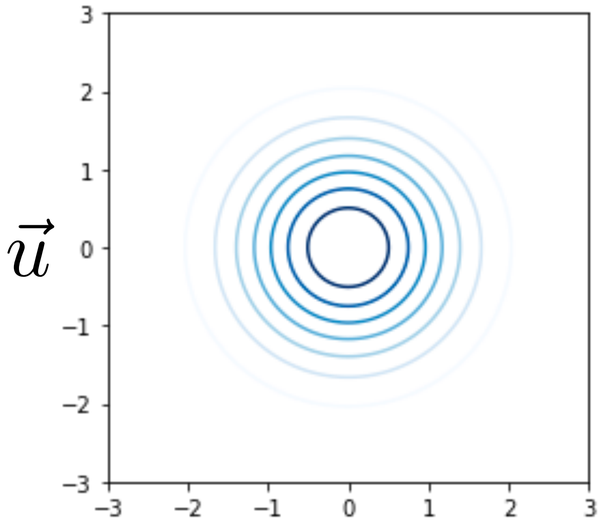
Given $\vec{u} = (u_1, \cdots, u_k),\ u_i \overset{iid}{\sim} \mathcal{N}(0,1)$

$\vec{z} = (z_1, \cdots, z_k)$ is M.V.G. iff

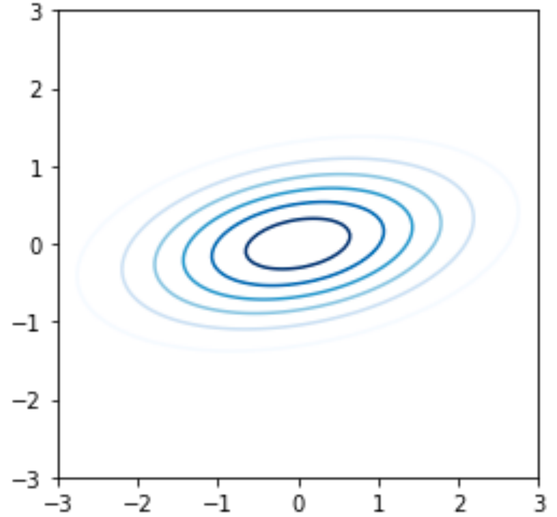
there's a matrix R and vector mu such that

$$\vec{z} = R\vec{u} + \mu$$

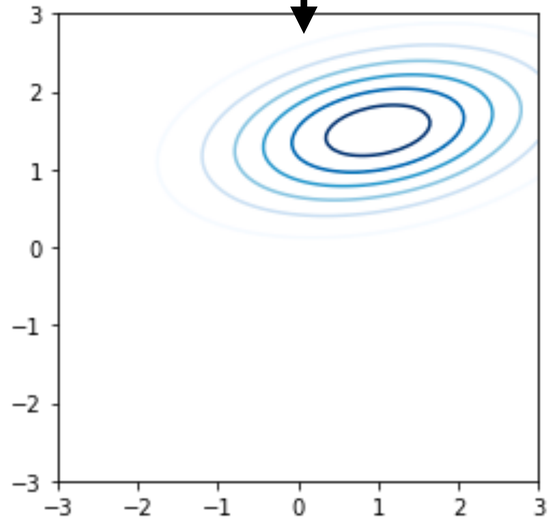
Equivalent definition intuition



$\vec{u}$



$R\vec{u}$



$\vec{z} = R\vec{u} + \mu$

Estimating MVG from data

Maximum likelihood estimate

$$\hat{\mu}_{\mathrm{MLE}}, \hat{\Sigma}_{\mathrm{MLE}} = \arg\max_{\mu,\Sigma} p(\mathcal{D}; \mu, \Sigma)$$

$$= \arg\max_{\mu,\Sigma} \sum_{i=1}^{n} \log p(x_i; \mu, \Sigma)$$

$$= \arg\min_{\mu,\Sigma} \sum_{i=1}^{n} \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) - \frac{n}{2}\log|\Sigma| + C$$

Estimating MVG from data

Maximum likelihood estimate

$$\nabla_\mu \Big(\sum_{i=1}^{n} \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \; -\frac{n}{2}\log|\Sigma| + C \Big)$$

$$= \sum_{i=1}^{n} \nabla_\mu \left(\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \right)$$

$$= \sum_{i=1}^{n} \Sigma^{-1}(x_i-\mu) = 0$$

$$\implies \sum_{i=1}^{n} x_i \; = \sum_{i=1}^{n} \mu \implies \mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Estimating MVG from data

Maximum likelihood estimate

$$\hat{\mu}_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\hat{\Sigma}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{\mathrm{MLE}})(x_i - \hat{\mu}_{\mathrm{MLE}})^T$$

Properties of MVG

Matrix transformations

$$z \sim \mathcal{N}(A\mu_z, Z\Sigma_z A^T)$$

proof:

$$\mu_{Az} = \mathbb{E}\left[AZ\right] = A\mathbb{E}\left[Z\right] = A\mu_z$$

$$\begin{aligned}
\Sigma_{Az} &= \mathbb{E}\left[(Az - \mathbb{E}[Az])(Az - \mathbb{E}[Az])^T\right] \\
&= \mathbb{E}\left[A(z - \mathbb{E}[z])(z - \mathbb{E}[z])^T A^T\right] \\
&= A\mathbb{E}\left[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^T\right] A^T \\
&= A\Sigma_z A^T
\end{aligned}$$

Properties of MVG

Marginal distribution

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

$$\Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx})$$

Properties of MVG

Conditional distribution

$$x \mid y$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx}$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

Prior mean

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

‘error’ in y

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \Sigma_{xy}\underline{\Sigma_{yy}^{-1}(y - \mu_y)}, \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

‘corrected error’ in y

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Properties of MVG

Conditional distribution

$$x \mid y \sim \mathcal{N}(\mu_x + \underline{\Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)}, \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

Implication for x of error in y

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

Where to learn more about MVGs

- Previous semester lecture notes
- More rigorous proofs: Michael Jordan's WIP textbook
  http://people.eecs.berkeley.edu/~jordan/prelims/chapter13.pdf