REGRESSION

BST228 Applied Bayesian Analysis

Speaker notes

No notes on this slide.

OFFICE HOUR

Room 434 in Building 2 Thursdays 11:30am to noon. Zoom at https://harvard.zoom.us/j/5482315734.

Speaker notes

No notes on this slide.

RECAP

- Metropolis with symmetric proposal distribution.
- Gibbs sampler with conditional distributions.
- Convergence diagnostics ($\hat{R}$, autocorrelation, trace plots, effective sample size).
- Tweaking sampling algorithms (proposal distribution, blocking, initialization).

Speaker notes

- Metropolis algo. proposes new parameter values θ' and accepts with prob. $\frac{p(\theta')}{p(\theta)}$, where θ are current parameter values.
- Gibbs samples the posterior by iteratively sampling from conditional dist.; can be more tractable than full joint dist.
- $\hat{R}$ is ratio of between- to within-chain variance. If $\hat{R} > 1.1$, likely not converged because chains are not sampling the same dist.
- Autocorrelation and effective sample size (ESS) measure efficiency of the algorithm. High autocorrelation and low ESS → many iterations needed.
- Trace plots can be instructive but are not feasible for models with many parameters.
- Tweaking samplers is important to explore posterior efficiently, e.g., scale of proposal for Metropolis, order and sampling parameters together for Gibbs.

```
> sample_one <- function(log_target, x, scale) {
+   # Draw proposal and calculate log ratio.
+   proposal <- rnorm(length(x), x, scale)
+   log_ratio <- log_target(proposal) - log_target(x)
+   # Accept-reject step.
+   if (log(runif(1)) < log_ratio) {
+     result <- list(accept = 1, value = proposal)
+   } else {
+     result <- list(accept = 0, value = x)
+   }
+   result
+ }
>
> set.seed(2024)
> log_target <- function(x) {
+   sum(dnorm(x, c(1, 2), c(.5, 2), log = TRUE))
+ }
> sample_one(log_target, c(0, -1), c(1, 1))
$accept
[1] 1

$value
[1]  0.9819694 -0.5312850

>
```

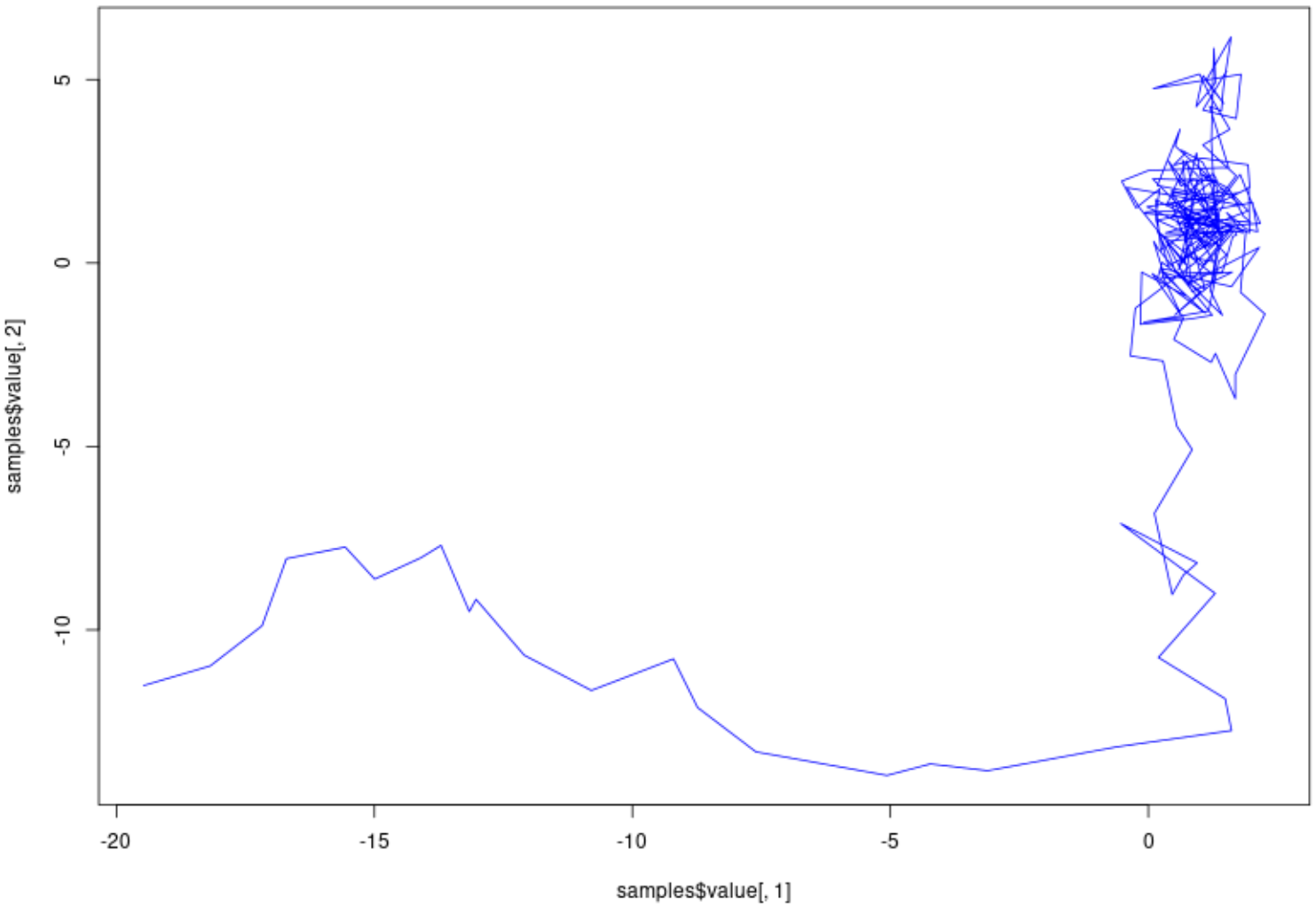
Speaker notes

- Want generic sampling algorithm so we don't need to start from scratch.
- `sample_one` runs one Metropolis iteration normal proposal dist. (line #3), ratio of densities (#4) using log for numerical stability, accept-reject (#6-10).
- #15-17 declares a target distribution (normal with mean `c(1, 2)` and scale `c(.5, 2)`).
- We run one sampling step in #18 which is accepted.
- Note: This sampler can propose values outside the support of the target dist., e.g., for positive parameters. Two solutions: (a) reject all proposals outside support or (b) make a change of variables such that all parameters are real. The latter is the approach taken by Stan which we will use later in the course.

```
> source("mh_one.R")
>
> sample_n <- function(log_target, x, scale, n) {
+   # Initialize the samples.
+   samples <- list(
+     value = matrix(nrow = n, ncol = length(x)),
+     accept = numeric(n)
+   )
+   # Run the sampling loop.
+   for (i in 1:n) {
+     current <- sample_one(log_target, x, scale)
+     samples$value[i, ] <- current$value
+     samples$accept[i] <- current$accept
+     x <- current$value
+   }
+   samples
+ }
>
> # Draw samples and plot them.
> samples <- sample_n(log_target, c(-20, -11), c(1, 1), 500)
> mean(samples$accept)
[1] 0.48
> png("samples.png", width = 800, height = 600)
> plot(samples$value[, 1], samples$value[, 2], type = "l", col = "blue")
> dev.off()
null device
          1
>
```

Speaker notes

- `sample_n` uses `sample_one` to run a sampling loop.
- Lines #5-8 initialize variables to keep track of samples.
- #10-15 runs the sampling by iteratively calling `sample_one`.
- #20 runs the sampler on the previously defined target distribution; #21 evaluates the mean acceptance prob.
- #23-25 plots the trajectory of the sampler (see next slide).



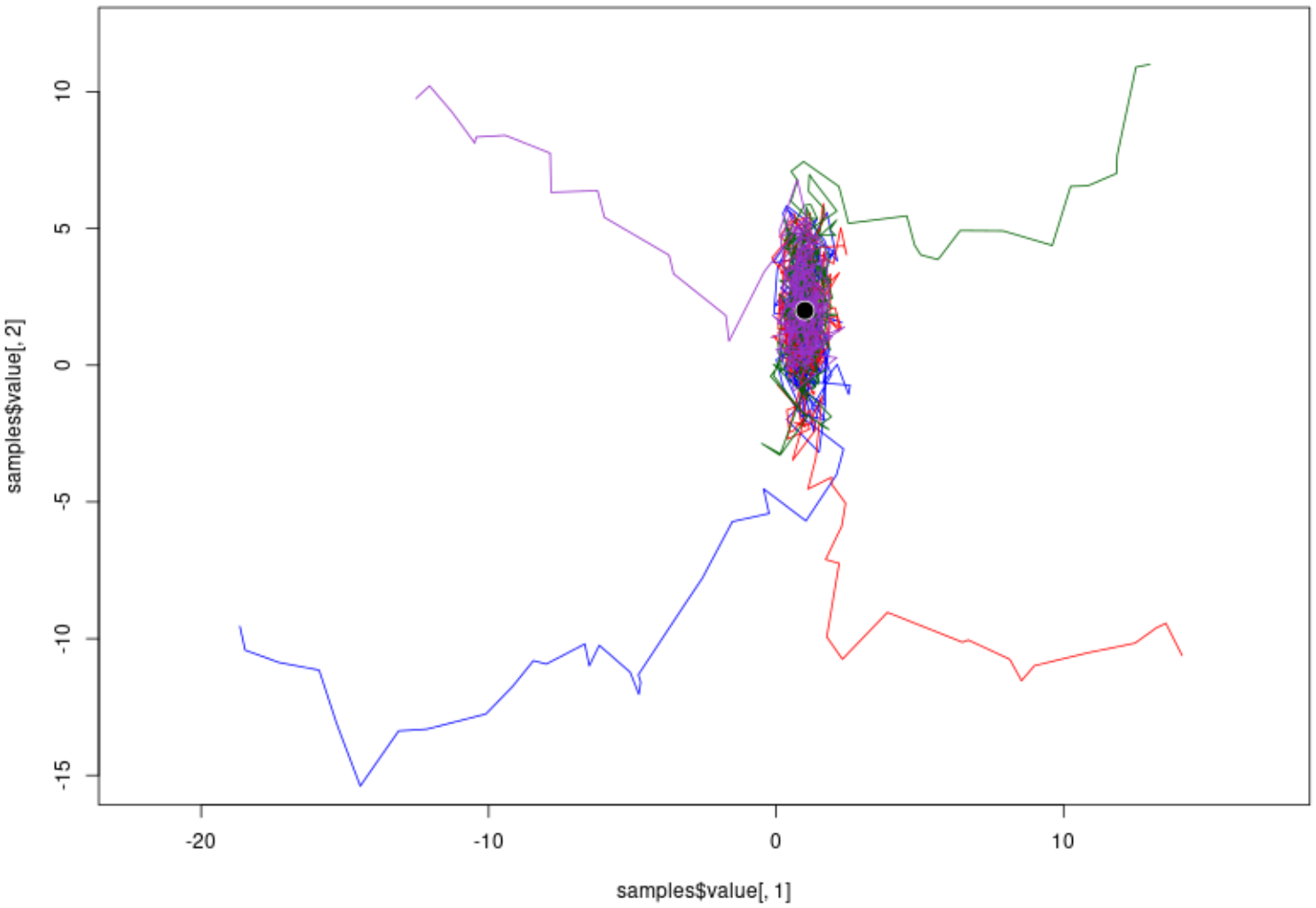
Speaker notes

- Figure shows trajectory of the sampler, starting in the lower left corner of the plot.
- Early *burnin* samples should be discarded until the sampler reaches the target dist.
- After a sufficient number of *burnin* samples, the sampler explores the target dist. well.

```
> source("mh_n.R")
>
> # Draw samples and plot them.
> png("samples_multiple.png", width = 800, height = 600)
> samples <- sample_n(log_target, c(-20, -11), c(1, 1), 500)
> plot(samples$value[, 1], samples$value[, 2], type = "l", col = "blue",
+      xlim = c(-22, 17), ylim = c(-15, 12))
> samples <- sample_n(log_target, c(15, -11), c(1, 1), 500)
> lines(samples$value[, 1], samples$value[, 2], col = "red")
> samples <- sample_n(log_target, c(13, 11), c(1, 1), 500)
> lines(samples$value[, 1], samples$value[, 2], col = "darkgreen")
> samples <- sample_n(log_target, c(-13, 11), c(1, 1), 500)
> lines(samples$value[, 1], samples$value[, 2], col = "darkorchid")
> points(1, 2, col = "white", bg = "black", pch = 21, cex = 2)
> dev.off()
null device
          1
>
```

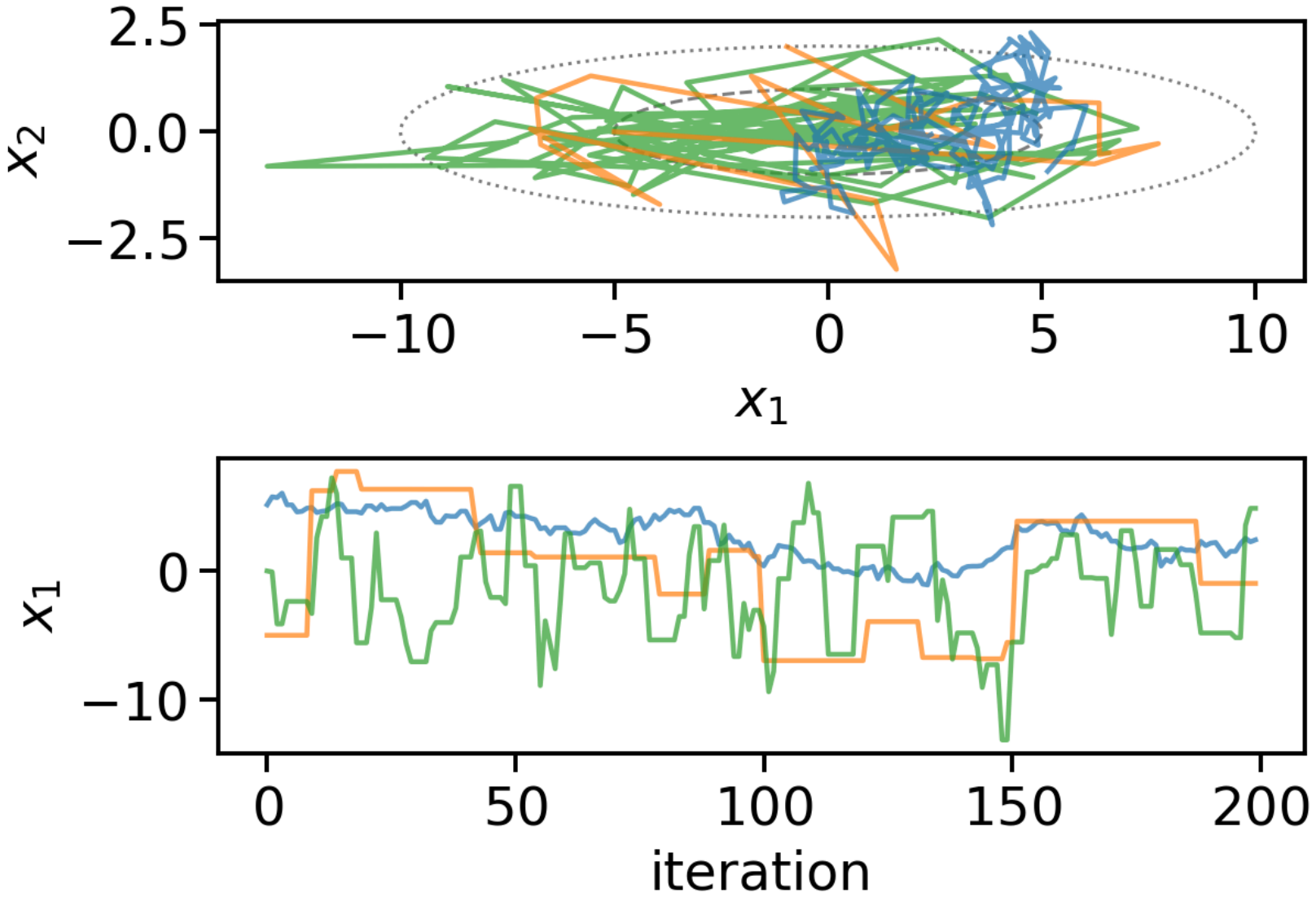
Speaker notes

- We run and plot four different chains with different initial conditions (lines #5-13).
- #14 shows the true posterior mean as a black circle with white edge.



Speaker notes

- Samplers with different initial conditions approach the distribution from different directions.
- For early samples, $\hat{R}$ would be large because the variance between chains >> variance within chains.
- Running multiple chains from different starting points is always a good idea to check for convergence.



Speaker notes

- Implementation supports different scales (`scale` argument can be a vector). Important for posteriors that are not isotropic.
- Top panel shows a normal bivariate posterior with different scales for each dimension. Using the same proposal scale for both parameters is inefficient. If `scale` is too small (blue), exploration is slow in the larger dimension (high autocorrelation). If `scale` is too big (orange), exploration is slow because samples are often rejected.
- Bottom panel shows trace plots for x_1: orange has low acceptance rate, blue has high autocorrelation, and green is well-tuned (uses proposal scales tuned to the target dist.).
- There are arguments for optimal scales. In practice, we often use a warmup phase to find the right scales using software like Stan.

LINEAR REGRESSION

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}\left(0, \tau^{-1}\right)$$

or

$$y_i \sim \text{Normal}\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tau^{-1}\right).$$

Speaker notes

- Linear regression is one of the most commonly used models.
- y_i is the response of the i^{th} observation.
- x_{ij} is the j^{th} feature of the i^{th} observation.
- ϵ_i captures residual noise but also "swallows" model misspecification; τ is the precision of observations.
- β_0 is the intercept and β_j for $j > 0$ are regression coefficients for each of p features.
- Alternative formulation is often easier for figuring out what the likelihood is: Independent noisy observations of the predictor $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.

A more concise representation in vector notation is

$$\mathbf{y} \sim \text{Normal}\left(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}\right),$$

where $\mathbf{I}$ is the identity matrix and

$$\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1q} \\ 1 & x_{21} & \ldots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nq} \end{pmatrix}$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^\mathsf{T}.$$

Speaker notes

- Why vector notation? We don't have to write as much; easier to manipulate and interpret.
- We need to further investigate the distribution for $\mathbf{y}$ because $\mathbf{y}$ is a vector. The distribution is a *multivariate* normal distribution (MVN).
- MVNs generalize normal distributions and support correlated outcomes.

MULTIVARIATE NORMAL DISTRIBUTION

A multivariate normal random variable $\mathbf{y} \in \mathbb{R}^n$ has density

$$p\left(\mathbf{y} \mid \boldsymbol{\nu}, \boldsymbol{\kappa}\right) = \left(2\pi\right)^{n/2} \left|\boldsymbol{\kappa}\right|^{1/2} \times \exp\left(-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\nu}\right)^{\mathsf{T}} \boldsymbol{\kappa} \left(\mathbf{y} - \boldsymbol{\nu}\right)\right)$$

with mean $\boldsymbol{\nu}$ and precision $\boldsymbol{\kappa}$.

Speaker notes

- MVN density has the same functional form as regular normal distribution except scalars are replaced by vectors ($\mathbf{y}$ and $\boldsymbol{\nu}$) and matrices ($\boldsymbol{\kappa}$).

MULTIVARIATE NORMAL TO INDEPENDENT NORMAL

For $\boldsymbol{\kappa} = \tau\mathbf{I}$, the elements of $\mathbf{y}$ reduce to independent samples from a normal distribution, where $\mathbf{I}$ is the identity matrix. Consider the density

$$\begin{aligned} p\left(\mathbf{y} \mid \boldsymbol{\nu}, \boldsymbol{\kappa}\right) &= \left(2\pi\right)^{n/2} \left|\boldsymbol{\kappa}\right|^{1/2} \exp\left[-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\nu}\right)^{\mathsf{T}} \boldsymbol{\kappa}\left(\mathbf{y}-\boldsymbol{\nu}\right)\right] \\ &= \left(2\pi\right)^{n/2} \left|\tau\mathbf{I}\right|^{1/2} \exp\left[-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\nu}\right)^{\mathsf{T}} \tau\mathbf{I}\left(\mathbf{y}-\boldsymbol{\nu}\right)\right] \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{\tau}{2}\left(\mathbf{y}-\boldsymbol{\nu}\right)^{\mathsf{T}}\left(\mathbf{y}-\boldsymbol{\nu}\right)\right], \end{aligned}$$

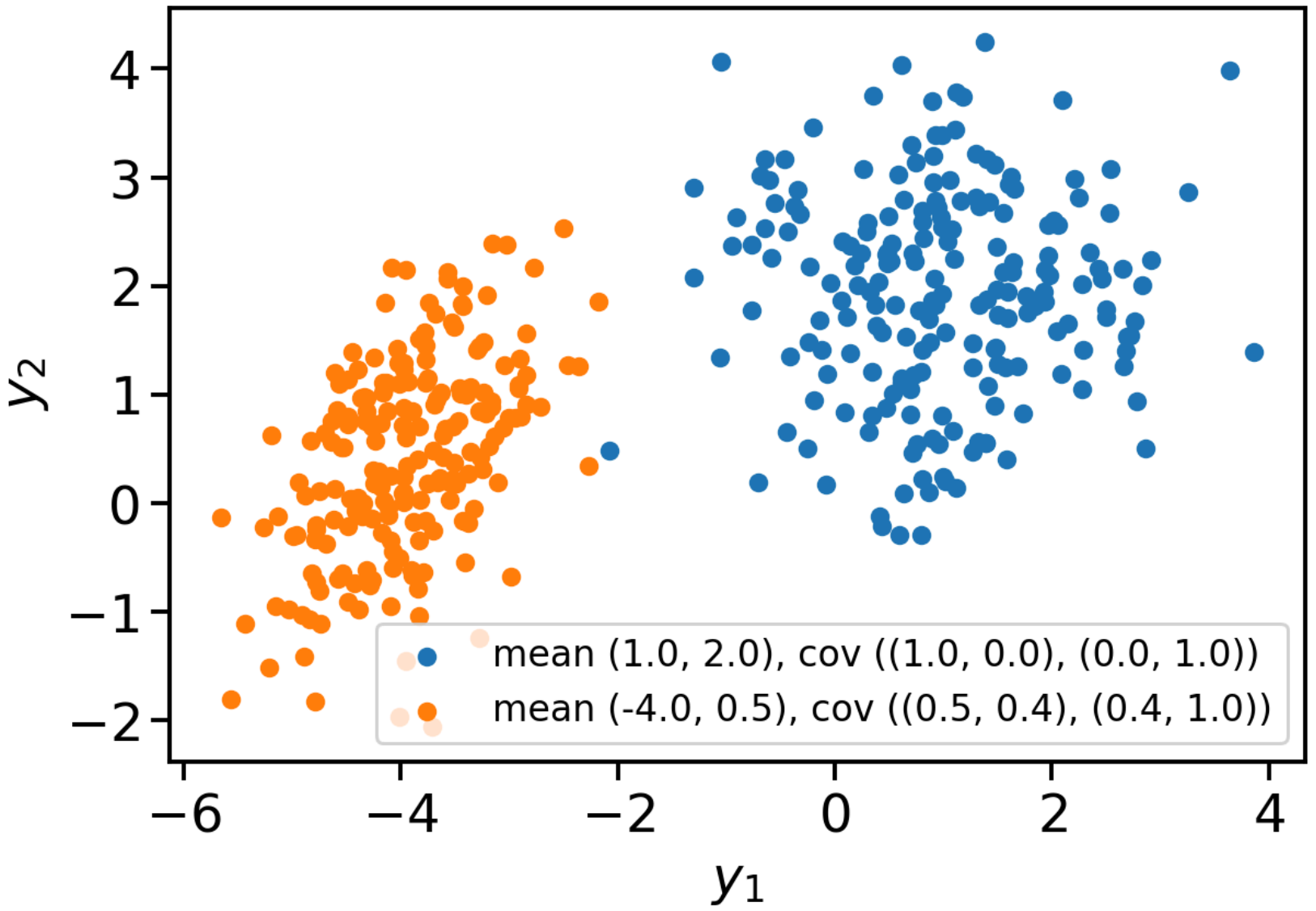
where the factor of $\tau^{n/2}$ follows from the identity $\left|a\mathbf{B}\right| = a^n \left|\mathbf{B}\right|$ for any scalar a and square matrix $\mathbf{B}$ and $\left|\mathbf{I}\right| = 1$. The expression in brackets follows because the scalar τ commutes with inner products.

The inner product can be expressed as $\left(\mathbf{y}-\boldsymbol{\nu}\right)^{\mathsf{T}}\left(\mathbf{y}-\boldsymbol{\nu}\right) = \sum_{i=1}^{n}\left(y_i-\nu_i\right)^2$, and we recover the likelihood of n independent normal samples

$$p\left(\mathbf{y} \mid \boldsymbol{\nu}, \boldsymbol{\kappa}\right) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}\left(y_i-\nu_i\right)^2\right].$$

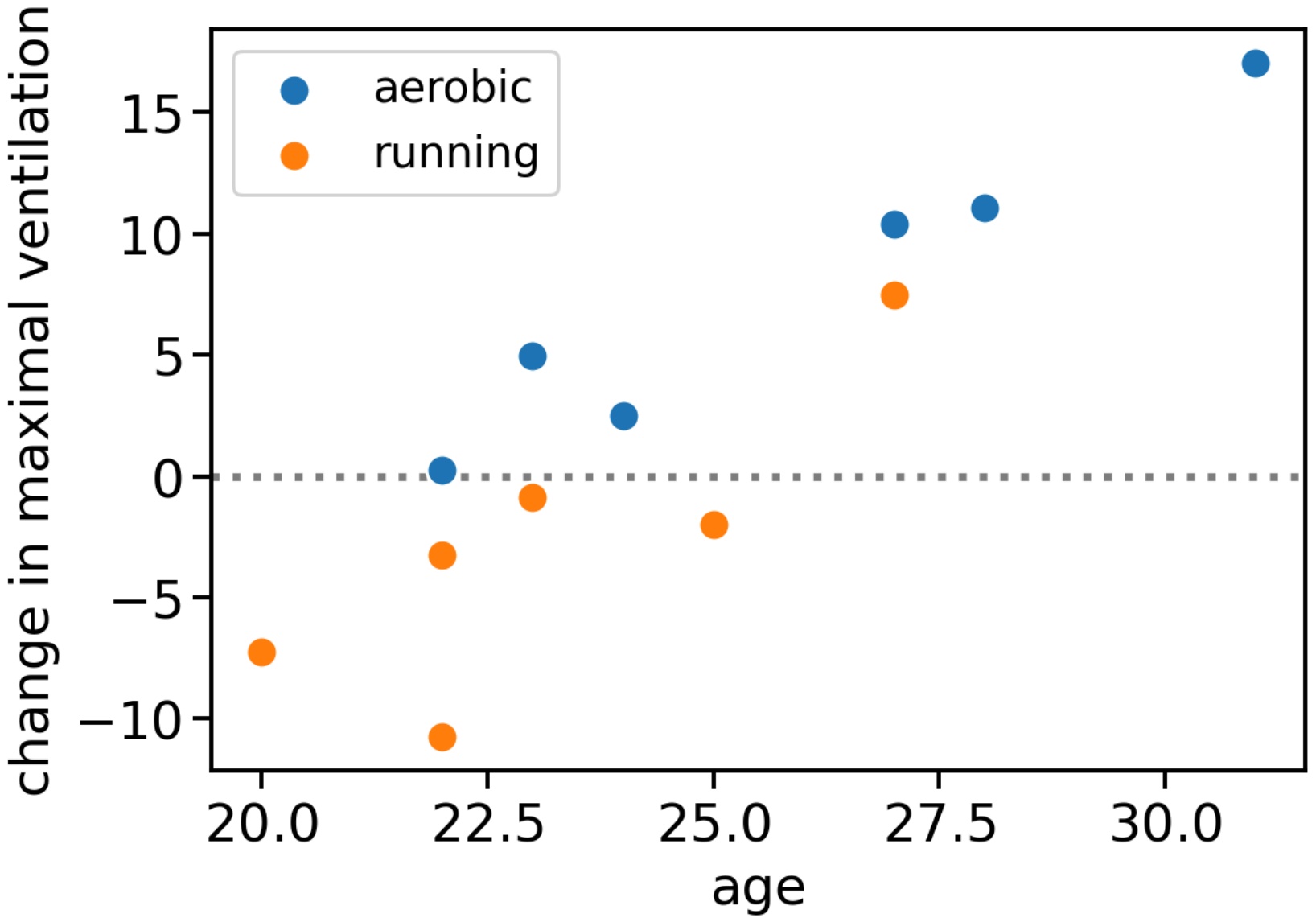
Speaker notes

No notes on this slide.



Speaker notes

- Examples of samples from MVN.
- When covariance matrix $\mathbf{\Sigma} = \boldsymbol{\tau}^{-1}$ is diagonal, we have independent samples (blue).
- Non-zero off-diagonals induce correlation between samples (orange). We also have a “stretched” ellipse because the diagonal elements are different. The distribution is elongated in the y_2 dimension because $\mathbf{\Sigma}_{22} > \mathbf{\Sigma}_{11}$.
- In the regression sample, we assume independent observations. So why should we bother with the more complex MVN distribution? Because it simplifies the algebra for deriving conditional distributions for a Gibbs sampler.



Speaker notes

- We consider the example of a sample of 12 people participating in a randomized control trial.
- Each is assigned to one of two exercise regimes: `running` on a flat service or `aerobic` exercise.
- The change in maximal O_2 exchange $\mathbf{y}$ in L/min was recorded, comparing pre and post treatment.
- We also have other data $\mathbf{Z}$ comprising the assigned treatment and participant age.
- The change in O_2 exchange conditional on other data is of primary interest, e.g., to predict changes for other members of the population.

VENTILATION DATA

- responses $\mathbf{y} = (17.05, 4.96, \ldots, -7.25)^\intercal$.
- other data

$$\mathbf{Z} = \begin{pmatrix} 1 & 31 \\ 1 & 23 \\ \vdots & \vdots \\ 0 & 20 \end{pmatrix},$$

where the first column is an `aerobic` indicator and the second is age.

Speaker notes

- Data comprise a response vector $\mathbf{y} \in \mathbb{R}^{12}$ and other data $\mathbf{Z} \in \mathbb{R}^{12 \times 2}$.
- We use $\mathbf{Z}$ instead of $\mathbf{X}$ here because we may want to use other features beyond the raw data for regression.

DATA TO FEATURES

We transform the data to features:

- $x_{i0} = 1$ is the intercept,
- $x_{i1} = z_{i1}$ is the `aerobic` indicator,
- $x_{i2} = z_{i2}$ is the age,
- $x_{i3} = z_{i1} \times z_{i2}$ is an interaction term.

Speaker notes

- We use index notation for the elements of the features (aka design matrix), i.e., x_{ij} is the element in the i^{th} row and j^{th} column of $\mathbb{X}$.
- The `aerobic` indicator captures common changes between the two treatment groups independent of age.
- The interaction term allows for different slopes with respect to age in the two treatment groups.

COMPLETING THE MODEL

We have the likelihood

$$\mathbf{y} \sim \text{Normal}\left(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}\right)$$

and complete the model with priors

$$\boldsymbol{\beta} \sim \text{Normal}\left(\boldsymbol{\nu}_0, \boldsymbol{\kappa}_0^{-1}\right)$$
$$\tau \sim \text{Gamma}\left(a_0, b_0\right).$$

Speaker notes

- This looks very much like the structure of the painful algebra we went through two weeks ago.
- We could do that again and get a closed-form posterior for the regression.
- Instead, we consider a Gibbs sampler because the conditional distributions aren't too bad–but still not great.

CONDITIONAL POSTERIOR FOR REGRESSION COEFFICIENTS β (1 / 2)

The conditional posterior for regression coefficients $\boldsymbol{\beta}$ is

$$
\begin{aligned}
p\left(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right) &= p\left(\mathbf{y} \mid \mathbf{X}, \tau, \boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}\right) \\
&\propto \exp\left[-\frac{\tau}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\intercal}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) - \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\nu}_0\right)^{\intercal}\boldsymbol{\kappa}_0\left(\boldsymbol{\beta} - \boldsymbol{\nu}_0\right)\right] \\
&\propto \exp\left[-\frac{\tau}{2}\left(\mathbf{y}^{\intercal}\mathbf{y} - \mathbf{y}^{\intercal}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} + \boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\beta}\right) - \frac{1}{2}\left(\boldsymbol{\beta}^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\nu}_0 - \boldsymbol{\nu}_0^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\beta} + \boldsymbol{\nu}_0^{\intercal}\boldsymbol{\kappa}_0\boldsymbol{\nu}_0\right)\right],
\end{aligned}
$$

where the second line follows by substitution of the likelihood and prior from the previous slide using the multivariate normal density from slide 11. The third line follows by distributing the inner products. Collecting terms gives

$$
p\left(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^{\intercal}\boldsymbol{\kappa}_n\boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal}\left(\boldsymbol{\kappa}_0\nu_0 + \tau\mathbf{X}^{\intercal}\mathbf{y}\right) - \left(\boldsymbol{\kappa}_0\nu_0 + \tau\mathbf{X}^{\intercal}\mathbf{y}\right)^{\intercal}\boldsymbol{\beta}\right)\right],
$$

where we have defined $\boldsymbol{\kappa}_n = \left(\boldsymbol{\kappa}_0 + \tau\mathbf{X}^{\intercal}\mathbf{X}\right)$. This term looks just like the precision matrix of a multivariate normal distribution. On the next slide, we consider the linear terms in $\boldsymbol{\beta}$.

Speaker notes

No notes on this slide.

CONDITIONAL POSTERIOR FOR REGRESSION COEFFICIENTS β (2 / 2)

Without changing the result, we insert $\boldsymbol{\kappa}_n \boldsymbol{\kappa}_n^{-1} = \mathbf{I}$ between $\boldsymbol{\beta}$ and $(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y})$ to get

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau) &\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\beta} - \boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\kappa}_n^{-1}(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y}) - (\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y})^\intercal \boldsymbol{\kappa}_n^{-1} \boldsymbol{\kappa}_n \boldsymbol{\beta}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\beta} - \boldsymbol{\beta}^\intercal \boldsymbol{\kappa}_n \boldsymbol{\nu}_n - \boldsymbol{\nu}_n^\intercal \boldsymbol{\kappa}_n \boldsymbol{\beta}\right)\right],
\end{aligned}
$$

where we defined $\boldsymbol{\nu}_n = \boldsymbol{\kappa}_n^{-1}(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y})$. We can now complete the square to obtain

$$
p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\nu_n})^\intercal \boldsymbol{\kappa}_n (\boldsymbol{\beta} - \boldsymbol{\nu_n})\right]
$$

and the conditional distribution is multivariate normal:

$$
\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau \sim \text{Normal}\left((\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X})^{-1}(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y}), (\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X})^{-1}\right).
$$

Speaker notes

No notes on this slide.

CONDITIONAL POSTERIOR FOR OBSERVATION PRECISION τ

The conditional posterior for observation precision τ is

$$\begin{aligned} p\left(\tau \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\beta}\right) &= p\left(\mathbf{y} \mid \mathbf{X}, \tau, \boldsymbol{\beta}\right) p\left(\tau\right) \\ &\propto \tau^{n/2} \exp\left[-\frac{\tau}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right] \tau^{a_0 - 1} \exp\left[-b_0 \tau\right]. \end{aligned}$$

Collecting terms, we recognize the kernel of a gamma distribution with parameters

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). \end{aligned}$$

Speaker notes

No notes on this slide.

CONDITIONAL DISTRIBUTIONS

The conditional Gibbs updates are

$$\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau \sim \text{Normal}\left(\left(\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X}\right)^{-1} \left(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y}\right), \left(\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X}\right)^{-1}\right)$$

$$\tau \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\beta} \sim \text{Gamma}\left(a_0 + \frac{n}{2}, b_0 + \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\intercal \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}{2}\right).$$

Speaker notes

- We have now arrived at the conditional posterior distributions we need to sample from the full posterior using a Gibbs sampler.

LIMITING CASES (1 / 2)

- For large prior precision $\boldsymbol{\kappa}_0$, we recover our prior best guess at the regression coefficients:

$$\begin{aligned}\lim_{\boldsymbol{\kappa}_0 \to \infty} \mathbb{E}\left[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right] &= \lim_{\boldsymbol{\kappa}_0 \to \infty} \left(\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X}\right)^{-1} \left(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y}\right) \\ &= \lim_{\boldsymbol{\kappa}_0 \to \infty} \boldsymbol{\kappa}_0^{-1} \boldsymbol{\kappa}_0 \boldsymbol{\nu}_0 \\ &= \boldsymbol{\nu}_0.\end{aligned}$$

- For large observation precision τ, we recover the maximum likelihood estimate:

$$\begin{aligned}\lim_{\tau \to \infty} \mathbb{E}\left[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \tau\right] &= \lim_{\tau \to \infty} \left(\boldsymbol{\kappa}_0 + \tau \mathbf{X}^\intercal \mathbf{X}\right)^{-1} \left(\boldsymbol{\kappa}_0 \nu_0 + \tau \mathbf{X}^\intercal \mathbf{y}\right) \\ &= \lim_{\tau \to \infty} \tau^{-1} \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \tau \mathbf{X}^\intercal \mathbf{y} \\ &= \left(\mathbf{X}^\intercal \mathbf{X}\right)^{-1} \mathbf{X}^\intercal \mathbf{y},\end{aligned}$$

  where the second equality follows because $\left(a\mathbf{B}\right)^{-1} = a^{-1}\mathbf{B}^{-1}$ and the third because scalars commute with inner products.

Speaker notes

- We consider limiting cases as sanity checks for the derivation of the conditional distributions.

LIMITING CASES (2 / 2)

For the limit $n \to \infty$, we need to rearrange the expression for ν_n slightly because it does not explicitly depend on n:

$$\boldsymbol{\nu}_n = \left(\boldsymbol{\kappa}_0 + \tau n \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_j^\intercal \right] \right)^{-1} \left(\boldsymbol{\kappa}_0 \nu_0 + \tau n \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i \right] \right).$$

In the limit, the expressions in brackets converge to expectations under an infinite population:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_j^\intercal = \mathbb{E}\left[\mathbf{x}\mathbf{x}^\intercal\right]$$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i = \mathbb{E}\left[\mathbf{x}y\right]$$

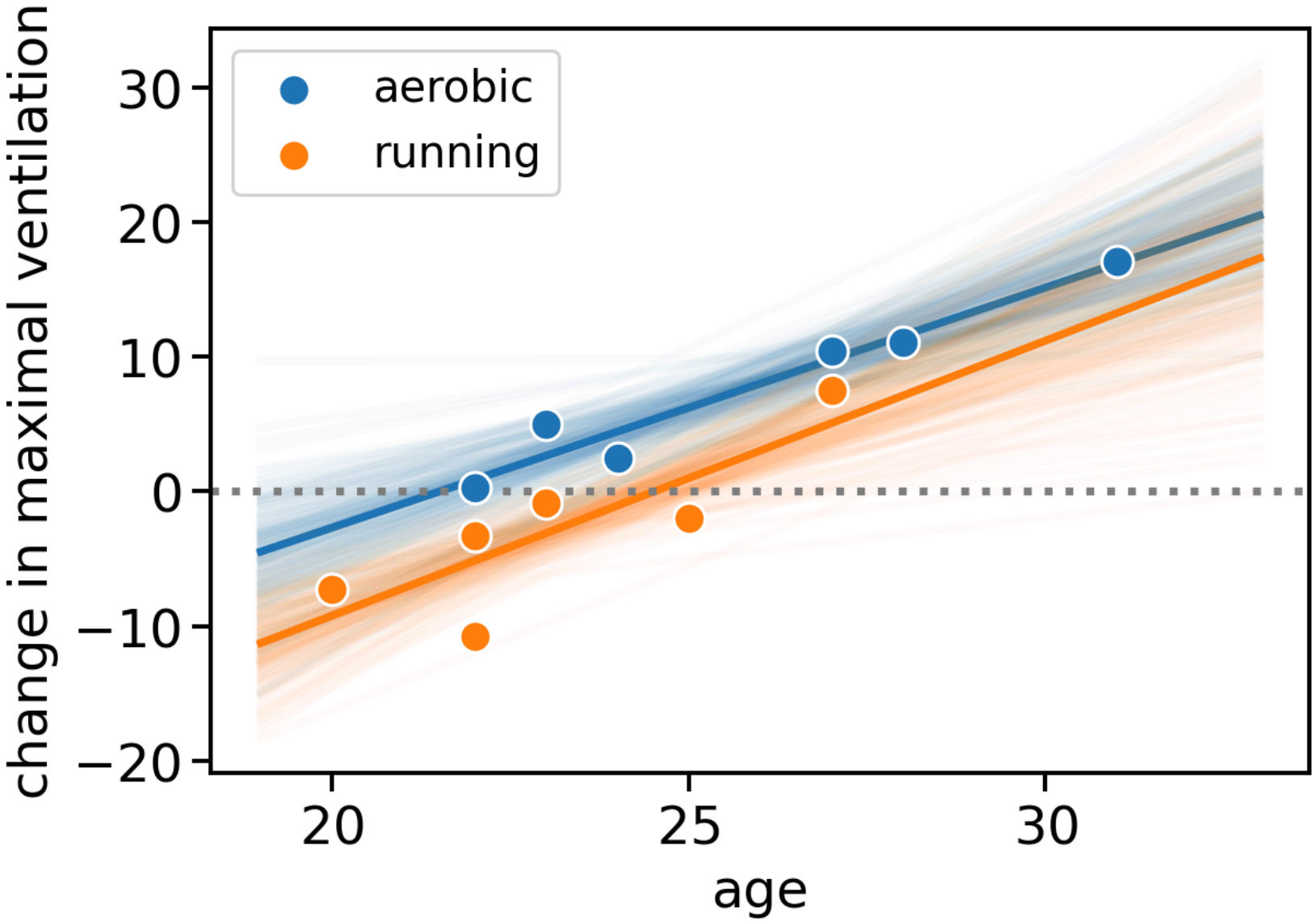
Substituting yields

$$\begin{aligned}
\lim_{n \to \infty} \boldsymbol{\nu}_n &= \lim_{n \to \infty} \left(\boldsymbol{\kappa}_0 + \tau n \mathbb{E}\left[\mathbf{x}\mathbf{x}\right] \right)^{-1} \left(\boldsymbol{\kappa}_0 \nu_0 + \tau n \mathbb{E}\left[\mathbf{x}y\right] \right) \\
&= \lim_{n \to \infty} \left(\tau n \right)^{-1} \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}\right] \right)^{-1} \tau n \mathbb{E}\left[\mathbf{x}y\right] \\
&= \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}\right] \right)^{-1} \mathbb{E}\left[\mathbf{x}y\right].
\end{aligned}$$

The second and third equalities follow the same argument as for the limit $\tau \to \infty$ on the previous slide.

Speaker notes

- In a Bayesian setting, we rarely think about infinite populations except for limiting cases such as this one.



Speaker notes

- Semi-transparent lines are posterior samples of $\mathbf{X}\boldsymbol{\beta}$ consistent with the observed data.
- We have two lines for the two different treatment regimes.
- Solid lines are posterior means averaged over all samples, i.e., the solid blue line is the average of all semi-transparent blue lines.
- Uncertainties reflect our intuition, e.g., the response for participants of the running regime with large age is poorly constrained.
- Here we capture uncertainties *in the predictor* $\mathbf{X}\boldsymbol{\beta}$. Variance would differ for the posterior predictive distribution which we consider in a future lecture.