# CS189: Intro to Machine Learning
# Summer 2018

Lecture 6: Maximum Likelihood and Maximum a Posteriori, pt. 2

Josh Tobin
UC Berkeley EECS

# Goal for today

**Generic MAP regression**

- multivariate gaussian observation noise
- multivariate gaussian prior on parameters

# Approach we'll take

**Generalize our current MAP estimate a bit at a time**

- *Weighted least squares:* Remove the "identically distributed" assumption from the output noise z (where y = f(x) + z)
- *Generalized least squares:* Remove the "independent" assumption from z
- *MAP with colored noise*: Remove i.i.d assumptions from parameter prior

# Outline

- Weighted least squares
- Generalized least squares
- MAP with colored noise

# Outline

- **Weighted least squares**
- Generalized least squares
- MAP with colored noise

# Revisiting the iid assumptions

**OLS assumptions**

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

Equivalent to the following:

$$\vec{z} \sim \mathcal{N}(\vec{0}, \sigma^2 I)$$

# Revisiting the iid assumptions

## OLS assumptions

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

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i$$

$$\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

i.e.,

$$z_i \sim \mathcal{N}(0, \sigma_i^2)$$

# Revisiting the iid assumptions

## OLS assumptions

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

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

i.e.,

Independent, but not identically distributed $\longrightarrow$ $z_i \sim \mathcal{N}(0, \sigma_i^2)$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Key idea

Find a transformation that turns this problem into OLS.

The *reparameteriziation trick.*

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Perform the following transformation

$$\frac{y_i}{\sigma_i} = \frac{f(\vec{x}_i)}{\sigma_i} + \frac{z_i}{\sigma_i}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Perform the following transformation

$$\frac{y_i}{\sigma_i} = \frac{f(\vec{x}_i)}{\sigma_i} + \frac{z_i}{\sigma_i}$$

$$\frac{z_i}{\sigma_i}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Perform the following transformation

$$\frac{y_i}{\sigma_i} = \frac{f(\vec{x}_i)}{\sigma_i} + \frac{z_i}{\sigma_i}$$

$$\frac{z_i}{\sigma_i} \overset{iid}{\sim}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Perform the following transformation

$$\frac{y_i}{\sigma_i} = \frac{f(\vec{x}_i)}{\sigma_i} + \frac{z_i}{\sigma_i}$$

$$\frac{z_i}{\sigma_i} \overset{iid}{\sim} \mathcal{N}(0, 1)$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Write the change of coordinates as follows:

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Write the change of coordinates as follows:

$$\Sigma_z^{-\frac{1}{2}} y$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Write the change of coordinates as follows:

$$\Sigma_z^{-\frac{1}{2}} y \sim$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Write the change of coordinates as follows:

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}($$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Write the change of coordinates as follows:

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Derivation of WLS

Write the change of coordinates as follows:

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS} = \arg\min_w$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS} = \arg\min_w \left( \sum_{i=1}^{n} \frac{(\frac{y_i}{\sigma_i} - \frac{\vec{x}_i^T}{\sigma_i} w)^2}{2} \right)$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS} = \arg\min_w \left( \sum_{i=1}^{n} \frac{(\frac{y_i}{\sigma_i} - \frac{\vec{x}_i^T}{\sigma_i} w)^2}{2} \right) + n \log \sqrt{2\pi}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS} = \arg\min_w \left( \sum_{i=1}^{n} \frac{(\frac{y_i}{\sigma_i} - \frac{\vec{x}_i^T}{\sigma_i} w)^2}{2} \right) + n \log \sqrt{2\pi}$$

$$= \arg\min_w$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS} = \arg\min_w \left( \sum_{i=1}^{n} \frac{(\frac{y_i}{\sigma_i} - \frac{\vec{x}_i^T}{\sigma_i} w)^2}{2} \right) + n \log \sqrt{2\pi}$$

$$= \arg\min_w \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

# Deriving the WLS optimization problem

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma_z^{-\frac{1}{2}} y \sim \mathcal{N}(\Sigma_z^{-\frac{1}{2}} X w, I)$$

## Derivation of WLS

$$\hat{w}_{WLS} = \arg\min_w \left( \sum_{i=1}^{n} \frac{(\frac{y_i}{\sigma_i} - \frac{\vec{x}_i^T}{\sigma_i} w)^2}{2} \right) + n \log \sqrt{2\pi}$$

$$= \arg\min_w \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \vec{x}_i^T w)^2$$

# WLS intuition

$$\hat{w}_{WLS} = \arg\min_{w} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \vec{x}_i^T w)^2$$

- Same as OLS, except the loss for each data point is scaled by some factor
- Probabilistically, this factor is the inverse of the variance
- Data points assumed to have  more noise should be given less weight

# WLS optimization problem

$$\hat{w}_{WLS} = \arg\min_w \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - \vec{x}_i^T w)^2$$

$$= \arg\min_w (y - Xw)^T \Omega (y - Xw) \qquad \left[\Omega = \mathrm{diag}\left(\frac{1}{\sigma_1^2}, \cdots, \frac{1}{\sigma_n^2}\right)\right]$$

$$= \arg\min_w (y - Xw)^T \Omega^{1/2} \Omega^{1/2} (y - Xw)$$

$$= \arg\min_w (\Omega^{1/2} y - \Omega^{1/2} Xw)^T (\Omega^{1/2} y - \Omega^{1/2} Xw)$$

$$= \left((\Omega^{1/2} X)^T (\Omega^{1/2} X)\right)^{-1} \left(\Omega^{1/2} X\right)^T \Omega^{1/2} y$$

$$= \left(X^T \Omega X\right)^{-1} X^T \Omega y$$

# WLS summary

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

## Optimization problem

$$\hat{w}_{WLS} = \arg\min_w \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (y_i - \vec{x}_i^T w)^2 = \arg\min_w (y - Xw)^T \Omega (y - Xw)$$

## Solution

$$\hat{w}_{WLS} = \left(X^T \Omega X\right)^{-1} X^T \Omega y$$

# Outline

- Weighted least squares
- **Generalized least squares**
- MAP with colored noise

# Revisiting the iid assumptions

## OLS assumptions

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

## WLS assumptions

$$y_i = f(\vec{x}_i) + z_i$$

$$\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\Sigma_z = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

i.e.,

$$z_i \sim \mathcal{N}(0, \sigma_i^2)$$

# Revisiting the iid assumptions

## OLS assumptions

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

## GLS assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

Neither independent nor identically distributed!

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Key idea

Find a transformation that turns this problem into OLS.

The *reparameteriziation trick.*

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$\hat{w}_{\mathrm{GLS}}$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\mathrm{GLS}} = \arg\max_{w}$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\mathrm{GLS}} = \arg\max_{w} \frac{1}{\sqrt{\det(\Sigma_z)}}$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\mathrm{GLS}} = \arg\max_{w} \frac{1}{\sqrt{\det(\Sigma_z)}} \frac{1}{\sqrt{(2\pi)^n}}$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\mathrm{GLS}} = \arg\max_{w} \frac{1}{\sqrt{\det(\Sigma_z)}} \frac{1}{\sqrt{(2\pi)^n}} \exp\left( -\frac{1}{2}(Y - Xw)^T \Sigma_z^{-1} (y - Xw) \right)$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\text{GLS}} = \arg\max_w \frac{1}{\sqrt{\det(\Sigma_z)}} \frac{1}{\sqrt{(2\pi)^n}} \exp\left( -\frac{1}{2}(Y - Xw)^T \Sigma_z^{-1} (y - Xw) \right)$$

$$= \arg\min_w$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\text{GLS}} = \arg\max_w \frac{1}{\sqrt{\det(\Sigma_z)}} \frac{1}{\sqrt{(2\pi)^n}} \exp\left( -\frac{1}{2}(Y - Xw)^T \Sigma_z^{-1} (y - Xw) \right)$$

$$= \arg\min_w (y - Xw)^T \Sigma_z^{-1} (y - Xw)$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$\hat{w}_{\text{GLS}} = \arg\min_w (y - Xw)^T \Sigma_z^{-1} (y - Xw)$$

Spectral decomposition of Sigma

$$\Sigma_z = Q \ \text{diag}(\sigma_1^2, \cdots \sigma_n^2) Q^T$$

$$\Sigma_z^{-\frac{1}{2}} = Q \ \text{diag}\left(\frac{1}{\sigma_1}, \cdots \frac{1}{\sigma_n}\right) Q^T$$

# Deriving the GLS optimization problem

## GLS assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Derivation of GLS

$$
\begin{aligned}
\hat{w}_{\mathrm{GLS}} &= \arg\min_w (y - Xw)^T \Sigma_z^{-1} (y - Xw) \\
&= \arg\min_w (y - Xw)^T \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (y - XW) \\
&= \arg\min_w (\Sigma^{-\frac{1}{2}} y - \Sigma^{-\frac{1}{2}} Xw)^T (\Sigma^{-\frac{1}{2}} y - \Sigma^{-\frac{1}{2}} XW) \\
&= \left( (\Sigma^{-\frac{1}{2}} X)^T (\Sigma^{-\frac{1}{2}} X) \right)^{-1} (\Sigma^{-\frac{1}{2}} X)^T \Sigma^{-\frac{1}{2}} y \\
&= \left( X^T \Sigma_z^{-1} X \right)^{-1} X^T \Sigma_z^{-1} y
\end{aligned}
$$

# GLS summary

## Probabilistic assumptions

$y_i = f(\vec{x}_i) + z_i$ $\quad$ $\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$ $\quad$ $\Sigma_z =$ Any covariance matrix

## Optimization problem

$$\hat{w}_{\mathrm{GLS}} = \arg\min_{w} (y - Xw)^T \Sigma_z^{-1} (y - Xw)$$

## Solution

$$\hat{w}_{\mathrm{GLS}} = \left(X^T \Sigma_z^{-1} X\right)^{-1} X^T \Sigma_z^{-1} y$$

# Outline

- Weighted least squares
- Generalized least squares
- **MAP with colored noise**

# Review of MAP

$f_\theta$ = *Family* of models, e.g., $f_\theta(x) = \theta^T x$

$$\hat{\theta} = \text{argmax}_\theta \mathcal{L}(\theta; \mathcal{D})$$

**MLE:** $\mathcal{L}(\theta; \mathcal{D}) = p(\text{data} = \mathcal{D} \mid \text{model} = f_\theta)$

**MAP:** $\mathcal{L}(\theta; \mathcal{D}) = p(\text{model} = f_\theta \mid \text{data} = \mathcal{D})$

# Review of MAP

$$\hat{\theta}_{\text{MLE}} = \arg\min_\theta - \log p(\text{data} = \mathcal{D} \mid \text{model} = f_\theta)$$

$$\hat{\theta}_{\text{MAP}} = \arg\min_\theta \underbrace{- \log p(\text{data} = \mathcal{D} \mid \text{model} = f_\theta)}_{\text{Same as MLE}} \underbrace{- \log p(\text{model} = f_\theta)}_{\textit{Prior}}$$

# Review of MAP

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta}$$

# Review of MAP

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$$

# Review of MAP

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 + \frac{\sigma^2}{\sigma^2_{\text{prior}}}$$

# Review of MAP

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 + \frac{\sigma^2}{\sigma^2_{\text{prior}}} \sum_{j=1}^{d} (\theta^j - \mu^j_{\text{prior}})^2$$

# Review of MAP

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 + \frac{\sigma^2}{\sigma^2_{\text{prior}}} \sum_{j=1}^{d} (\theta^j - \mu^j_{\text{prior}})^2$$

$$\text{assume } \mu^j_{\text{prior}} = 0 \qquad \text{let } \lambda = \frac{\sigma^2}{\sigma^2_{\text{prior}}}$$

$$\arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 + \lambda \sum_{j=1}^{d} (\theta^j)^2$$

**Ridge regression**

# MAP with dependent parameters

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

# MAP with dependent parameters

~~**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$~~

**Assumption:** $\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$

**Can we transform back to the original problem?**

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

## Derivation

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \qquad u \sim \mathcal{N}(0, I)$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

## Derivation

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \qquad u \sim \mathcal{N}(0, I)$$

$$\vec{y} = X\theta + \vec{z}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

## Derivation

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \qquad u \sim \mathcal{N}(0, I)$$

$$\vec{y} = X\theta + \vec{z} \implies \vec{y} = X\left(\Sigma_\theta^{\frac{1}{2}} u + \mu_\theta\right) + \vec{z}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \qquad u \sim \mathcal{N}(0, I)$$

$$\vec{y} = X\theta + \vec{z} \implies \vec{y} = X\left(\Sigma_\theta^{\frac{1}{2}} u + \mu_\theta\right) + \vec{z}$$

$$\implies \vec{y} - X\mu_\theta = X\Sigma_\theta^{\frac{1}{2}} u + \vec{z}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i$$

$$y_i = \vec{x_i}^T \theta + z_i$$

$$\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$\Sigma_z =$ Any covariance matrix

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\underbrace{\vec{y} - X\mu_\theta}_{\tilde{y}} = \underbrace{\left(X\Sigma_\theta^{\frac{1}{2}}\right)}_{\tilde{X}} u + \vec{z}$$

$$\tilde{y} = \tilde{X}u + \vec{z}$$

→ Standard ridge regression with lambda = 1

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i$$

$$y_i = \vec{x_i}^T \theta + z_i$$

$$\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$\Sigma_z =$ Any covariance matrix

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\underbrace{\vec{y} - X\mu_\theta}_{\tilde{y}} = \underbrace{\left(X\Sigma_\theta^{\frac{1}{2}}\right)}_{\tilde{X}} u + \vec{z}$$

$$\tilde{y} = \tilde{X}u + \vec{z}$$

$$\hat{u} = (\tilde{X}^T\tilde{X} + I)^{-1}\tilde{X}^T\tilde{y}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\underbrace{\vec{y} - X\mu_\theta}_{\tilde{y}} = \underbrace{\left(X\Sigma_\theta^{\frac{1}{2}}\right)}_{\tilde{X}} u + \vec{z}$$

$$\hat{u} = (\tilde{X}^T\tilde{X} + I)^{-1}\tilde{X}^T\tilde{y}$$

$$\hat{u}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\underbrace{\vec{y} - X\mu_\theta}_{\tilde{y}} = \underbrace{\left(X\Sigma_\theta^{\frac{1}{2}}\right)}_{\tilde{X}} u + \vec{z}$$

$$\hat{u} = (\tilde{X}^T \tilde{X} + I)^{-1} \tilde{X}^T \tilde{y}$$

$$\hat{u} = \left((\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I\right)^{-1}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\underbrace{\vec{y} - X\mu_\theta}_{\tilde{y}} = \underbrace{\left(X\Sigma_\theta^{\frac{1}{2}}\right)}_{\tilde{X}} u + \vec{z}$$

$$\hat{u} = (\tilde{X}^T \tilde{X} + I)^{-1} \tilde{X}^T \tilde{y}$$

$$\hat{u} = \left((\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I\right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\underbrace{\vec{y} - X\mu_\theta}_{\tilde{y}} = \underbrace{\left(X\Sigma_\theta^{\frac{1}{2}}\right)}_{\tilde{X}} u + \vec{z}$$

$$\hat{u} = (\tilde{X}^T \tilde{X} + I)^{-1} \tilde{X}^T \tilde{y}$$

$$\hat{u} = \left((\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I\right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i$$

$$y_i = \vec{x_i}^T \theta + z_i$$

$$\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$\Sigma_z =$ Any covariance matrix

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{u} = \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

$$\hat{\theta} = \Sigma_\theta^{\frac{1}{2}} \hat{u} + \mu_\theta$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i$$

$$y_i = \vec{x_i}^T \theta + z_i$$

$$\vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z)$$

$$\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$\Sigma_z =$ Any covariance matrix

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{u} = \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

$$\hat{\theta} = \Sigma_\theta^{\frac{1}{2}} \hat{u} + \mu_\theta$$

$$= \Sigma_\theta^{\frac{1}{2}} \left( \right.$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{u} = \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

$$\hat{\theta} = \Sigma_\theta^{\frac{1}{2}} \hat{u} + \mu_\theta$$

$$= \Sigma_\theta^{\frac{1}{2}} \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{u} = \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

$$\hat{\theta} = \Sigma_\theta^{\frac{1}{2}} \hat{u} + \mu_\theta$$

$$= \Sigma_\theta^{\frac{1}{2}} \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta) + \mu_\theta$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{u} = \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta)$$

$$\hat{\theta} = \Sigma_\theta^{\frac{1}{2}} \hat{u} + \mu_\theta$$

$$= \Sigma_\theta^{\frac{1}{2}} \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta) + \mu_\theta$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{\theta} = \cancel{\Sigma_\theta^{\frac{1}{2}}} \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} \cancel{(\Sigma_\theta^{\frac{1}{2}})^T} X^T (y - X\mu_\theta) + \mu_\theta$$

$$\cancel{(\Sigma_\theta^{\frac{1}{2}})^T} \left( X^T X + (\Sigma_\theta^{-\frac{1}{2}})^T \Sigma_\theta^{-\frac{1}{2}} \right) \cancel{\Sigma_\theta^{\frac{1}{2}}}$$

$$\Sigma_\theta^{-1}$$

# MAP with dependent parameters

## Probabilistic assumptions

$$y_i = f(\vec{x}_i) + z_i \qquad \vec{z} \sim \mathcal{N}(\vec{0}, \Sigma_z) \qquad \Sigma_z = \text{Any covariance matrix}$$

$$y_i = \vec{x_i}^T \theta + z_i \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\theta = \Sigma_\theta^{\frac{1}{2}} u + \mu_\theta \quad u \sim \mathcal{N}(0, I)$$

## Derivation

$$\hat{\theta} = \Sigma_\theta^{\frac{1}{2}} \left( (\Sigma_\theta^{\frac{1}{2}})^T X^T X \Sigma_\theta^{\frac{1}{2}} + I \right)^{-1} (\Sigma_\theta^{\frac{1}{2}})^T X^T (y - X\mu_\theta) + \mu_\theta$$

$$= \left( X^T X + \Sigma_\theta^{-1} \right)^{-1} X^T (y - X\mu_\theta) + \mu_\theta$$

# Summary of linear regression

## MLE, iid noise

$$\hat{\theta}_{OLS} = (X^T X)^{-1} X^T y$$

## MLE, MVG noise

$$\hat{\theta}_{GLS} = (X^T \Sigma_z^{-1} X)^{-1} X^T \Sigma_z^{-1} y$$

## MAP, iid noise and parameters

$$\hat{\theta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

## MAP, iid noise, MVG parameters

$$\hat{\theta}_{\text{MAP}} = (X^T X + \Sigma_\theta^{-1})^{-1} X^T (y - X\mu_\theta) + \mu_\theta$$

## MAP, MVG noise, MVG parameters

$$\hat{\theta}_{\text{GMAP}} = (X^T \Sigma_z^{-1} X + \Sigma_\theta^{-1})^{-1} X^T \Sigma_z^{-1} y$$