6.7900: Machine Learning
Lecture 5

Lecture start: Tues/Thurs 2:35pm
Who's speaking today? Prof. Tamara Broderick
Course website: gradml.mit.edu
Questions? Ask here or on piazza.com/mit/fall2024/67900/
Materials: Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

Last Times

I. MLE, ERM, Bayes
II. Linear regression
    A. Motivation
    B. MLE & ERM

Upcoming

I. Challenges with MLE/ERM for linear regression
II. Bayesian linear regression
    A. Bayes with Gaussians
    B. Posterior, predictive

Recap

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$
- Empirical risk minimization: if we just minimize empirical risk over all decision rules, might not generalize well

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$
- Empirical risk minimization: if we just minimize empirical risk over all decision rules, might not generalize well
  - Idea: restrict to linear predictors $h(x) = \theta^\top x$

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*, *Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$
- Empirical risk minimization: if we just minimize empirical risk over all decision rules, might not generalize well
  - Idea: restrict to linear predictors $h(x) = \theta^\top x$
- In both cases (MLE and ERM), we get the following:

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*, *Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$
- Empirical risk minimization: if we just minimize empirical risk over all decision rules, might not generalize well
  - Idea: restrict to linear predictors $h(x) = \theta^\top x$
- In both cases (MLE and ERM), we get the following:

$$\hat{\theta} = \arg\min_\theta \mathrm{RSS}(\theta) = \arg\min_\theta (X\theta - Y)^\top (X\theta - Y)$$

RSS = residual sum of squares

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$
- Empirical risk minimization: if we just minimize empirical risk over all decision rules, might not generalize well
  - Idea: restrict to linear predictors $h(x) = \theta^\top x$
- In both cases (MLE and ERM), we get the following:

$$\hat{\theta} = \arg\min_\theta \mathrm{RSS}(\theta) = \arg\min_\theta (X\theta - Y)^\top (X\theta - Y)$$

  - If *N*>*D* & *X* is full rank, then the minimizer is:

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y$$

RSS = residual sum of squares

Recap

Proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- We don't know the distribution of a future data point, but if we assume past and future data are iid, training data can help
- Maximum likelihood: Model, e.g. $p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$
  - And use $p(y|x, \hat{\theta}, \hat{\sigma}^2) \longrightarrow h(x) = \hat{\theta}^\top x$
- Empirical risk minimization: if we just minimize empirical risk over all decision rules, might not generalize well
  - Idea: restrict to linear predictors $h(x) = \theta^\top x$
- In both cases (MLE and ERM), we get the following:

$$\hat{\theta} = \arg\min_\theta \mathrm{RSS}(\theta) = \arg\min_\theta (X\theta - Y)^\top (X\theta - Y)$$

RSS = residual sum of squares

  - If *N*>*D* & *X* is full rank, then the minimizer is:

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y$$

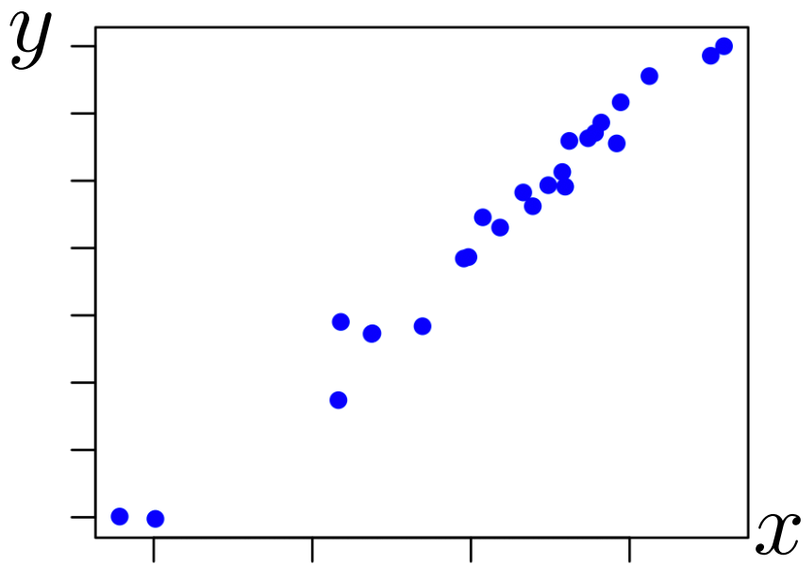
  - What if those conditions don't hold?

Visualizing the full-rank case

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$

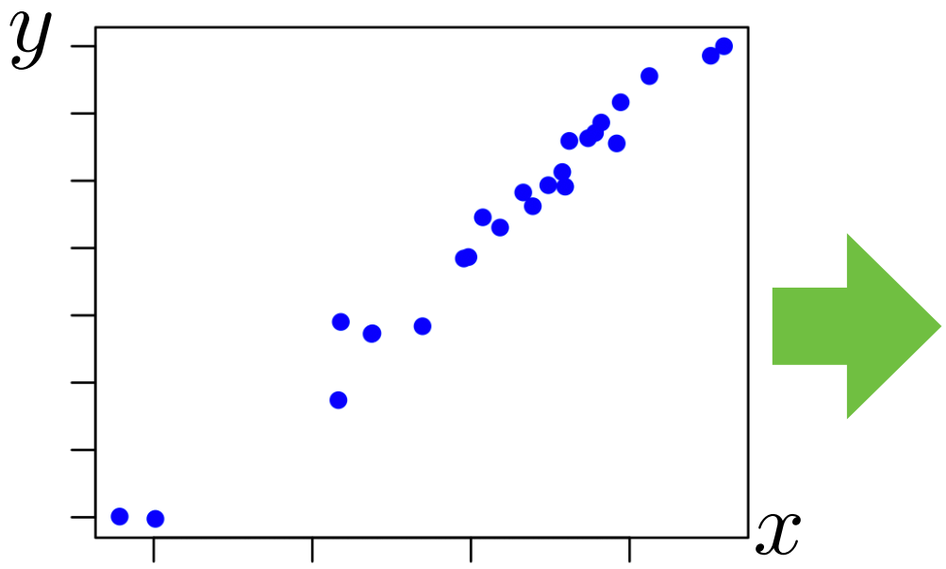
Visualizing the full-rank case

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$



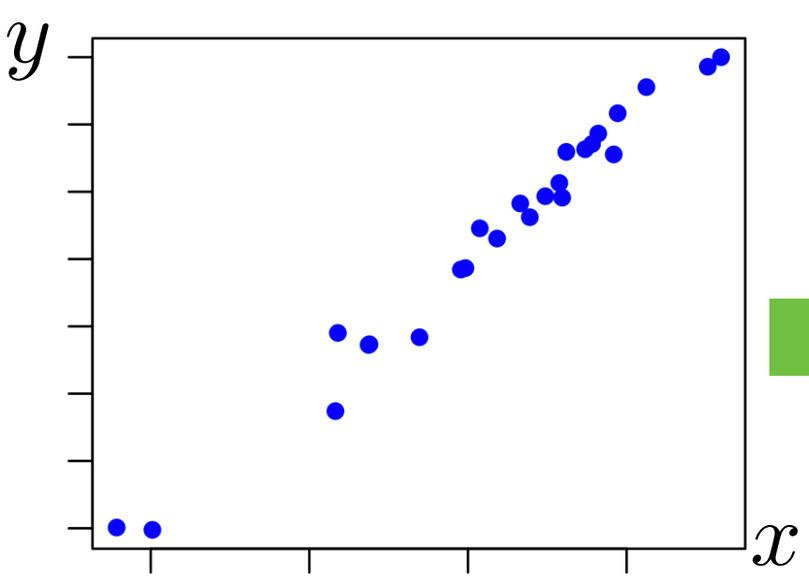
Visualizing the full-rank case

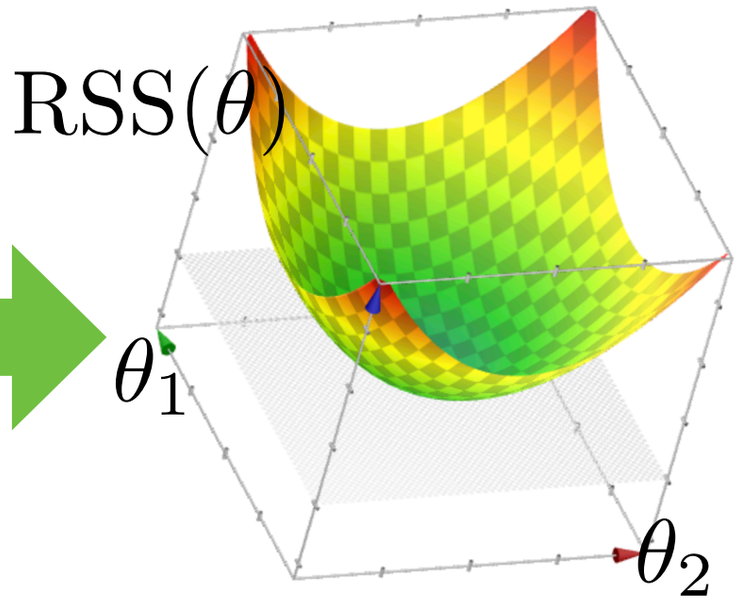
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$



Visualizing the full-rank case

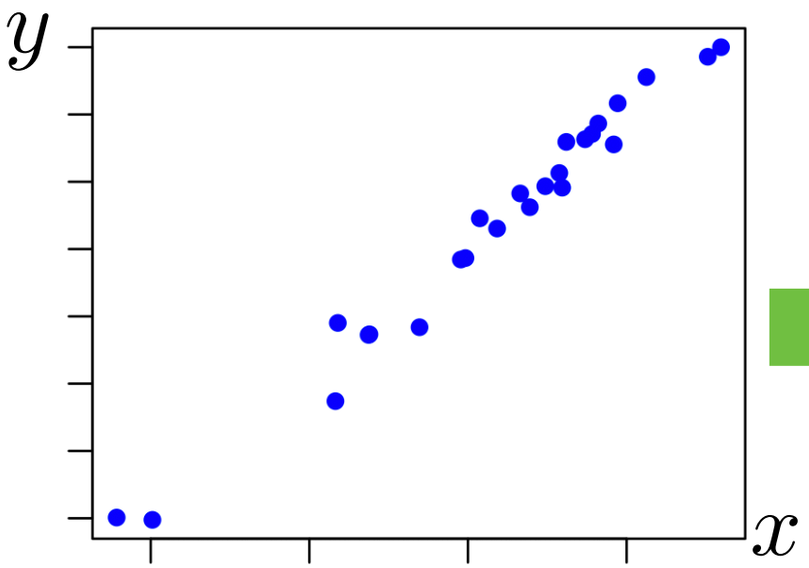
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$



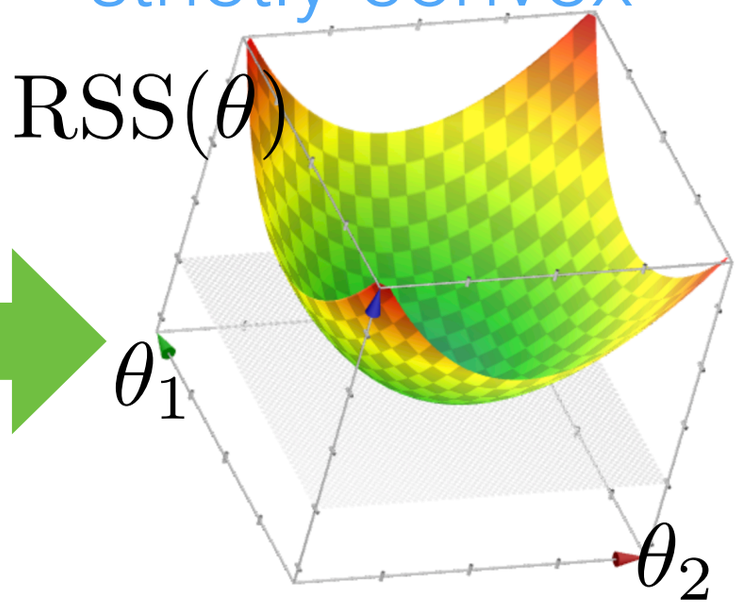


Visualizing the full-rank case

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$

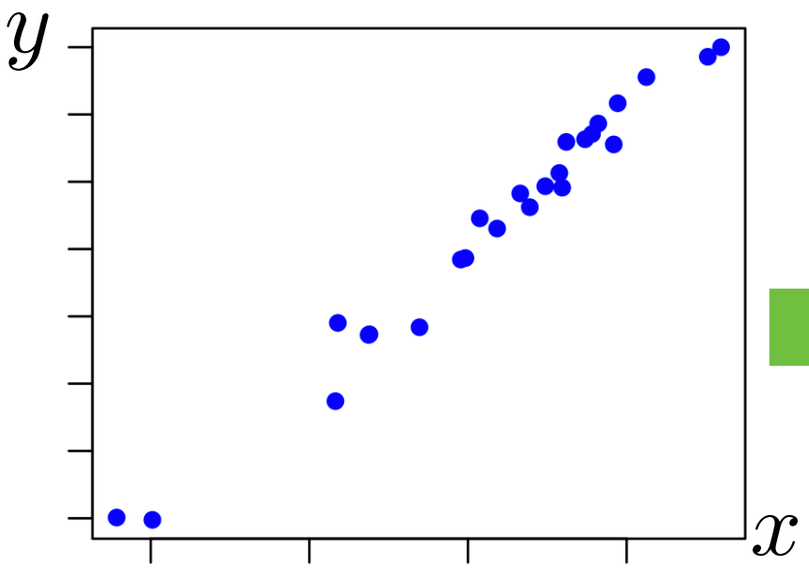


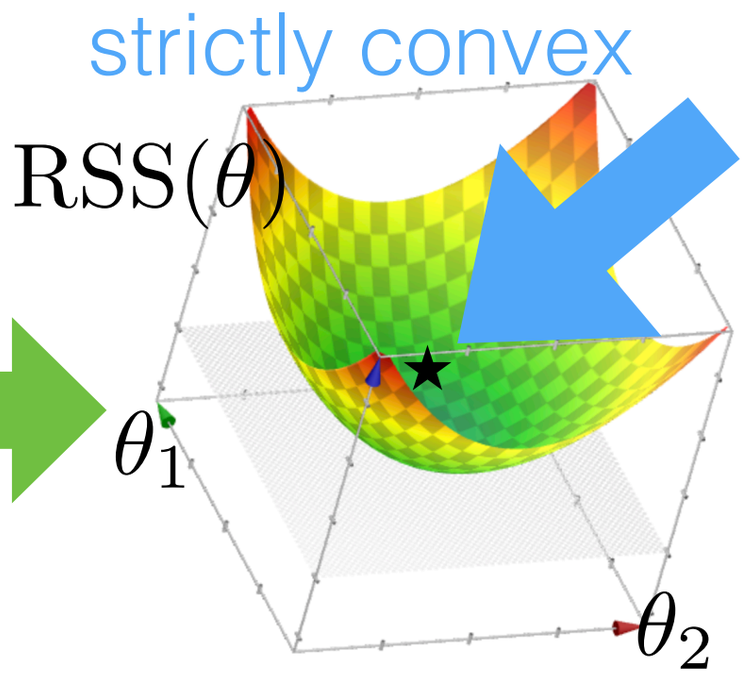
strictly convex



Visualizing the full-rank case

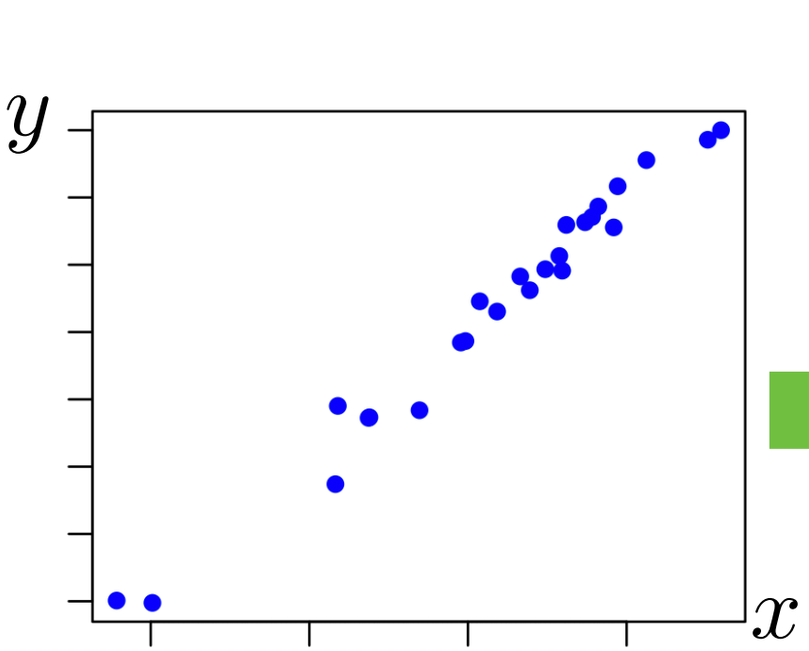
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$

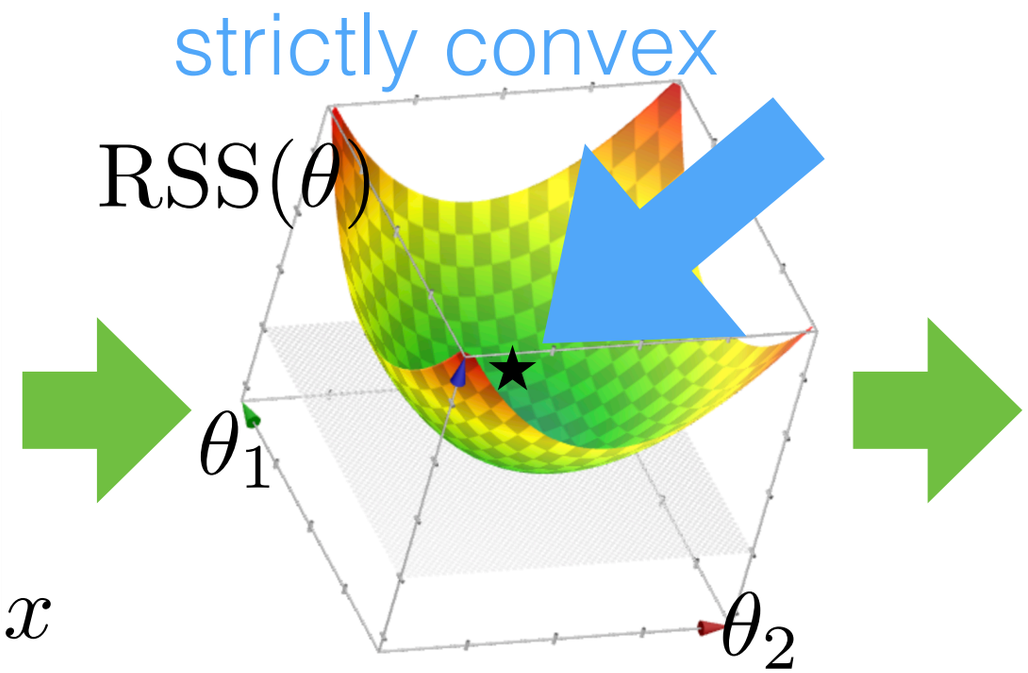




Visualizing the full-rank case

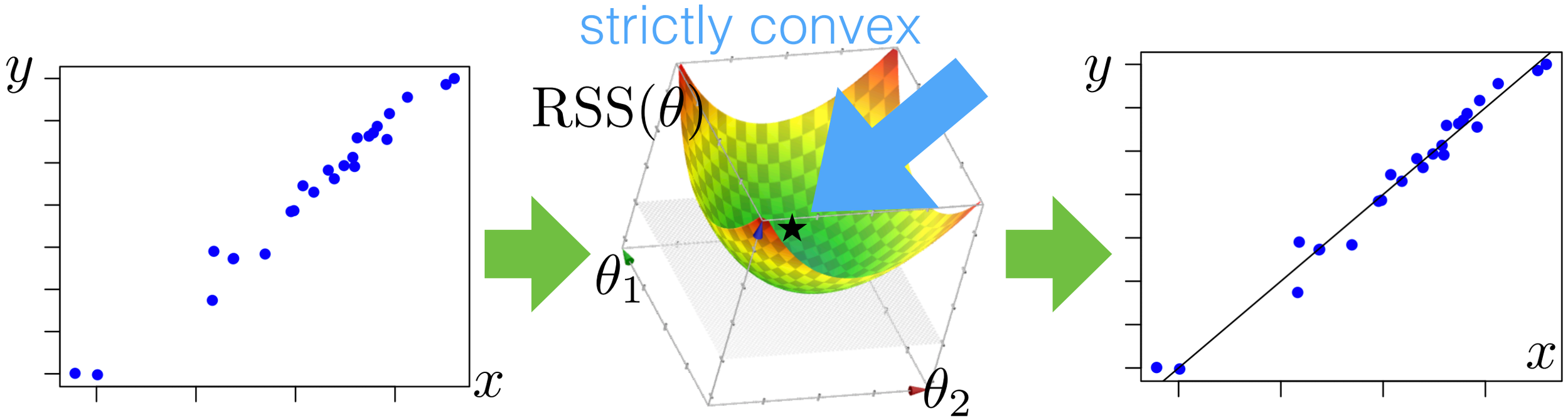
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$





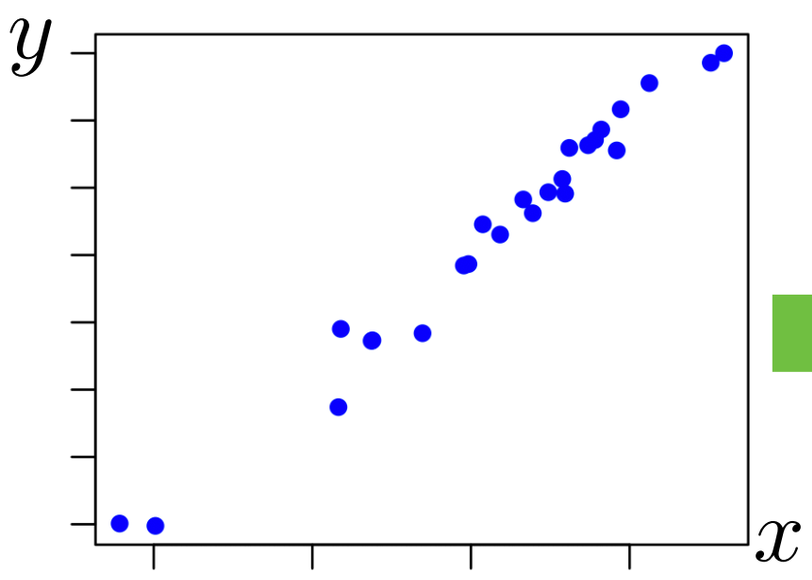
Visualizing the full-rank case

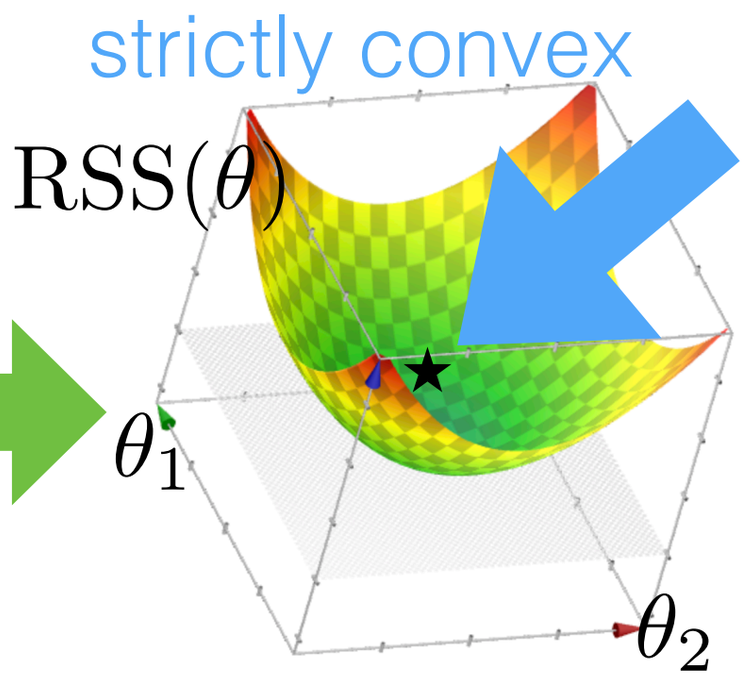
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$

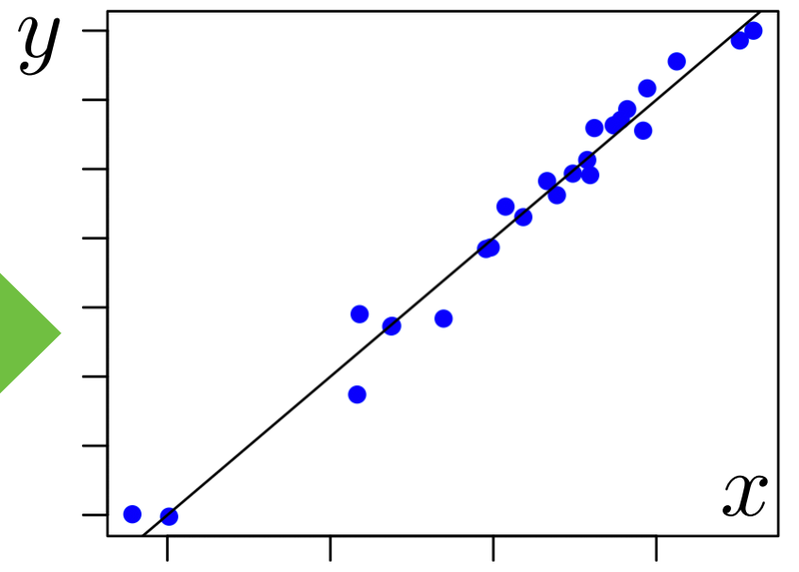


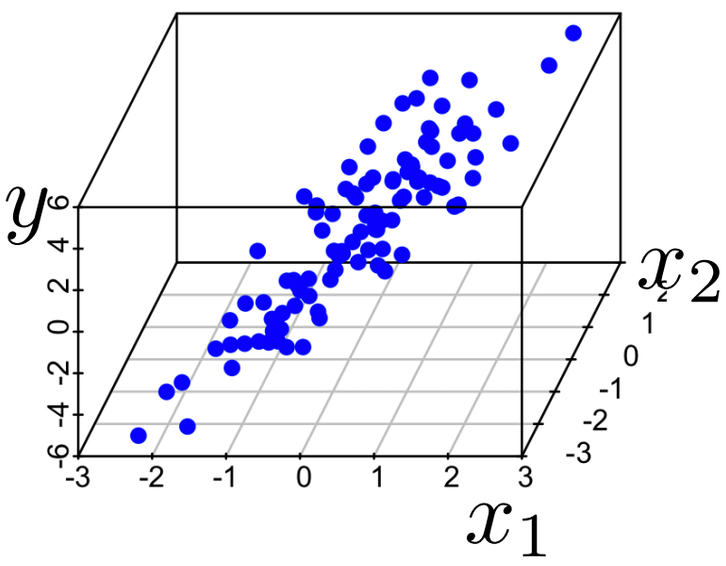
Visualizing the full-rank case

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$



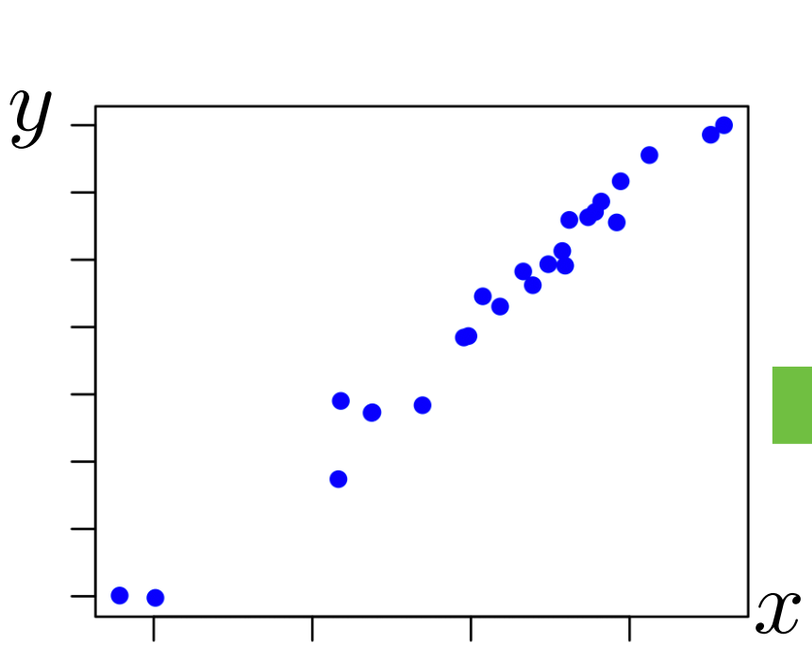


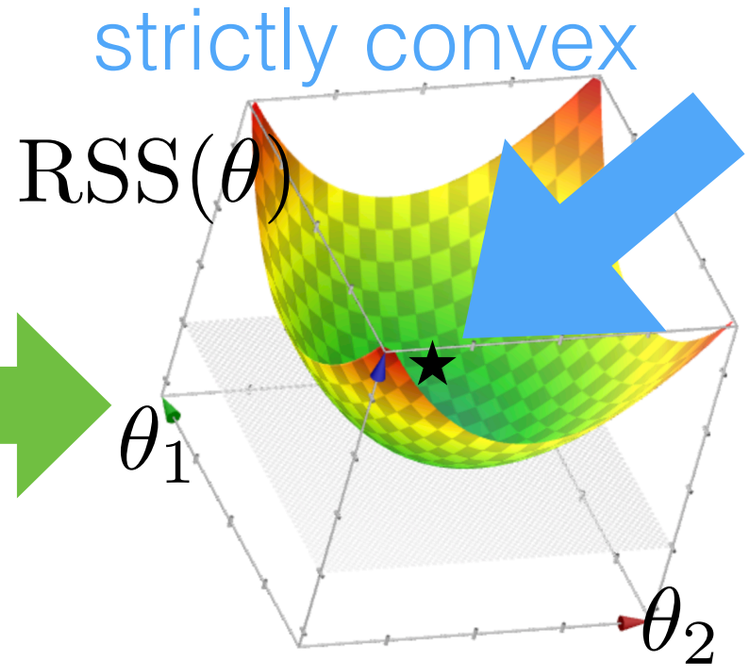


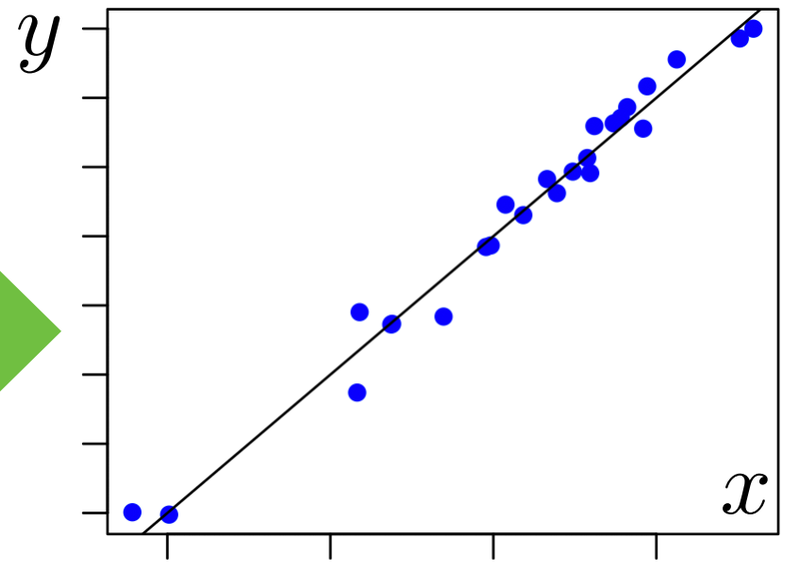


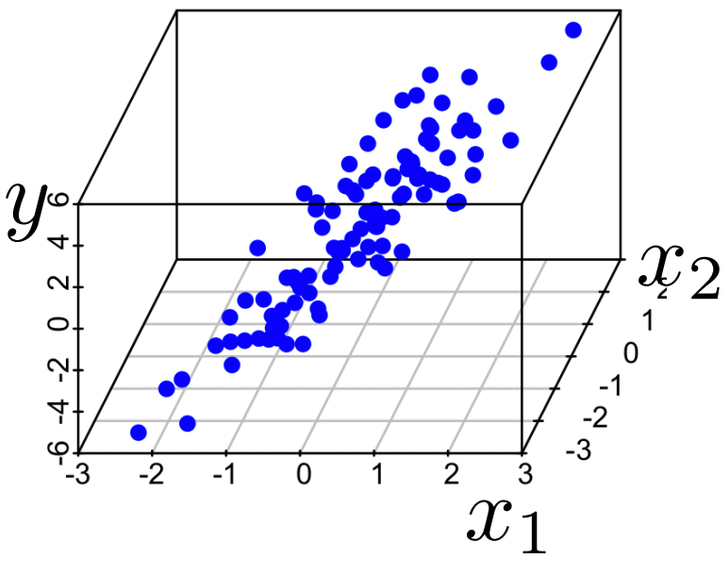
Visualizing the full-rank case

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$





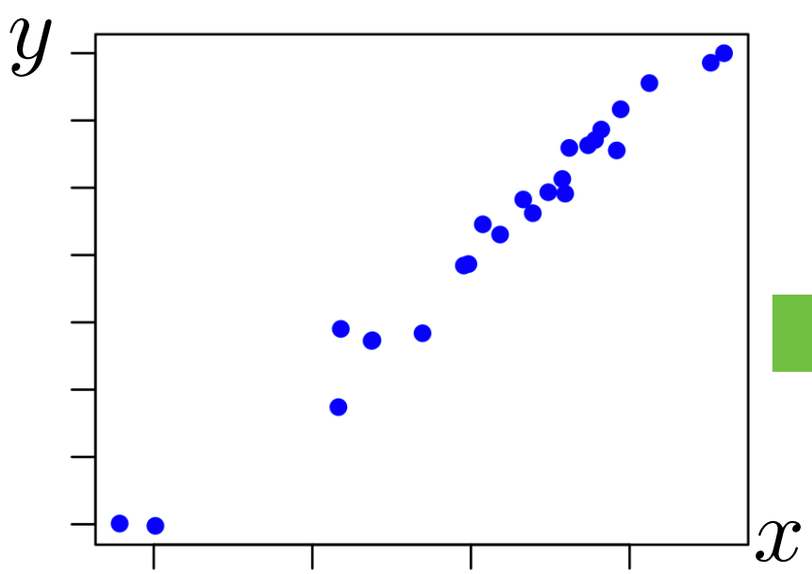


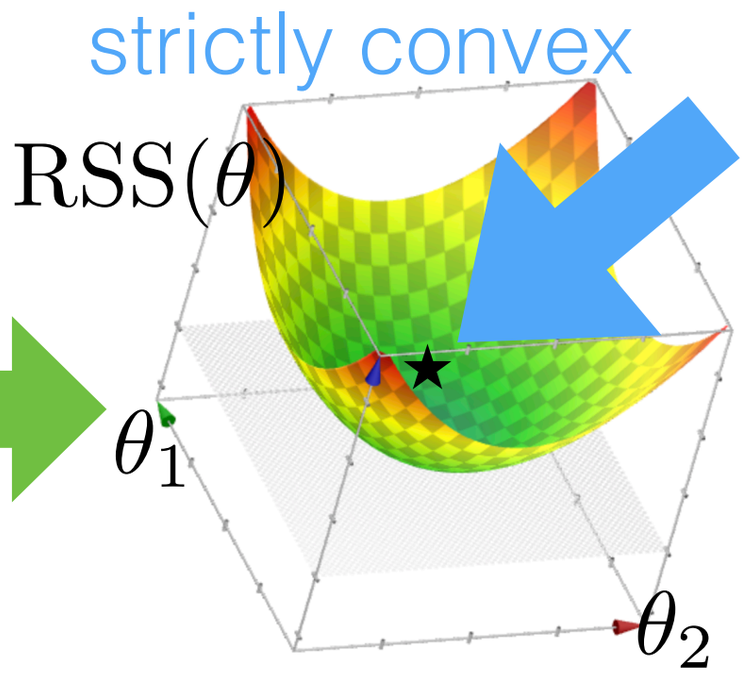


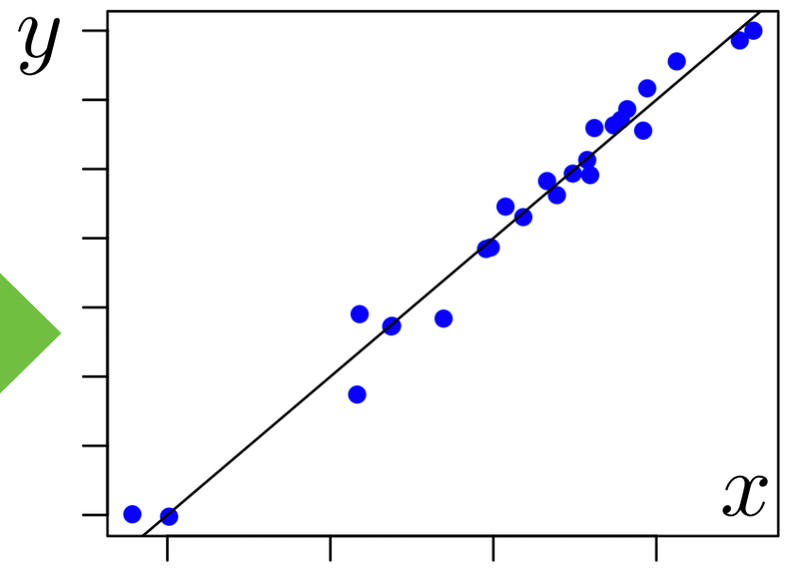
Visualizing the full-rank case

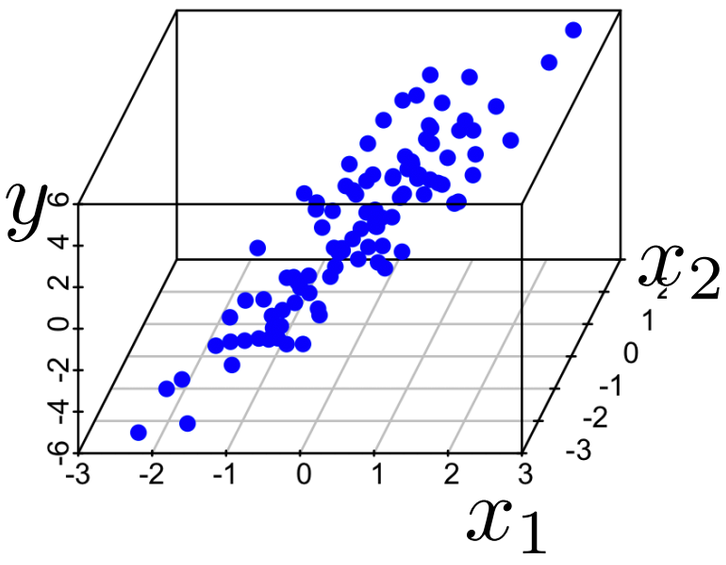
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$

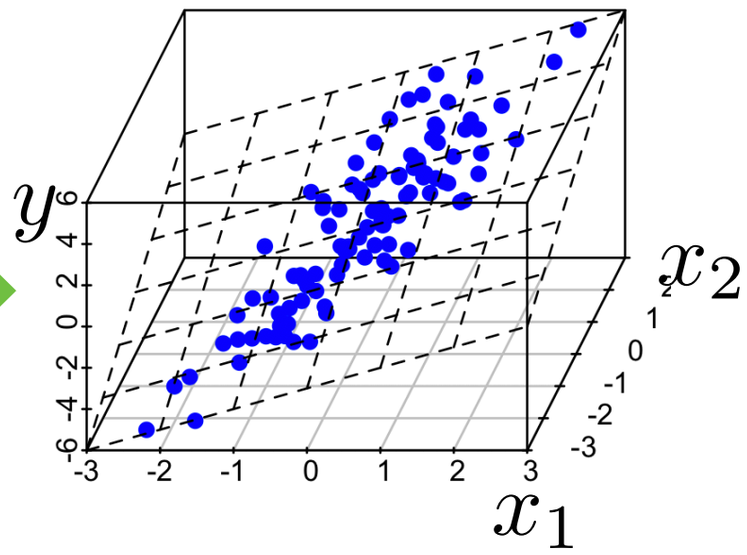
strictly convex





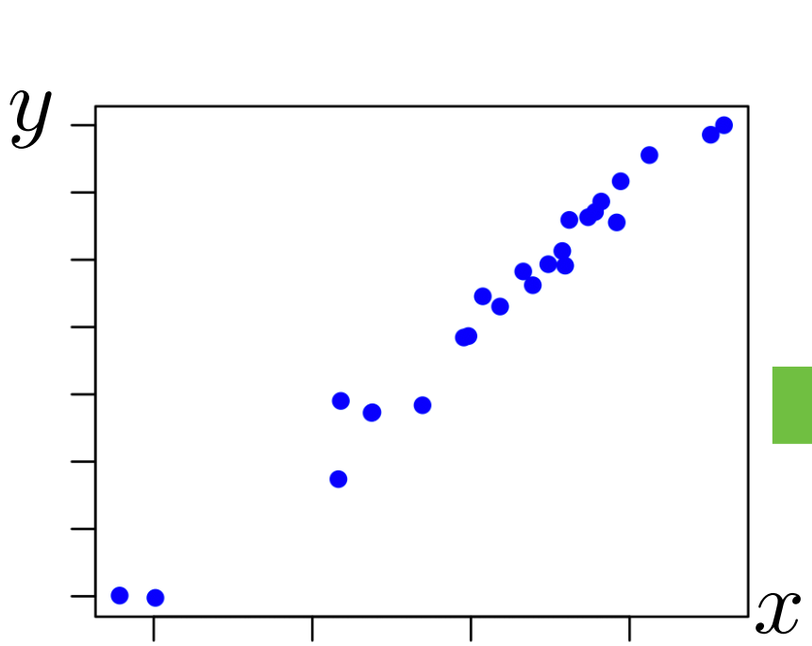


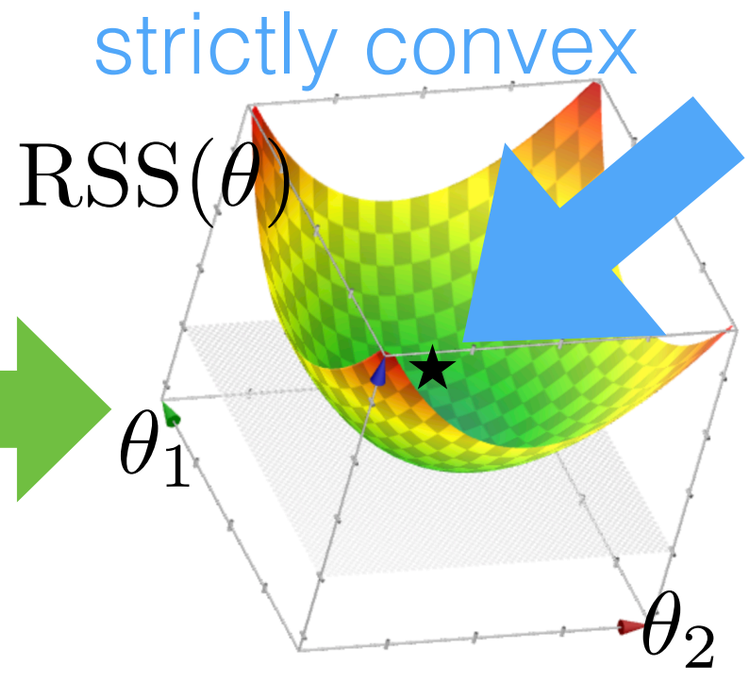


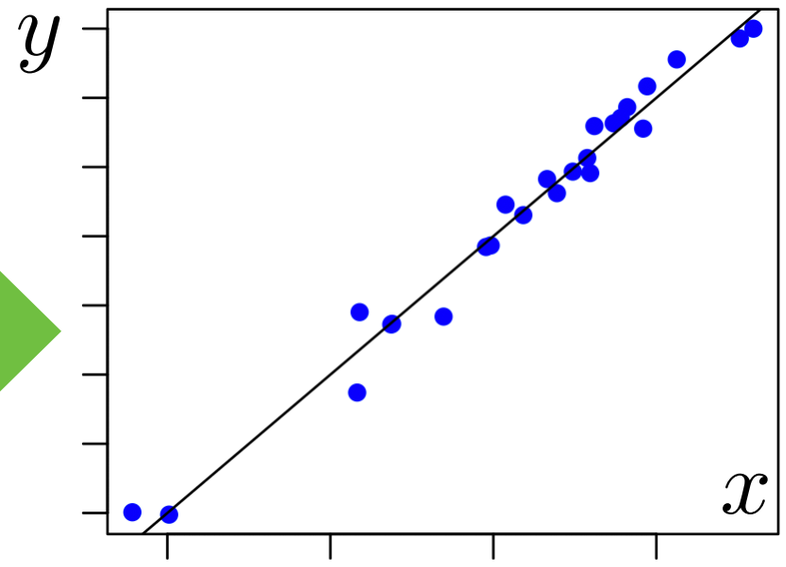


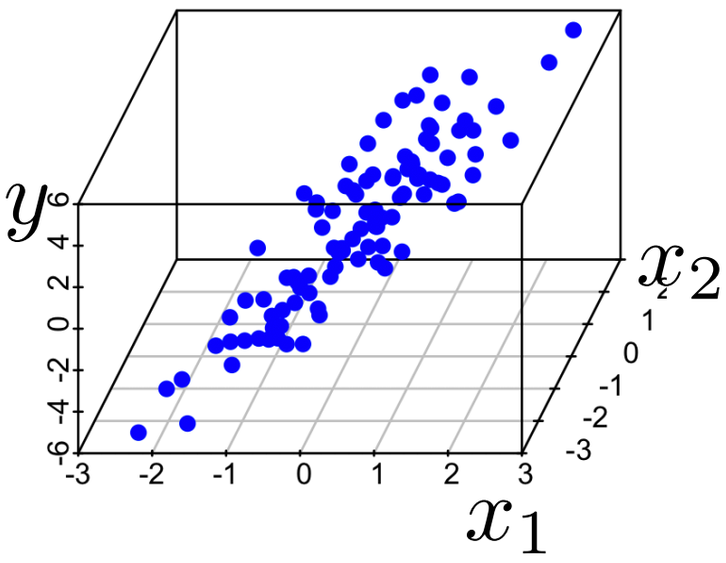
Visualizing the full-rank case

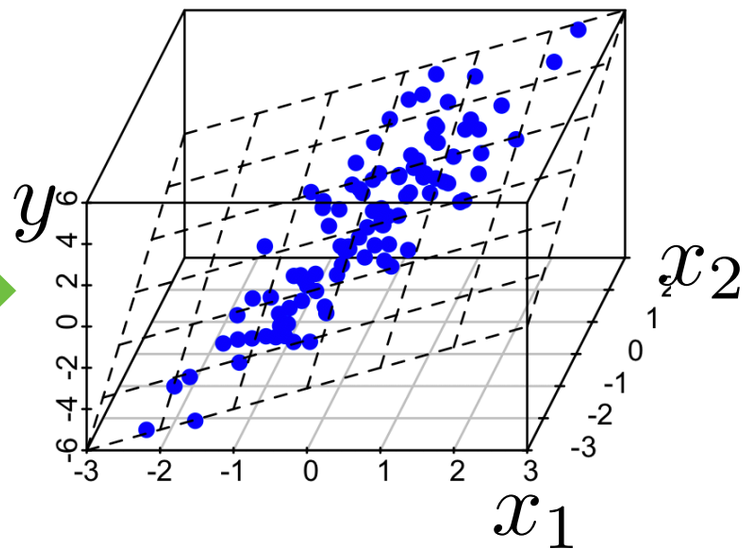
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$









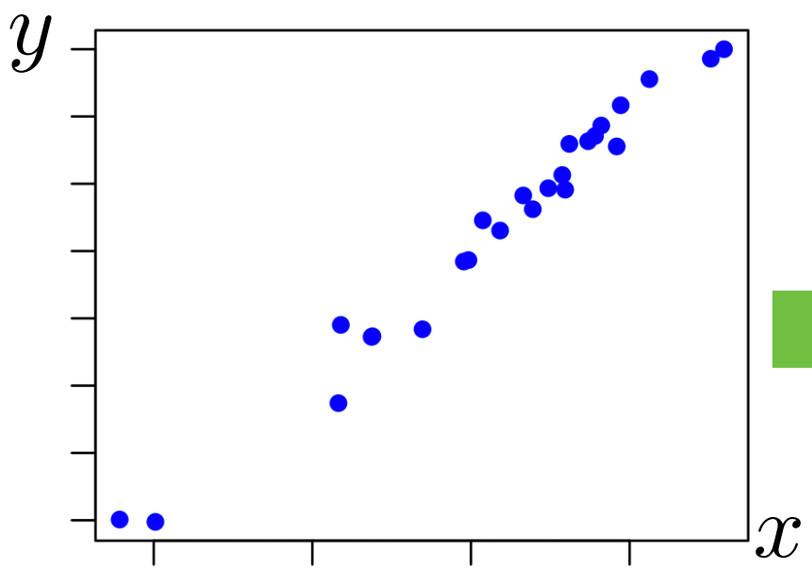


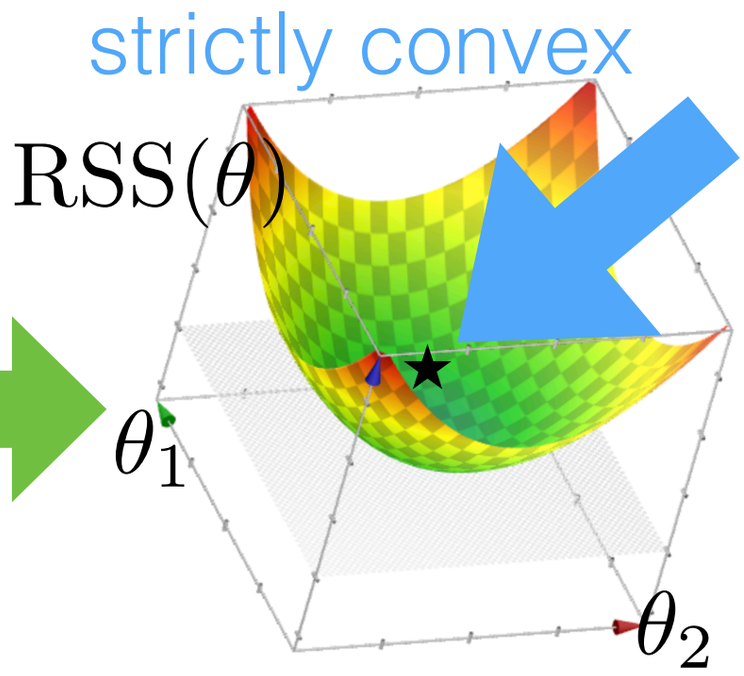
- What could go wrong?

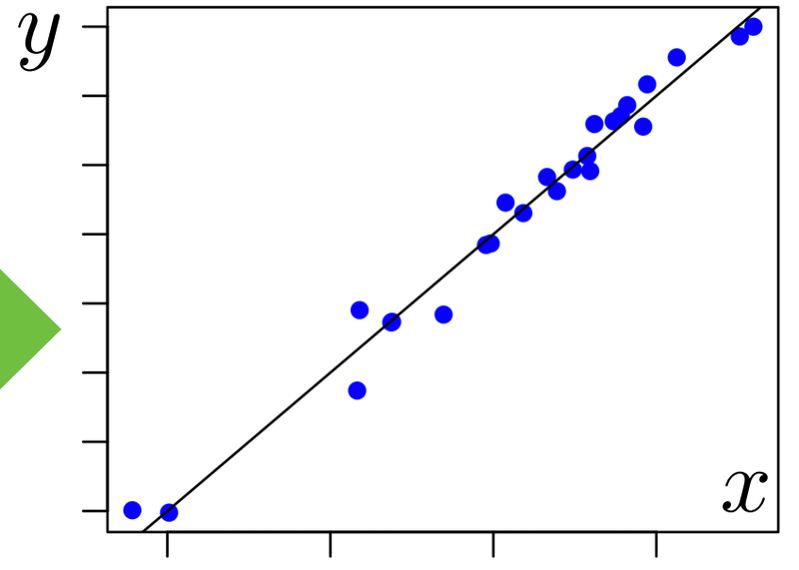
Visualizing the full-rank case

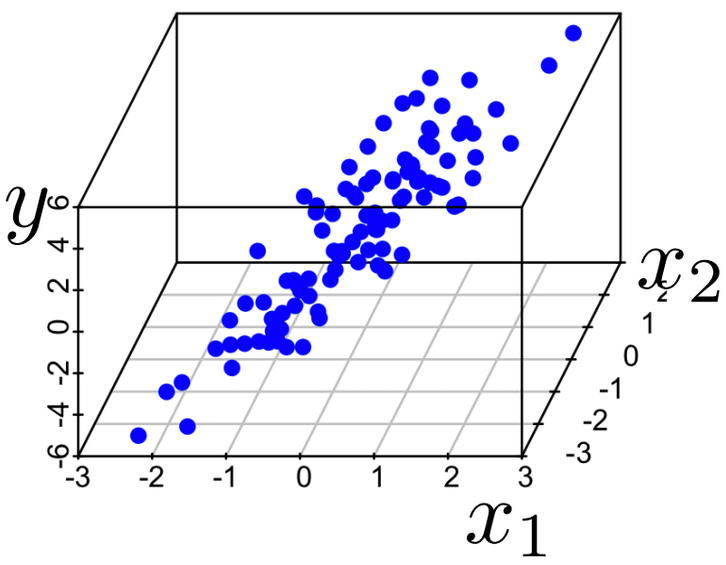
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - When *N*>*D* and *X* is full rank: $\hat{\theta} = (X^\top X)^{-1} X^\top Y$

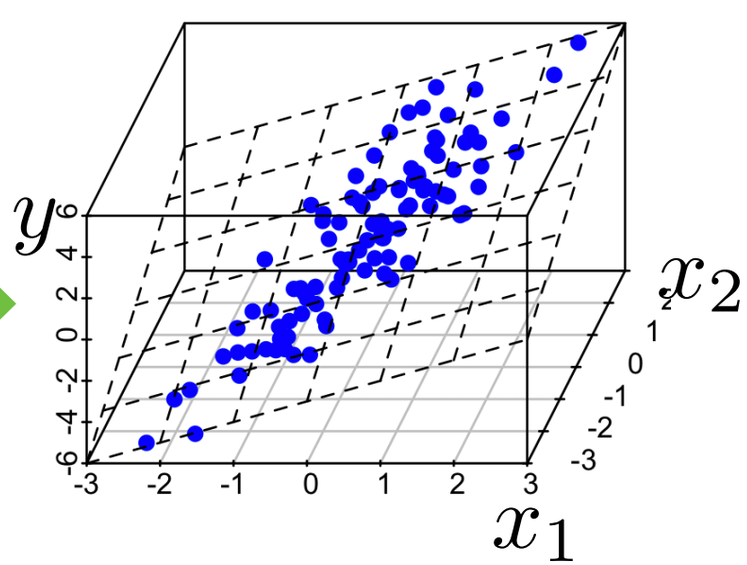
strictly convex



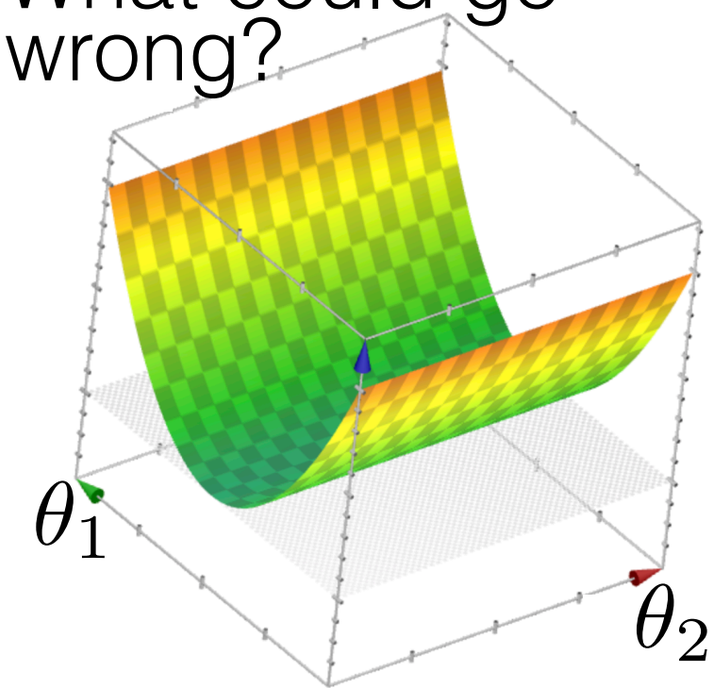








- What could go wrong?



What could go wrong?

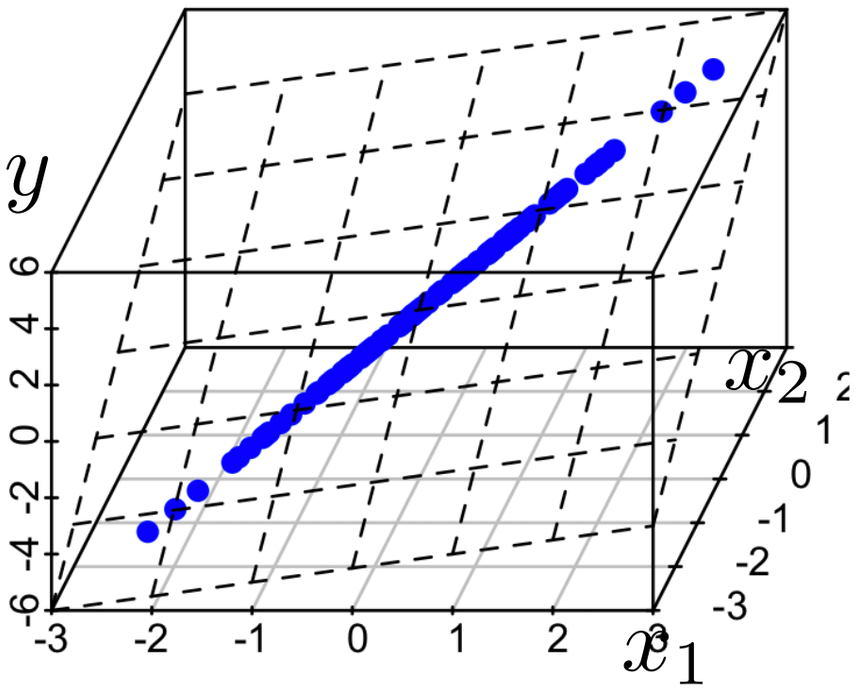
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$

What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)

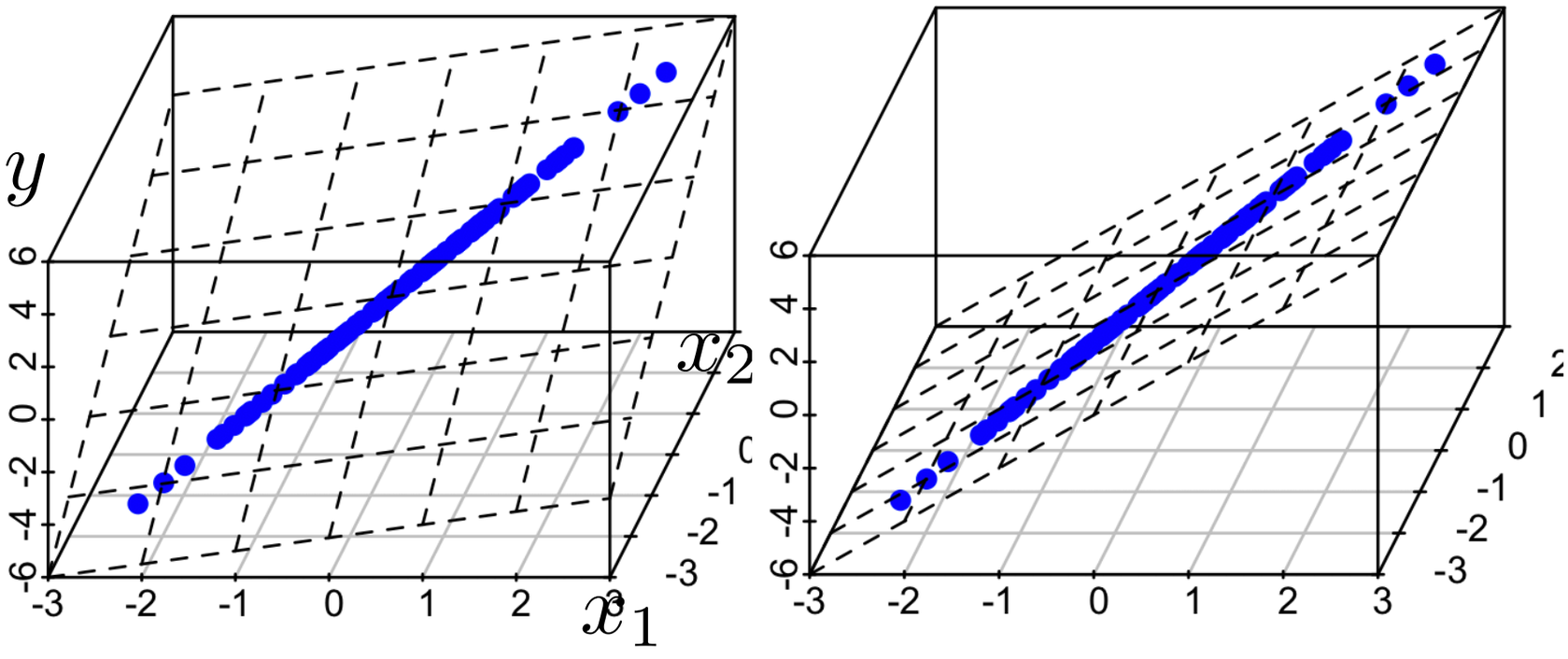
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)



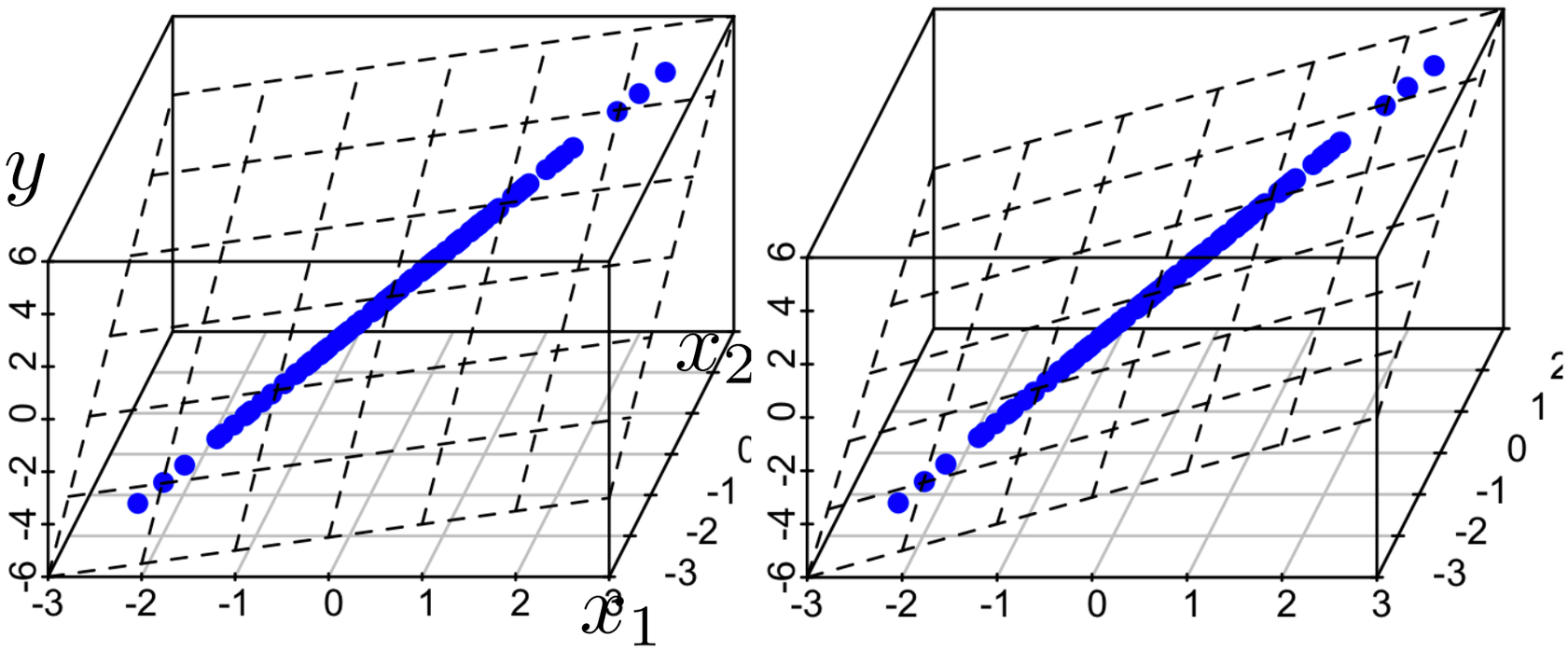
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)



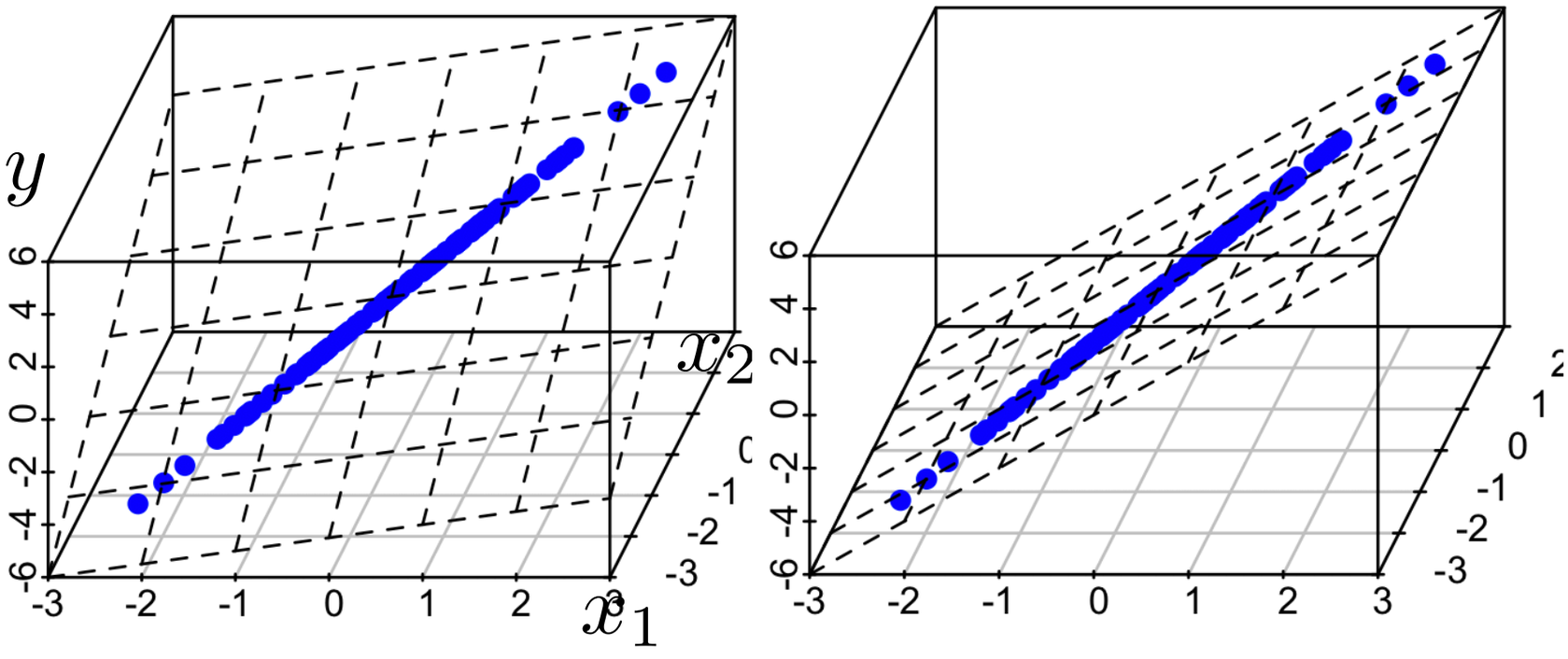
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)



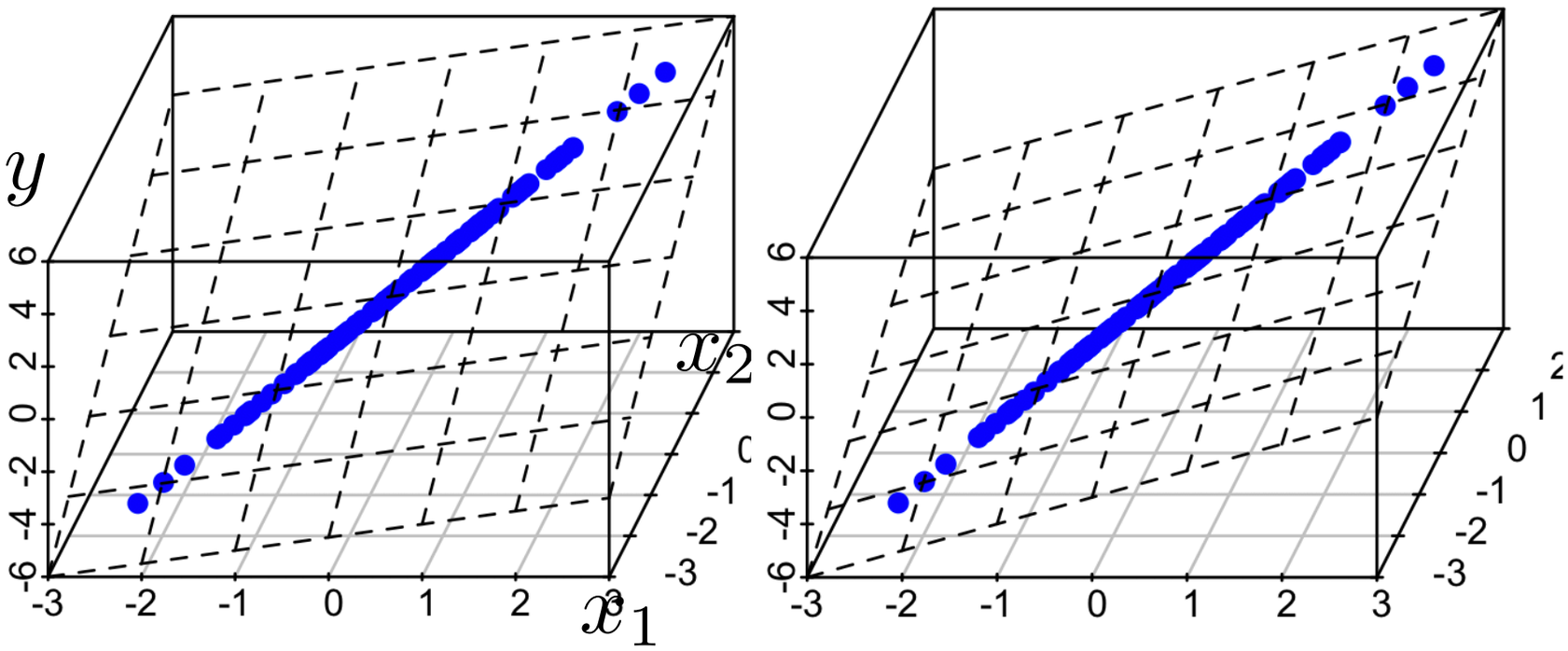
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)



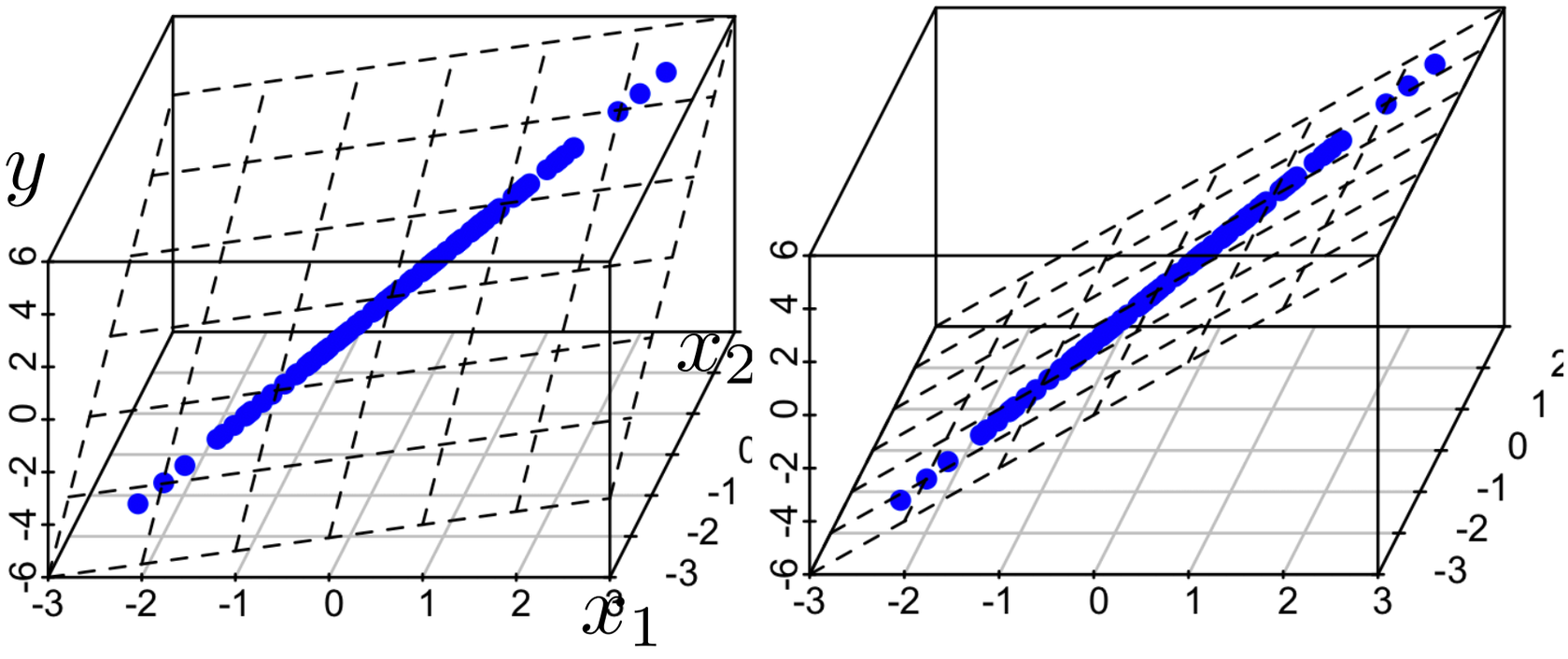
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)



What could go wrong?

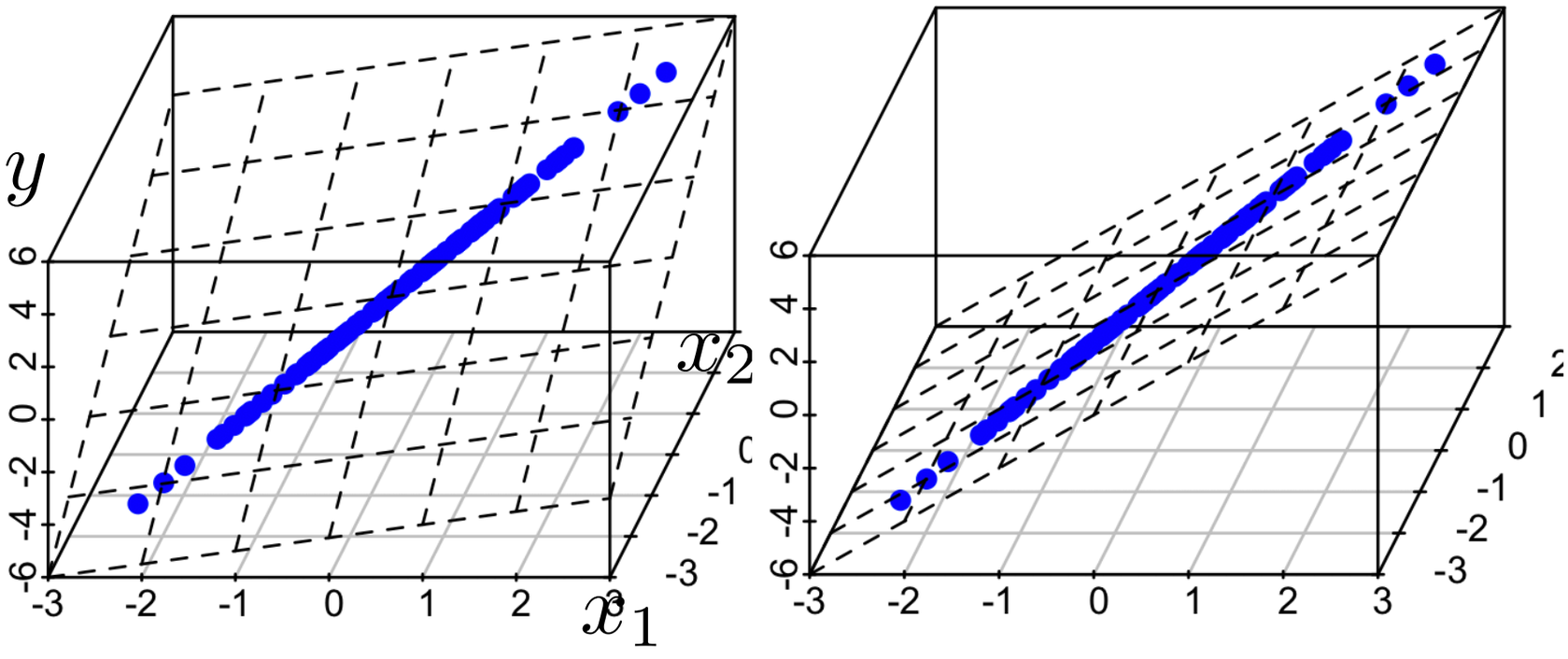
- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)



What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)

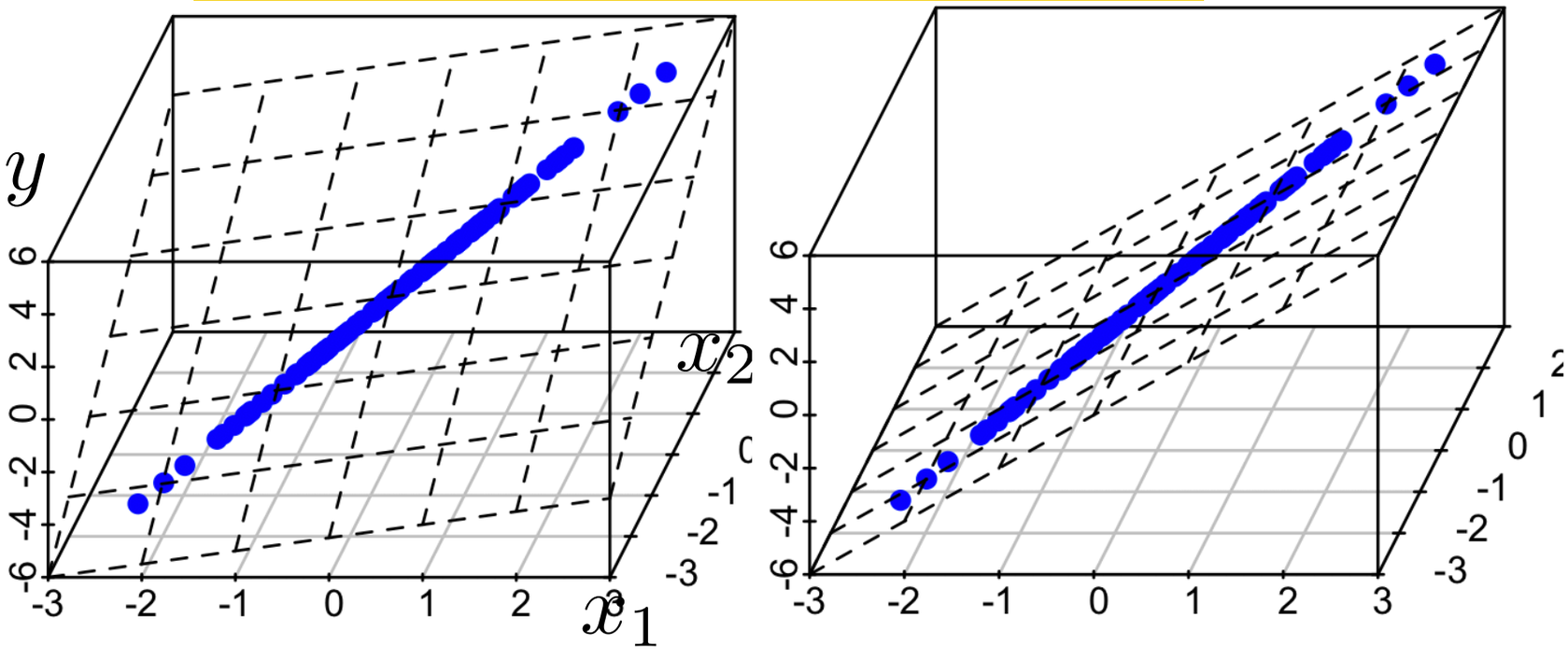
still true with any noise in *y*



What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

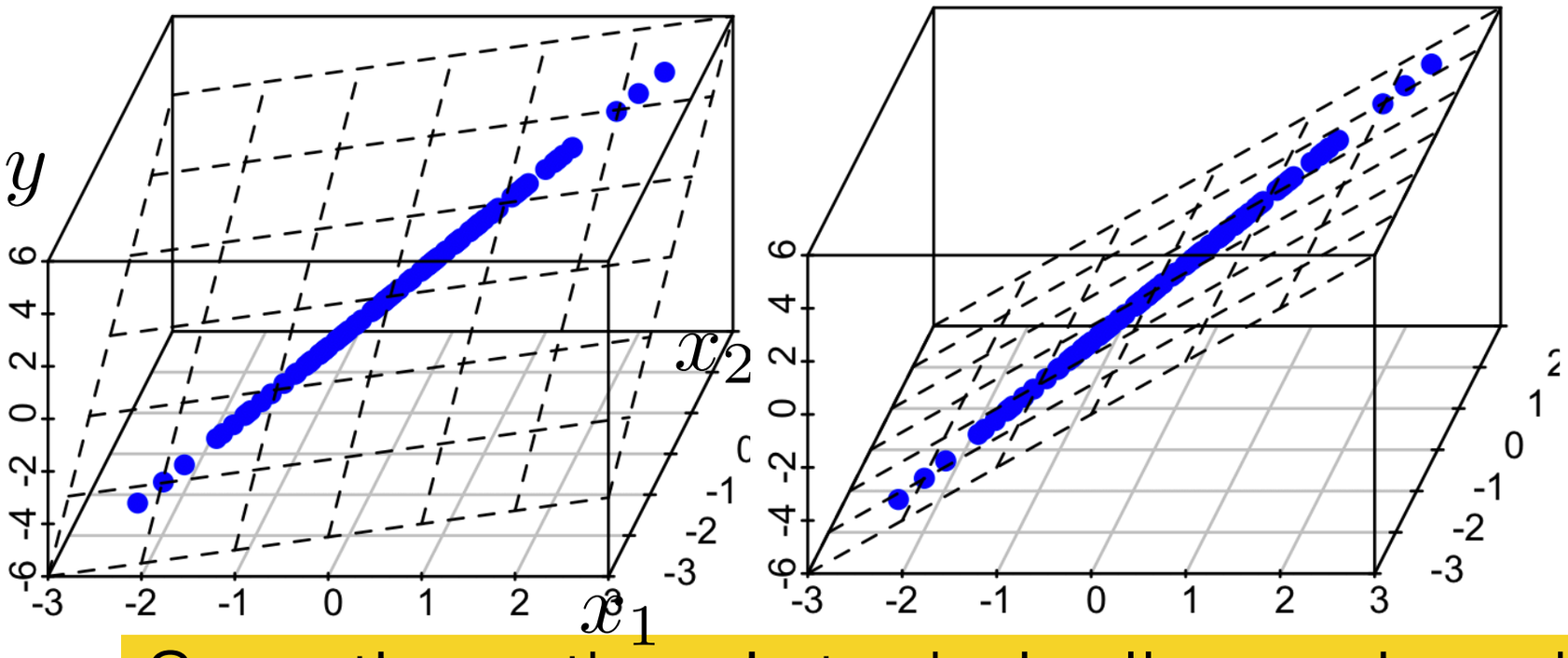
still true with any noise in *y*



What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*

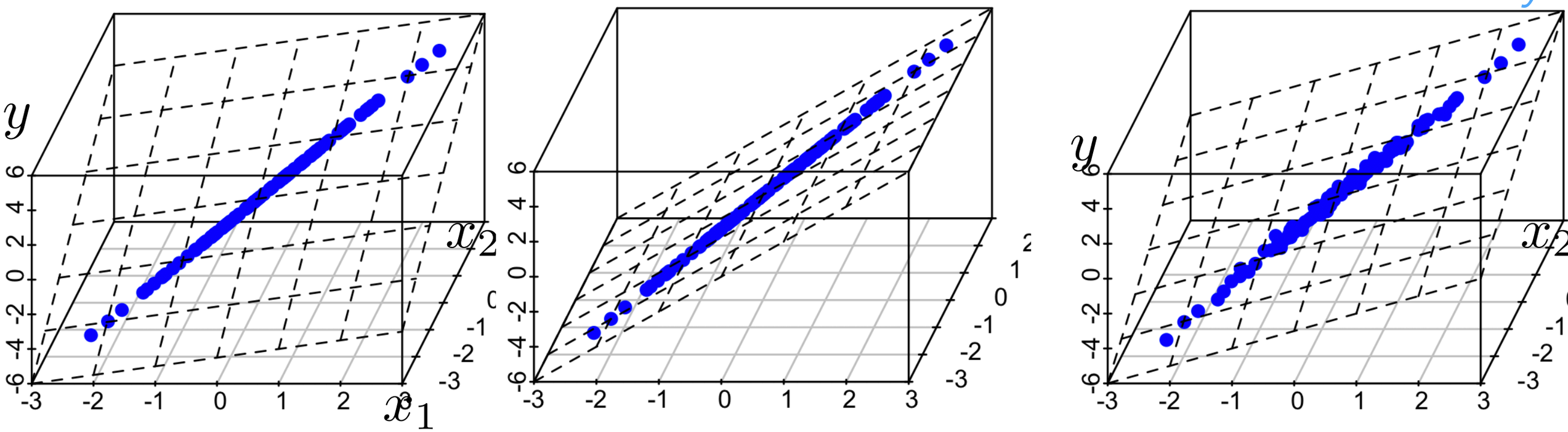


- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*

What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*

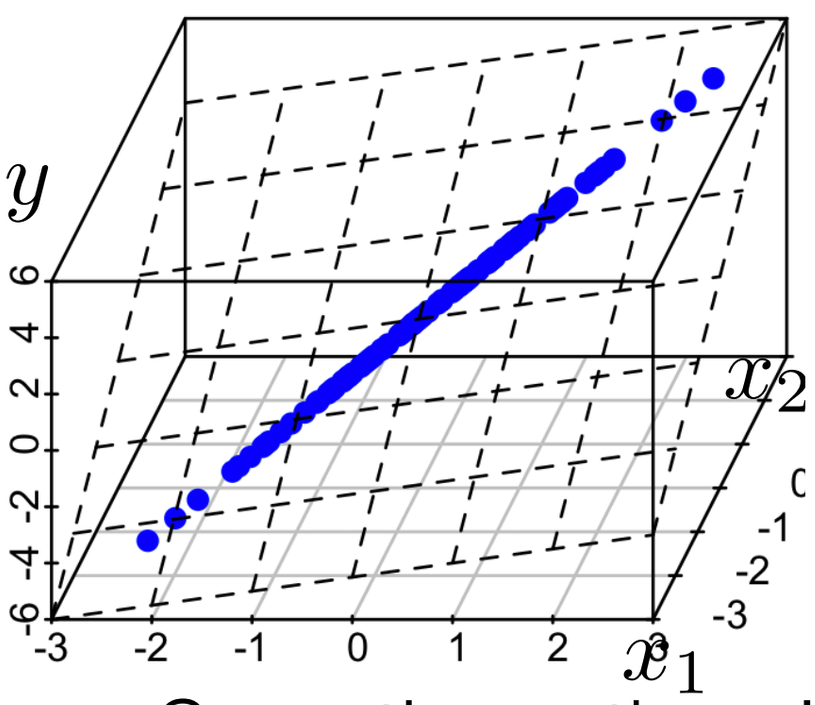


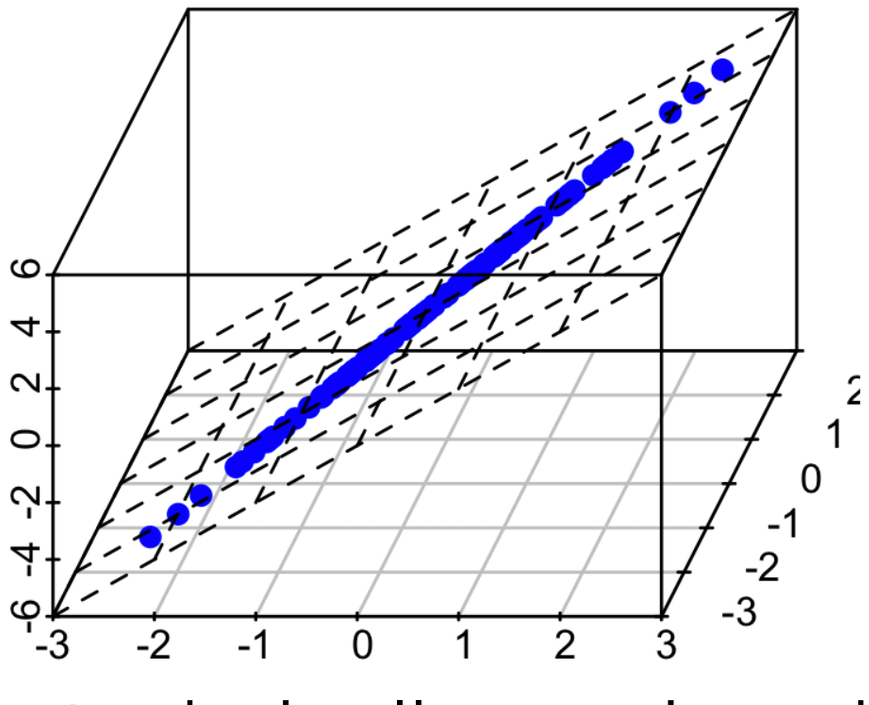
- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*

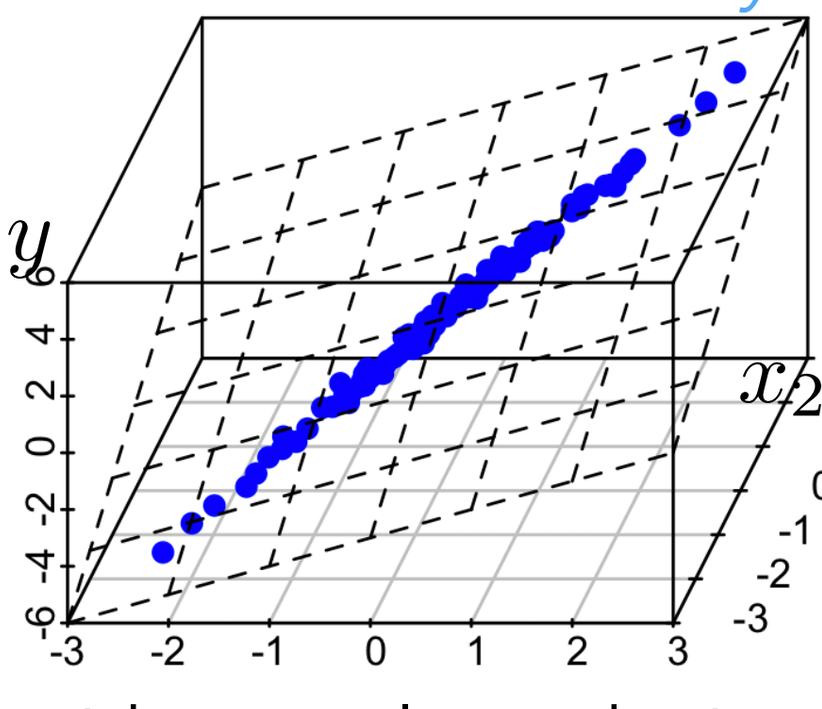
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*





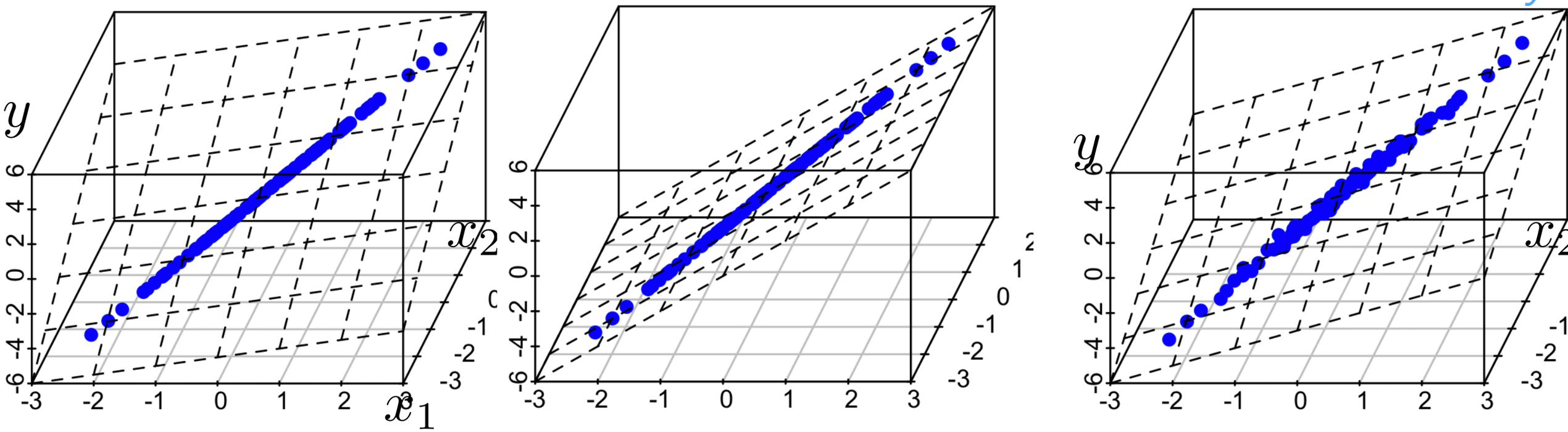


- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*
  - E.g. Predict credit card debt from credit limit and credit rating; limit and rating might be perfectly or imperfectly collinear.

What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*

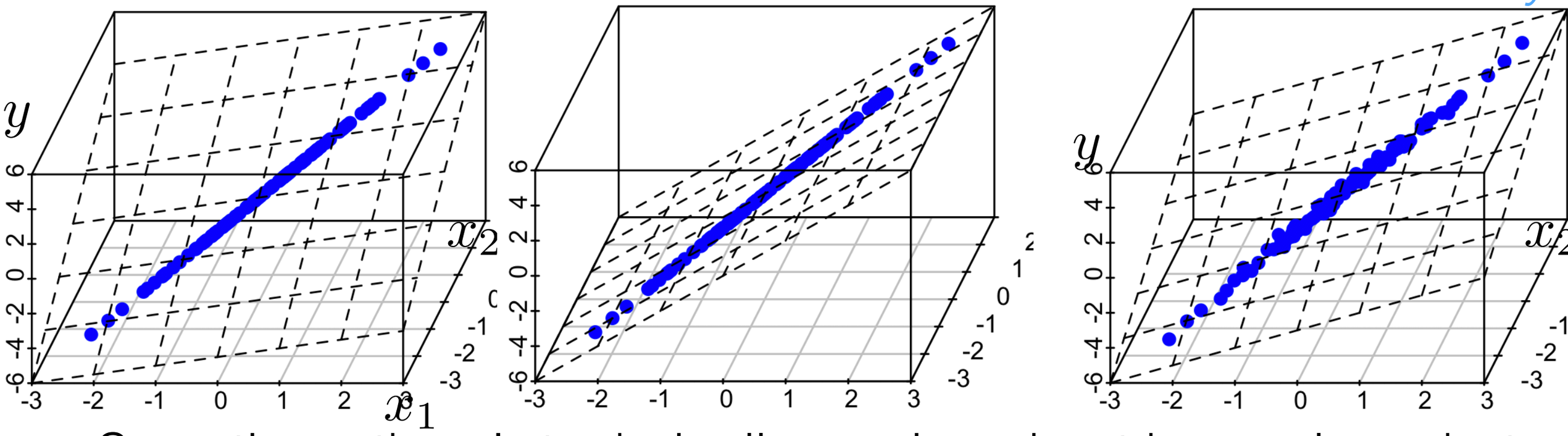


- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*
  - E.g. Predict credit card debt from credit limit and credit rating; limit and rating might be perfectly or imperfectly collinear. Or predict sales from TV and newspaper ads.

What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*

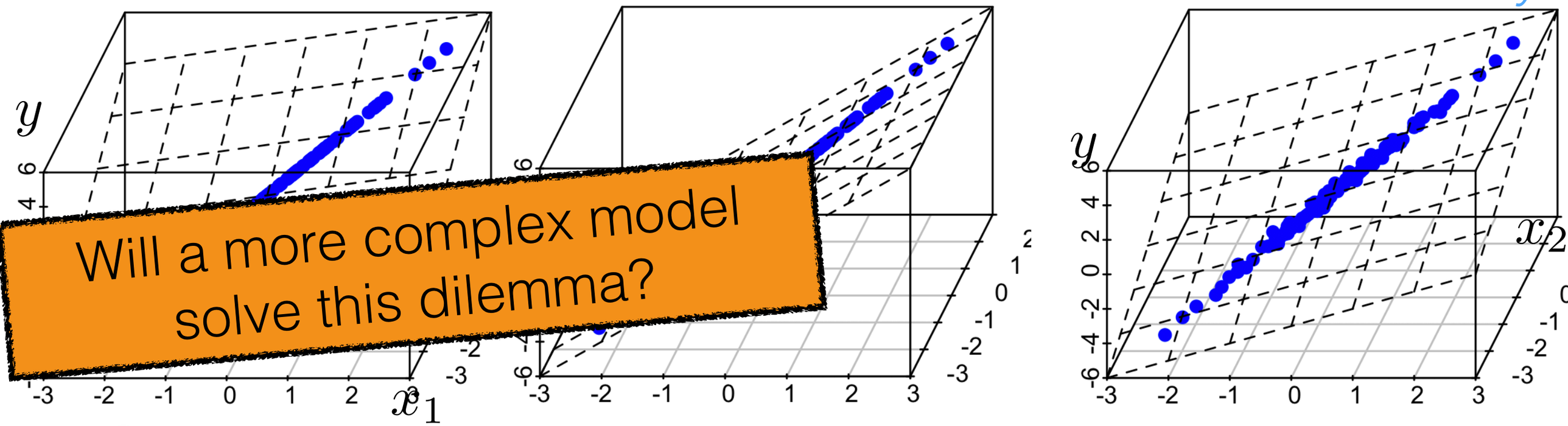


- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*
  - E.g. Predict credit card debt from credit limit and credit rating; limit and rating might be perfectly or imperfectly collinear. Or predict sales from TV and newspaper ads.
  - Recall Gaussian example with "flat" likelihood from Lec 2

What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*

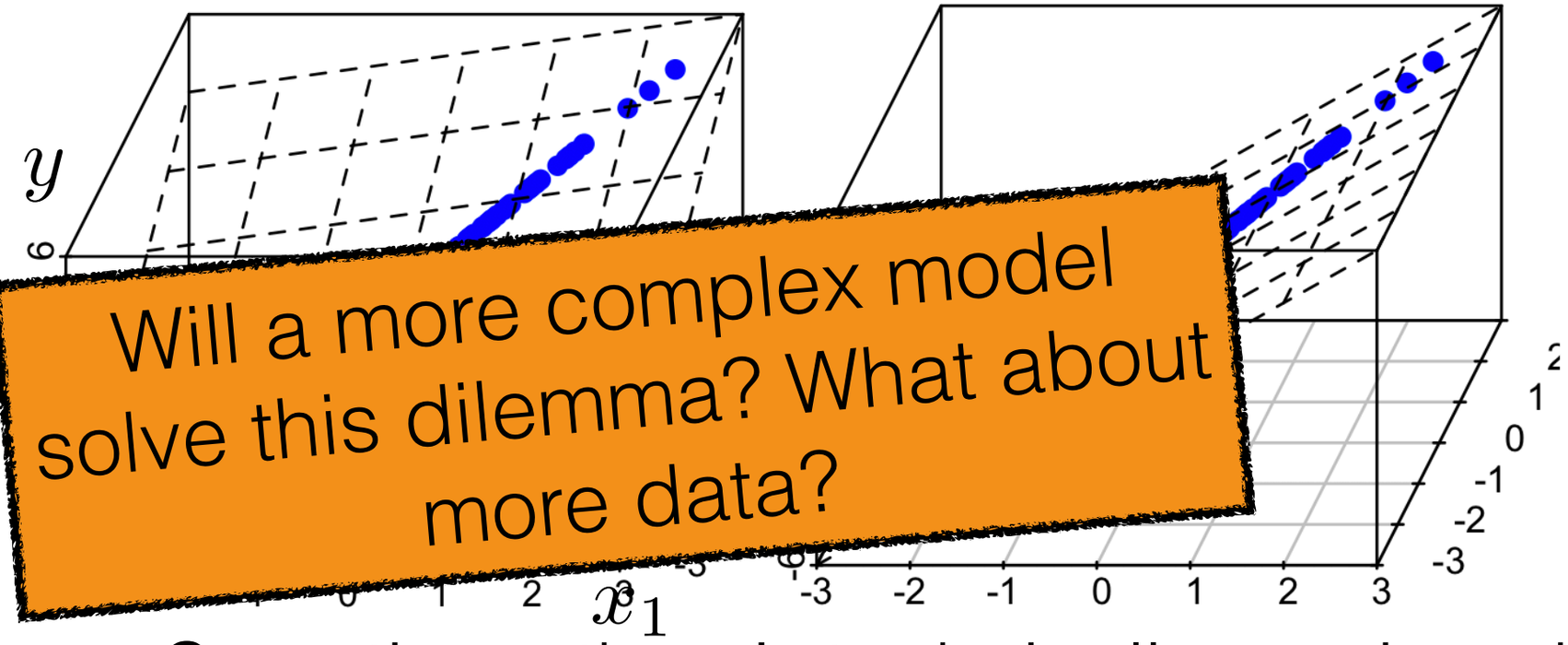


- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*
  - E.g. Predict credit card debt from credit limit and credit rating; limit and rating might be perfectly or imperfectly collinear. Or predict sales from TV and newspaper ads.
  - Recall Gaussian example with "flat" likelihood from Lec 2

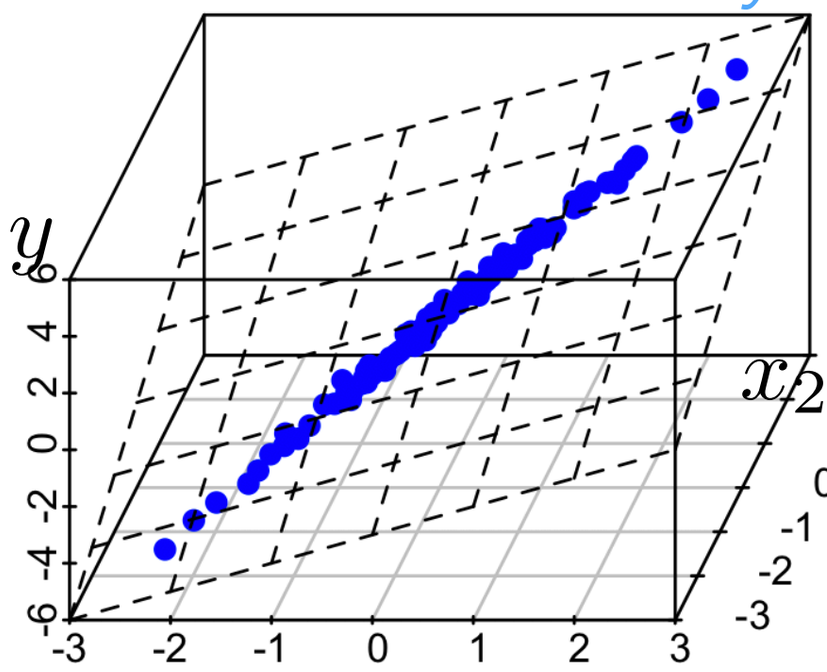
What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*



Will a more complex model solve this dilemma? What about more data?

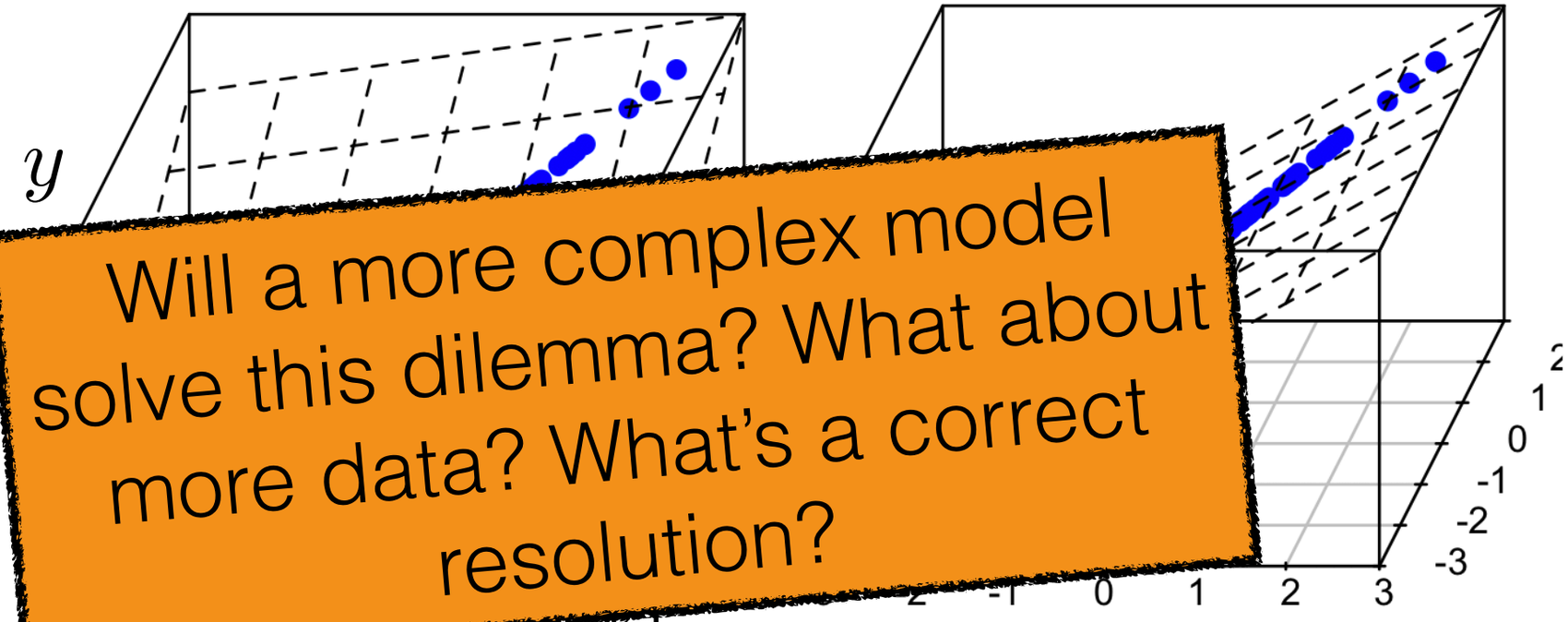


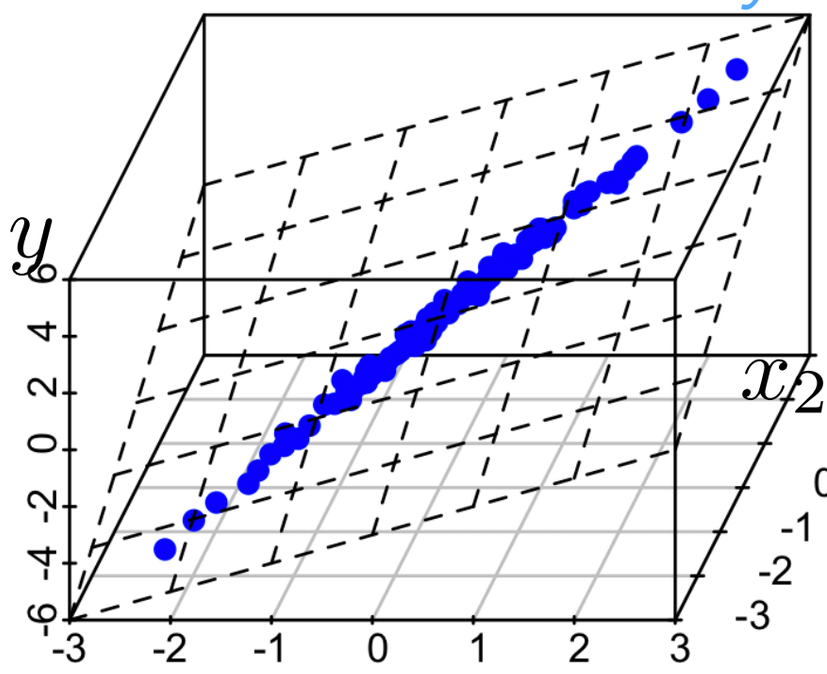
- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*
  - E.g. Predict credit card debt from credit limit and credit rating; limit and rating might be perfectly or imperfectly collinear. Or predict sales from TV and newspaper ads.
  - Recall Gaussian example with "flat" likelihood from Lec 2

What could go wrong?

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
- If two or more features are perfectly collinear, there isn't a unique best hyperplane (there are infinitely many)
  - Then $X^\top X$ is not invertible

still true with any noise in *y*





Will a more complex model solve this dilemma? What about more data? What's a correct resolution?

- Sometimes there's technically a unique best hyperplane, but just because of small/not-meaningful noise in *x*
  - E.g. Predict credit card debt from credit limit and credit rating; limit and rating might be perfectly or imperfectly collinear. Or predict sales from TV and newspaper ads.
  - Recall Gaussian example with "flat" likelihood from Lec 2

Bayes & multivariate Gaussians

Bayes & multivariate Gaussians

- Plan:

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^{N}$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$

D_yx1

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$

D_yx1 D_yx1

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$

D_yx1 D_yx1 D_yxD_y

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
  - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
  - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
  - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
  - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
  - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
    - "Normal means model"

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
  - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
  - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
    - "Normal means model"
    - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed
  - Then $p(y^{(n)}|\mu) =$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed
  - Then

$$p(y^{(n)}|\mu) = \frac{1}{\sqrt{(2\pi)^{D_y}|\Sigma|}} \exp\left\{-\frac{1}{2}(y^{(n)} - \mu)^\top \Sigma^{-1}(y^{(n)} - \mu)\right\}$$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed
  - Then

determinant

$$p(y^{(n)}|\mu) = \frac{1}{\sqrt{(2\pi)^{D_y}|\Sigma|}} \exp\left\{-\frac{1}{2}(y^{(n)} - \mu)^\top \Sigma^{-1}(y^{(n)} - \mu)\right\}$$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed
  - Then

determinant D_yxD_y

$$p(y^{(n)}|\mu) = \frac{1}{\sqrt{(2\pi)^{D_y}|\Sigma|}} \exp\left\{-\frac{1}{2}(y^{(n)} - \mu)^\top \Sigma^{-1}(y^{(n)} - \mu)\right\}$$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed
  - Then

determinant

D_yxD_y D_yx1

$$p(y^{(n)}|\mu) = \frac{1}{\sqrt{(2\pi)^{D_y}|\Sigma|}} \exp\left\{-\frac{1}{2}(y^{(n)} - \mu)^\top \Sigma^{-1}(y^{(n)} - \mu)\right\}$$

Bayes & multivariate Gaussians

- Plan:
  1. First go back to the case where data is just $\{y^{(n)}\}_{n=1}^N$
  2. Develop Bayesian inference for multivariate Gaussians
  3. Then apply those ideas to linear regression
- Now let's take a D_y-dimensional label $y^{(n)} = [y_1^{(n)}, \ldots, y_{D_y}^{(n)}]^\top$
  - E.g. D_y experiments
  - Suppose we posit a Gaussian likelihood $y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$
    - With $\mu \in \mathbb{R}^{D_y}$ & Σ positive definite D_yx1 D_yx1 D_yxD_y
    - Special case: $\Sigma = \sigma^2 I_{D_y \times D_y}$
      - "Normal means model"
      - Stein's phenomenon: can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs

[Stigler 1990, "A Galtonian perspective on shrinkage estimators"]

- To avoid notational overload, we'll treat Σ as fixed
  - Then

$$p(y^{(n)}|\mu) = \frac{1}{\sqrt{(2\pi)^{D_y}|\Sigma|}} \exp\left\{-\frac{1}{2}(y^{(n)} - \mu)^\top \Sigma^{-1}(y^{(n)} - \mu)\right\}$$

determinant; 1xD_y; D_yxD_y; D_yx1

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in ’23)

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

$p(\mu|y^{(1)}) =$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$
$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1} \mu = \mu^\top \Sigma^{-1} \mu + \mu^\top \Sigma_0^{-1} \mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

inverse covariance matrix is called the **precision matrix**

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

inverse covariance matrix is called the **precision matrix**

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right\}$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right.$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right\}$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

inverse covariance matrix is called the **precision matrix**

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

$$\Rightarrow \mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(n)}-\mu)^\top \Sigma^{-1}(y^{(n)}-\mu)\right\}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)\,, \mu_0 \in \mathbb{R}^{D_y}$ & Σ_0 pos def
  - How to choose hyperparameters in practice? Domain info
    - E.g. suppose I want to know log PM2.5 value at a sensor.
    - In past, NYC has seen near-0 to 117 μg m^{-3} (latter in '23)
- Posterior for one data point: $p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$

$$p(\mu|y^{(1)}) \propto_\mu \exp\left\{-\frac{1}{2}(y^{(1)}-\mu)^\top \Sigma^{-1}(y^{(1)}-\mu)\right\} \cdot \exp\left\{-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right\}$$

solve for posterior mean & var

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu \exp\left\{-\frac{1}{2}(\mu-\mu_1)^\top \Sigma_1^{-1}(\mu-\mu_1)\right\}$$

- So $\forall \mu, \mu^\top \Sigma_1^{-1}\mu = \mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma_0^{-1}\mu$

inverse covariance matrix is called the **precision matrix**

$$\Rightarrow \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Rightarrow \Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

sum of pos def matrices is pos def

- And $\forall \mu, \mu_1^\top \Sigma_1^{-1}\mu = (y^{(1)})^\top \Sigma^{-1}\mu + \mu_0^\top \Sigma_0^{-1}\mu$

$$\Rightarrow \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

$$\Rightarrow \mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1}$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1}$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

$$\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1$$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

$$\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1$$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1$

use old posterior as new prior

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

use old posterior as new prior

$$\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^{2} y^{(n)} + \Sigma_0^{-1}\mu_0$$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$ use old posterior as new prior

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$ use old posterior as new prior

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

$$\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$$

use old posterior as new prior

$$\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$$

- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \quad \text{i.e.} \quad \Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0) \quad \text{i.e.} \quad \Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

$$\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$$

use old posterior as new prior

$$\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$$

- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$$

use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1$ $= \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1$ $= \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

  $\mu_N = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\left(\frac{1}{\sigma^2}\sum_{n=1}^N y^{(n)} + \frac{1}{\sigma_0^2}\mu_0\right)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

  $\mu_N = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\left(\frac{1}{\sigma^2}\sum_{n=1}^N y^{(n)} + \frac{1}{\sigma_0^2}\mu_0\right)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

  $\mu_N = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \left(\frac{1}{\sigma^2}\sum_{n=1}^N y^{(n)} + \frac{1}{\sigma_0^2}\mu_0\right)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

  check: *N* = 0

  $\mu_N = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\left(\frac{1}{\sigma^2}\sum_{n=1}^N y^{(n)} + \frac{1}{\sigma_0^2}\mu_0\right)$

Bayes & multivariate Gaussians

- Likelihood: $p(y^{(n)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma), y^{(n)} \in \mathbb{R}^{D_y}$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one data point: $p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1)$

  $\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$ i.e. $\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$

  $\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$ i.e. $\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$
- Posterior for two data points: $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

  $\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^2 y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior
- Posterior for *N* data points: $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

  $\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$

  $\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^N y^{(n)} + \Sigma_0^{-1}\mu_0$

  use old posterior as new prior recursively
- Sense check: one label (D_y=1)

  $\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$

  $\mu_N = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\left(\frac{1}{\sigma^2}\sum_{n=1}^N y^{(n)} + \frac{1}{\sigma_0^2}\mu_0\right)$

  check: *N* = 0, *N* = really large

Bayesian linear regression

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} x^{(1)})^2\right\}$$

1x1

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{x}D} \underbrace{x^{(1)}}_{D\text{x}1})^2\right\}$$

1x1

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

  $p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\}$

  1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$

Note: new uses of notation here μ_n, Σ_n

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$

Note: new uses of notation here μ_n, Σ_n

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$

Note: new uses of notation here μ_n, Σ_n

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{x}D} \underbrace{x^{(1)}}_{D\text{x}1})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

Note: new uses of notation here μ_n, Σ_n

not bothering to write *x* since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{x}D} \underbrace{x^{(1)}}_{D\text{x}1})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

Note: new uses of notation here μ_n, Σ_n

not bothering to write *x* since conditioning on it everywhere

exercise: check

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

exercise: check

Note: new uses of notation here

μ_n, Σ_n

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1 1xD

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

Note: new uses of notation here μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

$$\underset{\text{DxD}}{\Sigma_1^{-1}\mu_1} = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1}x^{(1)}y^{(1)}$$

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

$$\underset{\text{DxD}}{\Sigma_1^{-1}\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}\mu_0} + \underset{\text{Dx1}}{(\sigma^2)^{-1} x^{(1)} y^{(1)}}$$

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

$$\underset{\text{DxD}}{\Sigma_1^{-1}}\underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}}\mu_0 + \underset{\text{1x1}}{(\sigma^2)^{-1}} x^{(1)} y^{(1)}$$

Note: new uses of notation here

μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1 1xD

exercise: check

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

DxD Dx1 1x1 Dx1

Note: new uses of notation here

μ_n, Σ_n

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1 1xD

exercise: check

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

DxD Dx1 1x1 Dx1 1x1

Note: new uses of notation here

μ_n, Σ_n

not bothering to write x since conditioning on it everywhere

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write x since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}}\underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}}\underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

Note: new uses of notation here μ_n, Σ_n

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

$$\underset{\text{DxD}}{\Sigma_1^{-1}}\underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}}\underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

not bothering to write x since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write x since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}}\underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}}\underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$

Note: new uses of notation here μ_n, Σ_n

- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1 1xD

exercise: check

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

DxD Dx1 1x1 Dx1 1x1

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1 1xD

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

DxD Dx1 1x1 Dx1 1x1

Note: new uses of notation here μ_n, Σ_n

exercise: check

not bothering to write x since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write x since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write x since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} X^\top X$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write x since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}}\underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}}\underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\begin{aligned}\Sigma_N^{-1} &= \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)}(x^{(n)})^\top \\ &= \Sigma_0^{-1} + (\sigma^2)^{-1} X^\top X\end{aligned}$$

recall: nth row of *X* matrix is nth data point's features

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write *x* since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\begin{aligned}\Sigma_N^{-1} &= \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top \\ &= \Sigma_0^{-1} + (\sigma^2)^{-1} X^\top X\end{aligned}$$

recall: nth row of *X* matrix is nth data point's features

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^\top x^{(1)})^2\right\}$$

1x1 1xD Dx1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} (x^{(1)})^\top$$

DxD DxD 1x1 Dx1 1xD

exercise: check

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

DxD Dx1 1x1 Dx1 1x1

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} X^\top X$$

NxD

recall: nth row of *X* matrix is nth data point's features

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write x since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\begin{aligned}\Sigma_N^{-1} &= \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)}(x^{(n)})^\top \\ &= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}\end{aligned}$$

recall: nth row of X matrix is nth data point's features

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write *x* since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

recall: nth row of *X* matrix is nth data point's features

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

$$\underset{\text{DxD}}{\Sigma_1^{-1}}\underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}}\underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}}\underset{\text{Dx1}}{x^{(1)}}\underset{\text{1x1}}{y^{(1)}}$$

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)}(x^{(n)})^\top$$

recall: nth row of *X* matrix is nth data point's features

check: is pos def

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(1)} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

DxD DxD 1x1 Dx1

exercise: check

not bothering to write *x* since conditioning on it everywhere

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

DxD Dx1 1x1 Dx1 1x1

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} X$$

DxN NxD

recall: nth row of *X* matrix is nth data point's features

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write *x* since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

recall: nth row of *X* matrix is nth data point's features

$$\Sigma_N^{-1} \mu_N = \Sigma_0^{-1} \mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write *x* since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

recall: nth row of *X* matrix is nth data point's features

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

$$\Sigma_N^{-1}\mu_N = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

$$= \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} X^\top Y$$

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

recall: nth row of *X* matrix is nth data point's features

check: is pos def

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

$$\Sigma_N^{-1}\mu_N = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

$$= \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} X^\top Y$$

recall: *Y* is a column vector of *N* labels

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

not bothering to write *x* since conditioning on it everywhere

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

recall: nth row of *X* matrix is nth data point's features

check: is pos def

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

$$\Sigma_N^{-1}\mu_N = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

$$= \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} X^\top Y$$

Nx1

recall: *Y* is a column vector of *N* labels

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

recall: nth row of *X* matrix is nth data point's features

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

$$\Sigma_N^{-1}\mu_N = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

$$= \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{Y}_{\text{Nx1}}$$

recall: *Y* is a column vector of *N* labels

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

recall: nth row of *X* matrix is nth data point's features

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

$$\underset{\text{DxD}}{\Sigma_N^{-1}} \mu_N = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

$$= \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{Y}_{\text{Nx1}}$$

recall: *Y* is a column vector of *N* labels

Bayesian linear regression

- Likelihood: $\mathcal{N}(y^{(n)}|\theta^\top x^{(n)}, \sigma^2)$; we'll assume $\sigma^2 > 0$ is known

$$p(y^{(1)}|x^{(1)}, \theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2}(y^{(1)} - \underbrace{\theta^\top}_{1\text{xD}} \underbrace{x^{(1)}}_{\text{Dx1}})^2\right\}$$

1x1

Note: new uses of notation here μ_n, Σ_n

- Conjugate prior: $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Posterior, one point: $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$
- Similar calculations to previous case for posterior:

$$\underset{\text{DxD}}{\Sigma_1^{-1}} = \underset{\text{DxD}}{\Sigma_0^{-1}} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underbrace{(x^{(1)})^\top}_{\text{1xD}}$$

exercise: check

$$\underset{\text{DxD}}{\Sigma_1^{-1}} \underset{\text{Dx1}}{\mu_1} = \underset{\text{DxD}}{\Sigma_0^{-1}} \underset{\text{Dx1}}{\mu_0} + \underset{\text{1x1}}{(\sigma^2)^{-1}} \underset{\text{Dx1}}{x^{(1)}} \underset{\text{1x1}}{y^{(1)}}$$

not bothering to write *x* since conditioning on it everywhere

- Recursively for *N* data points: $p(\theta|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} (x^{(n)})^\top$$

check: is pos def

recall: nth row of *X* matrix is nth data point's features

$$= \Sigma_0^{-1} + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underbrace{X}_{\text{NxD}}$$

$$\underset{\text{DxD Dx1}}{\Sigma_N^{-1} \mu_N} = \Sigma_0^{-1} \mu_0 + (\sigma^2)^{-1} \sum_{n=1}^N x^{(n)} y^{(n)}$$

$$= \Sigma_0^{-1} \mu_0 + (\sigma^2)^{-1} \underbrace{X^\top}_{\text{DxN}} \underset{\text{Nx1}}{Y}$$

recall: *Y* is a column vector of *N* labels