Markov Chain Monte Carlo and Couplings
Part IV

Pierre E. Jacob

Optimization-Conscious Econometrics Summer School

June 3-6th, 2024

Outline

Outline

Outline

Outline

Outline

Outline

How to design appropriate coupled chains?

To implement the methods of Part III,

we need to design a Markov transition kernel $\bar{P}$

such that, when (X_t, Y_t) is sampled from $\bar{P}\left((X_{t-1}, Y_{t-1}), \cdot\right)$,

- marginally $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$, and $Y_t | Y_{t-1} \sim P(Y_{t-1}, \cdot)$,
- it is possible that $X_t = Y_t$ exactly for some $t \geq 0$,
- if $X_{t-1} = Y_{t-1}$, then $X_t = Y_t$ almost surely.

How do we do that for our favorite MCMC algorithms?

Coupling of probability distributions

Taking a step back, from Markov chains to static distributions. Let p and q be two probability distributions on $\mathbb{X}$.

A joint distribution Γ on $\mathbb{X} \times \mathbb{X}$ is a coupling of p and q if its first marginal is p and its second marginal is q.

$$\int \Gamma(x, dy) = p(x), \quad \int \Gamma(dx, y) = q(y).$$

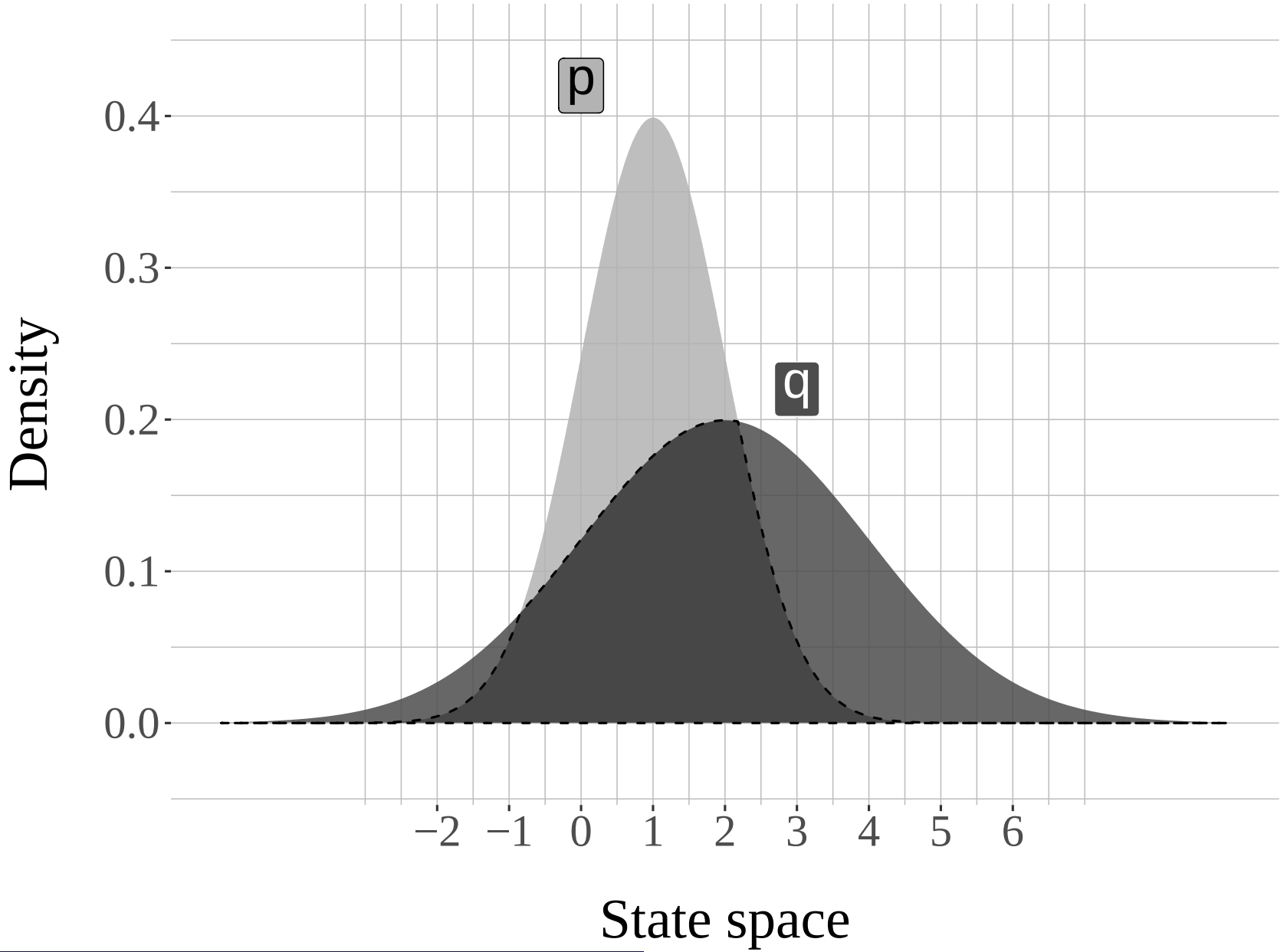
Total variation distance

For two distributions p and q,

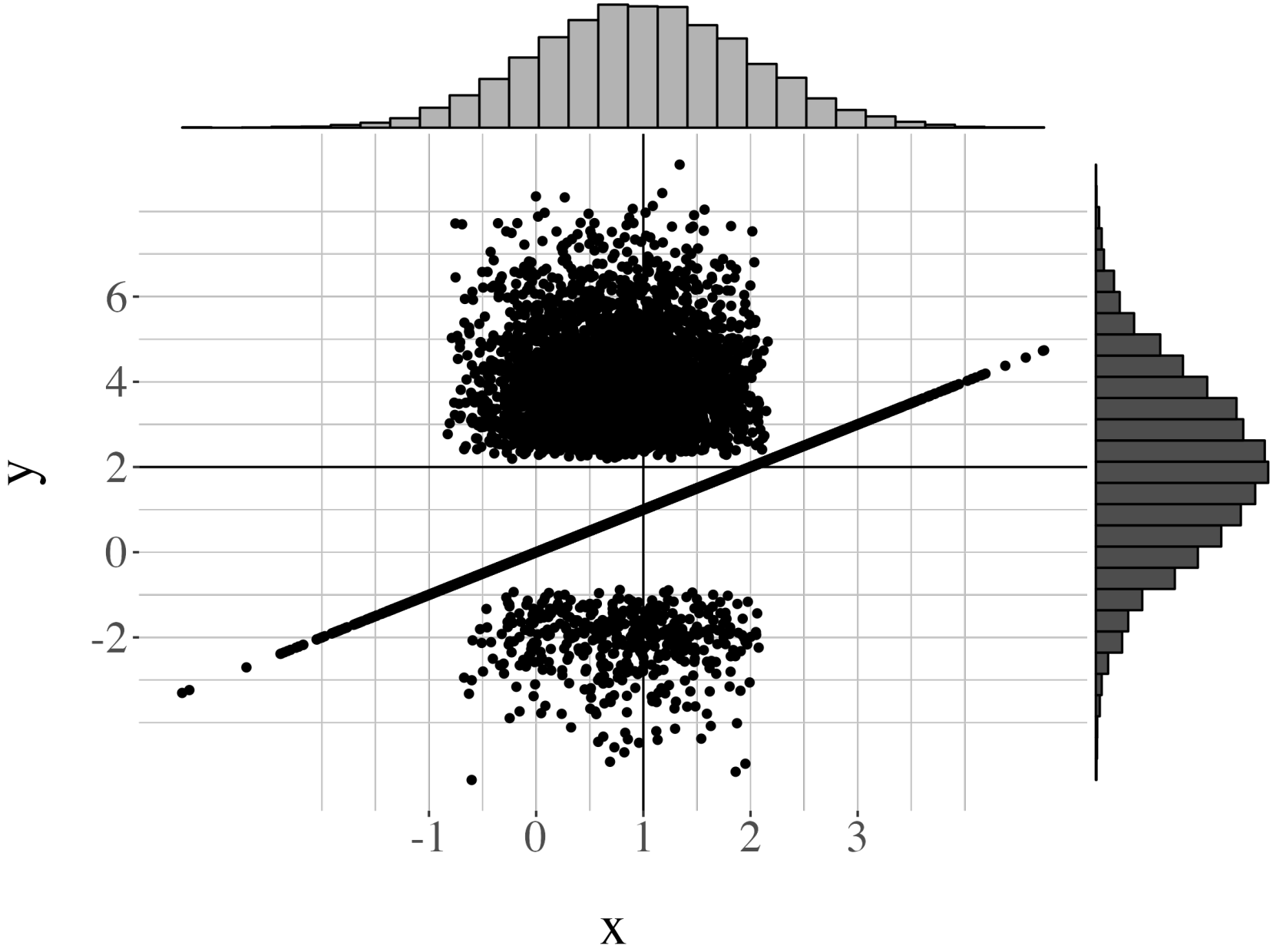
$$\begin{aligned} |p - q|_{\mathrm{TV}} &= \sup_{A \in \mathcal{X}} \; p(A) - q(A) \\ &= 1 - \int \min(p(x), q(x)) dx \\ &= \inf_{(X,Y) \in \text{coupling}(p,q)} \mathbb{P}(X \neq Y). \end{aligned}$$

The infimum is attained by "maximal" couplings (not unique).

Overlap



A maximal coupling



A maximal coupling: algorithm

Requires: evaluations of p and q, sampling from p and q.

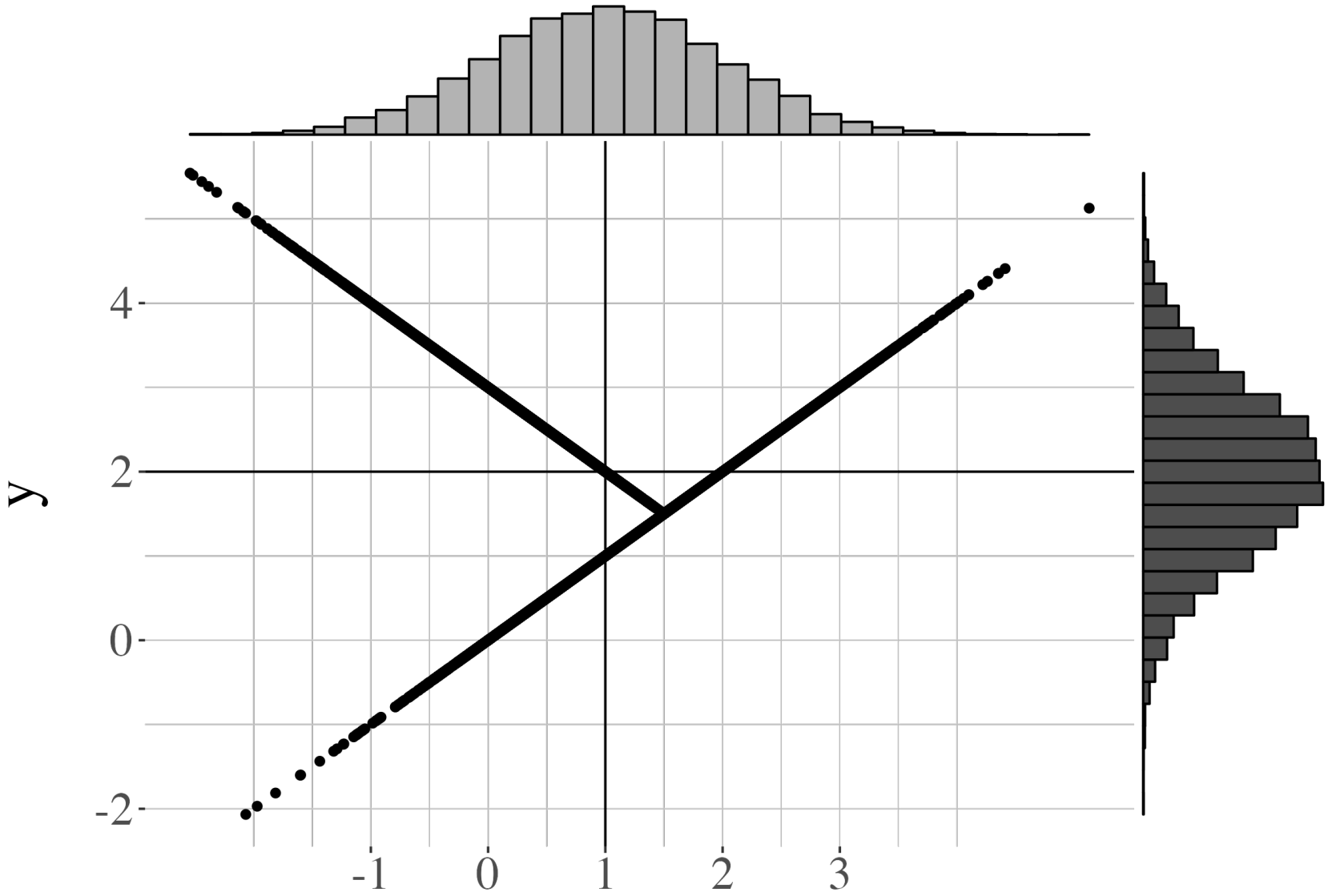
1. Sample $X \sim p$ and $W \sim \text{Uniform}(0, 1)$.
   If $W \leq q(X)/p(X)$, set $Y = X$, output (X, Y).

2. Otherwise, sample $Y^\star \sim q$ and $W^\star \sim \text{Uniform}(0, 1)$
   until $W^\star > p(Y^\star)/q(Y^\star)$, set $Y = Y^\star$ and output (X, Y).

Output: a pair (X, Y) such that $X \sim p$, $Y \sim q$
and $\mathbb{P}(X = Y)$ is maximal.

Another maximal coupling

For two Normals with the same variance.



Another maximal coupling: algorithm

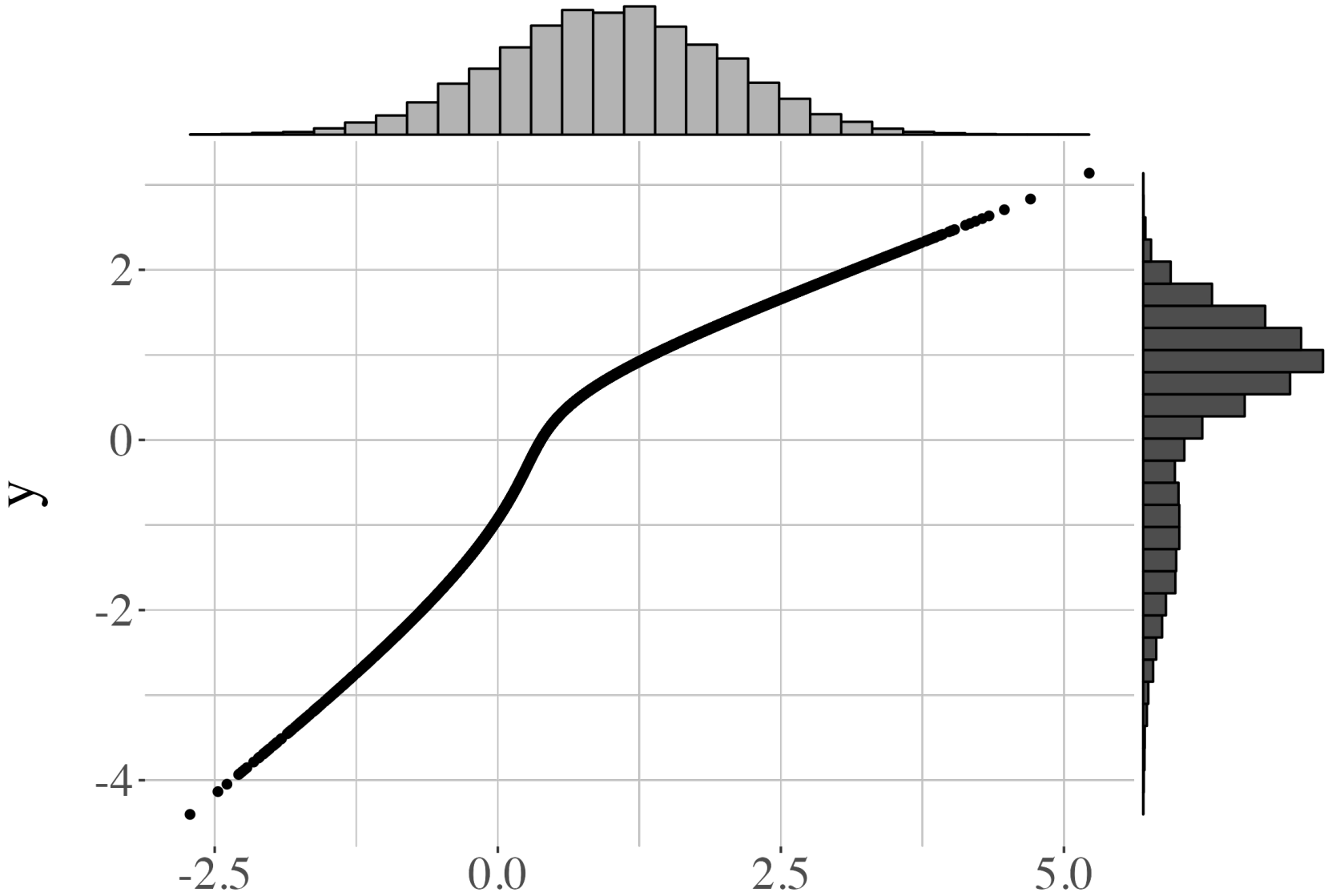
A coupling of Normal(μ_1, Σ) and Normal(μ_2, Σ).

1. Let $z = \Sigma^{-1/2}(\mu_1 - \mu_2)$ and $e = z/|z|$.
2. Sample $\dot{X} \sim \text{Normal}(0_d, I_d)$, and $W \sim \text{Uniform}(0, 1)$.
3. If $\varphi(\dot{X})W \leq \varphi(\dot{X} + z)$, set $\dot{Y} = \dot{X} + z$; else set $\dot{Y} = \dot{X} - 2(e^T \dot{X})e$.
4. Set $X = \Sigma^{1/2}\dot{X} + \mu_1, Y = \Sigma^{1/2}\dot{Y} + \mu_2$, and return (X, Y).

Bou-Rabee, Eberle & Zimmer, 2020, *Coupling and convergence for Hamiltonian Monte Carlo.*

An optimal transport coupling

Between a Normal and a mixture of Normals.



Outline

Example: probit regression

Albert & Chib, 1993, *Bayesian Analysis of Binary and Polychotomous Response Data.*

Probit model: $Y_i = \mathbb{1}\{Z_i \geq 0\}$, $Z_i \sim \text{Normal}(X_i^T \beta, 1)$.

Prior: $\beta \sim \text{Normal}(0, \Sigma)$.

Define joint distribution $\pi(\beta, Z | X, Y)$.

Data with $n = 12975$ rows and $p = 10$ columns, from Ramírez Hassan, Cardona Jiménez & Cadavid Montoya, 2013, *The Impact of Subsidized Health Insurance on the Poor in Colombia: Evaluating the Case of Medellín.*

Example: probit regression

Each Z_i has conditional distribution

$$\pi(Z_i \mid X_i, Y_i, \beta) = \begin{cases} \text{Normal}_+(X_i^T\beta, 1) \text{ if } Y_i = 1, \\ \text{Normal}_-(X_i^T\beta, 1) \text{ if } Y_i = 0, \end{cases}$$

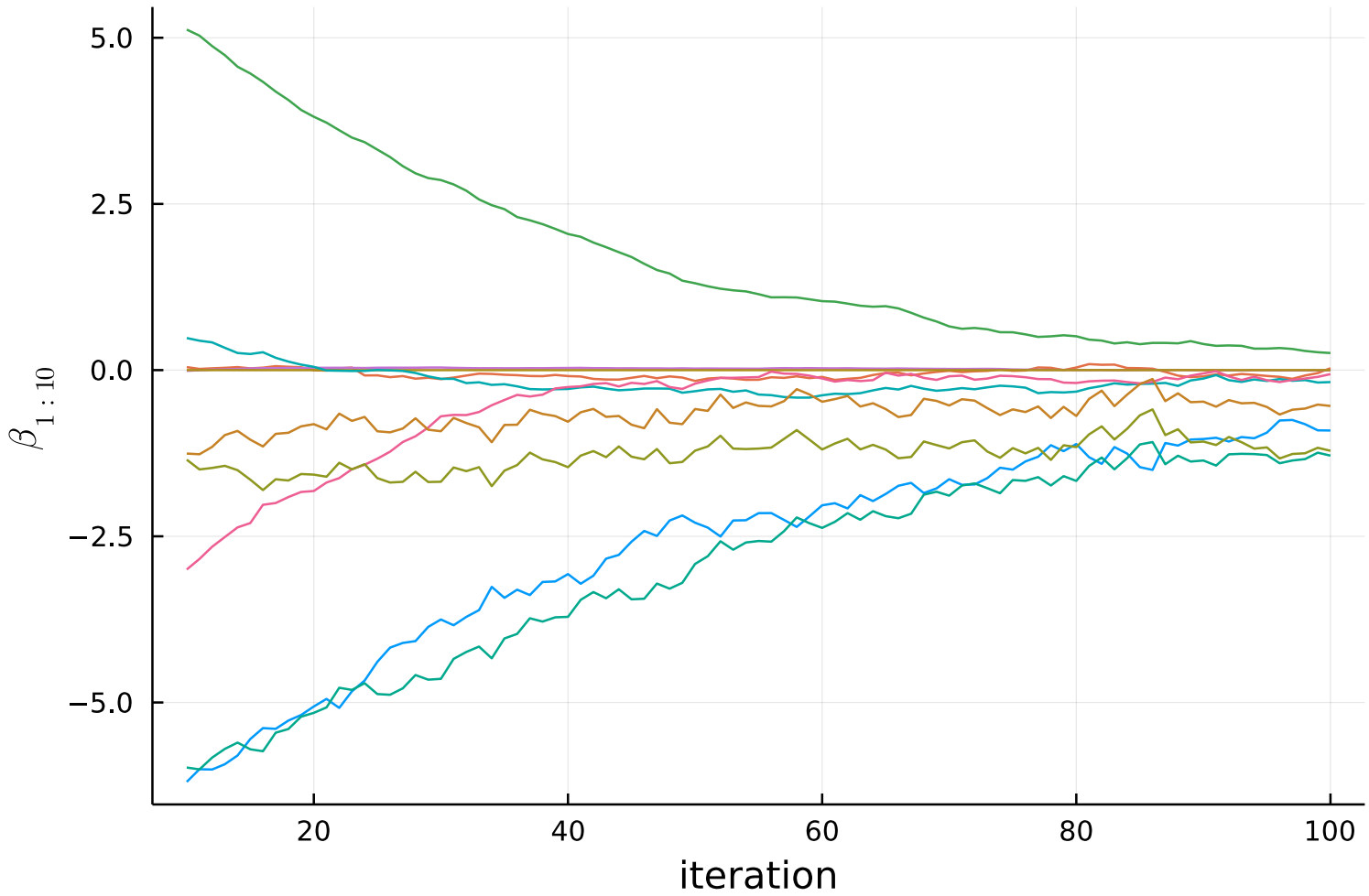
where Normal_+ refers to truncation on the left of 0, and Normal_- to truncation on the right of 0.

For β given X, Y, Z,

$$\pi(\beta \mid X, Y, Z) = \text{Normal}\left((\Sigma^{-1} + X^TX)^{-1}X^TZ, (\Sigma^{-1} + X^TX)^{-1}\right),$$

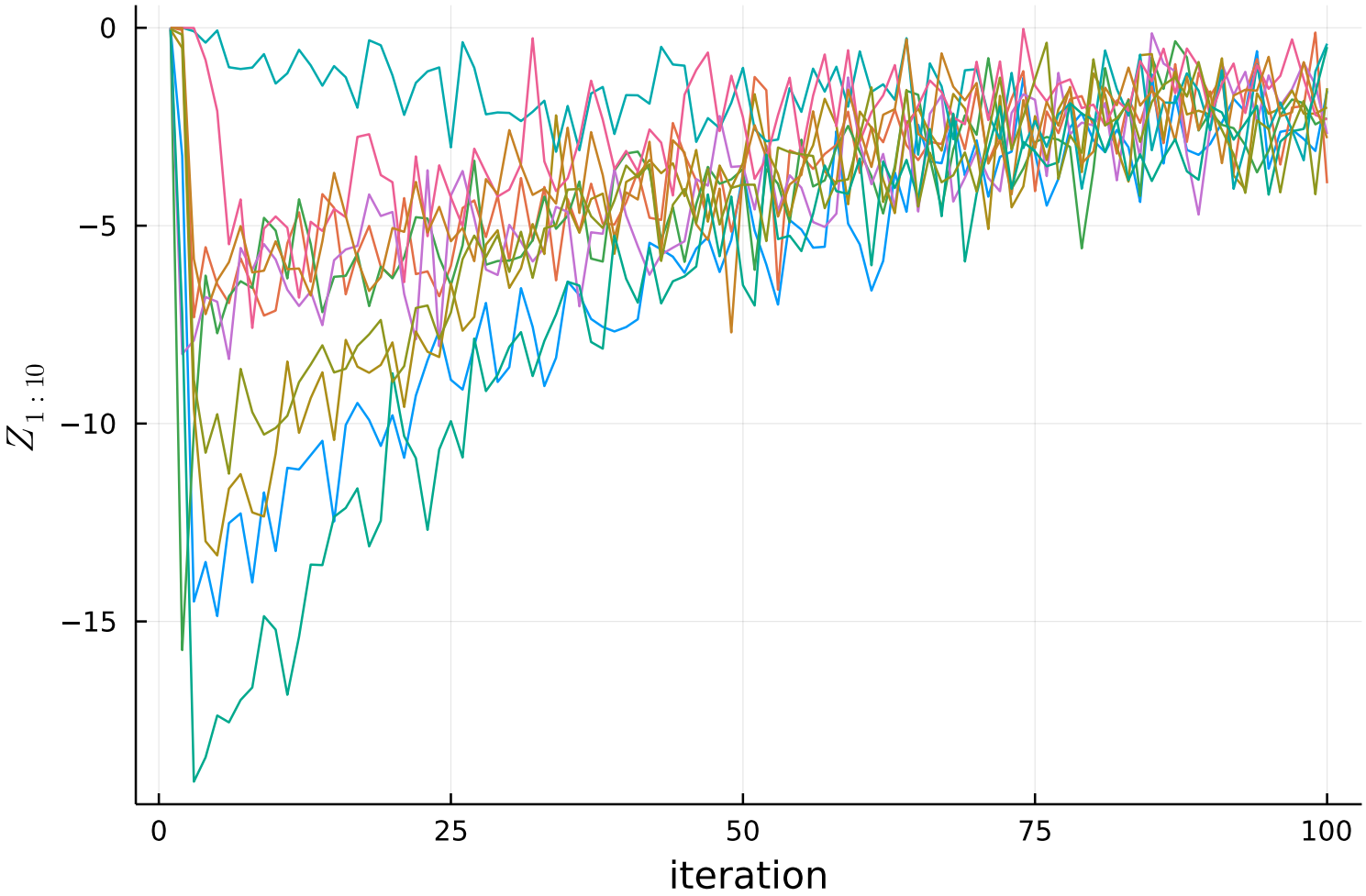
using the conjugacy of Normal distributions.

Example: probit regression



How long should we run this?

Example: probit regression



How long should we run this?

Example: probit regression

Recall: state $(\beta_1, \ldots, \beta_p, Z_1, \ldots, Z_n)$.

The components $(\beta_1, \ldots, \beta_p)$ are drawn from a multivariate Normal distribution which mean depends on $(Z_1, \ldots, Z_n)$.

The components Z_i are drawn one by one from a truncated Normal distribution which mean depends on $(\beta_1, \ldots, \beta_p)$.

Consider two states:

$$(\beta_1, \ldots, \beta_p, Z_1, \ldots, Z_n) \quad \text{and} \quad (\tilde{\beta}_1, \ldots, \tilde{\beta}_p, \tilde{Z}_1, \ldots, \tilde{Z}_n).$$

How to couple them?

Example: probit regression

For β given X, Y, Z,

$$\pi(\beta \mid X, Y, Z) = \text{Normal}\left((B^{-1} + X^T X)^{-1} X^T Z, (B^{-1} + X^T X)^{-1}\right).$$

For $\tilde{\beta}$ given $X, Y, \tilde{Z}$,

$$\pi(\tilde{\beta} \mid X, Y, \tilde{Z}) = \text{Normal}\left((B^{-1} + X^T X)^{-1} X^T \tilde{Z}, (B^{-1} + X^T X)^{-1}\right).$$

How to couple them?

Example: probit regression

Each Z_i has conditional distribution

$$\pi(Z_i \mid X_i, Y_i, \beta) = \begin{cases} \text{Normal}_+(X_i^T\beta, 1) \text{ if } Y_i = 1, \\ \text{Normal}_-(X_i^T\beta, 1) \text{ if } Y_i = 0. \end{cases}$$

Each $\tilde{Z}_i$ has conditional distribution

$$\pi(\tilde{Z}_i \mid X_i, Y_i, \tilde{\beta}) = \begin{cases} \text{Normal}_+(X_i^T\tilde{\beta}, 1) \text{ if } Y_i = 1, \\ \text{Normal}_-(X_i^T\tilde{\beta}, 1) \text{ if } Y_i = 0. \end{cases}$$

How to couple them?

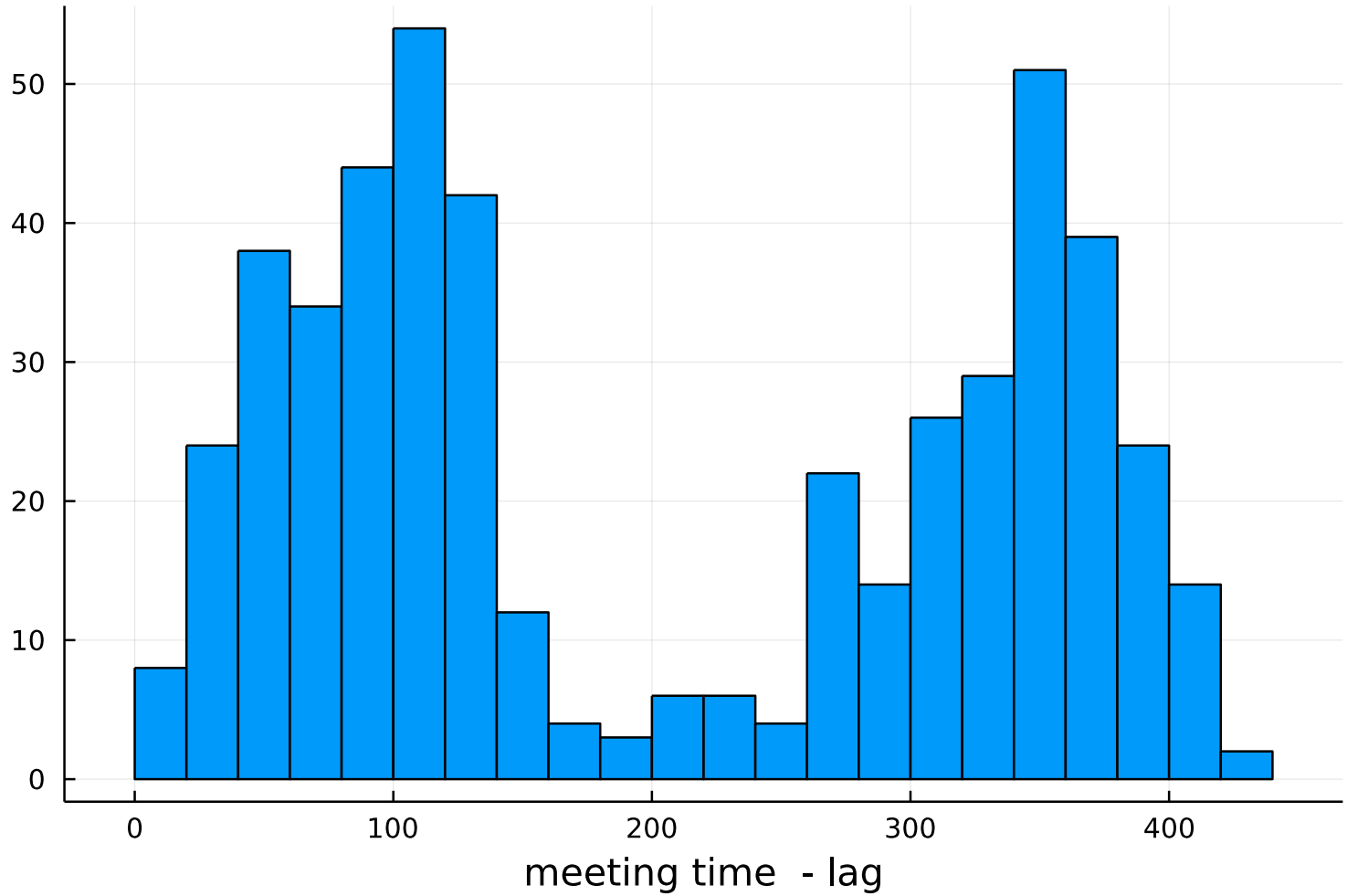
Example: probit regression

Many possible strategies, but here

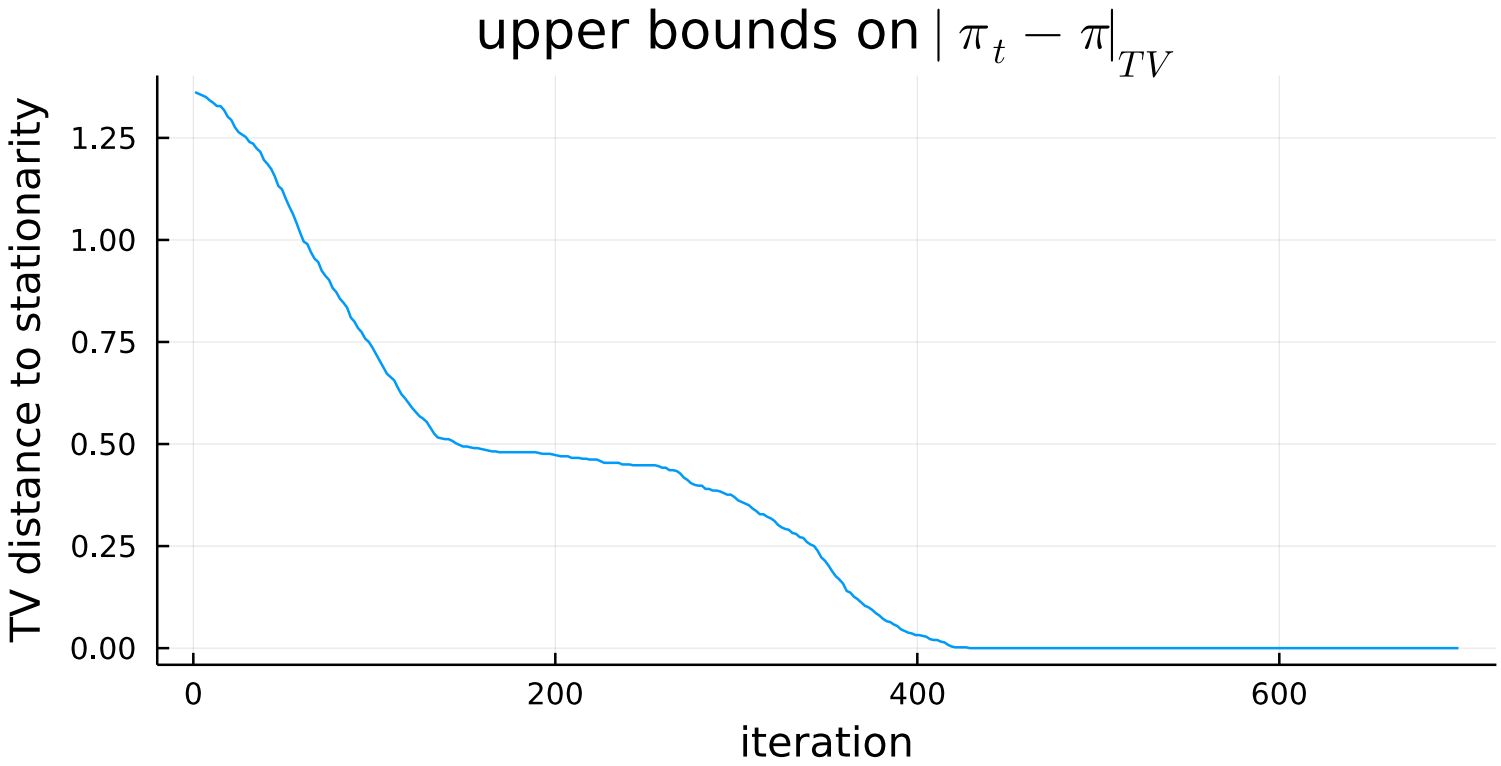
- common uniforms and inverse CDF transforms for Z_i, $\tilde{Z}_i$,
- reflection-maximal couplings for $\beta, \tilde{\beta}$,

seems to result in good performance.

Example: probit regression



Example: probit regression



Maybe the chain really converges in 100 steps, maybe not.
This plot suggests that it converges in less than ~ 400 steps.

Example: Bayesian quantile regression

Quantile regression:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta), \quad \rho_\tau : u \mapsto u(\tau - \mathbb{1}(u < 0)).$$

Asymmetric Laplace distribution:

$$f(u|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\rho_\tau\left(\frac{u-\mu}{\sigma}\right)\right).$$

Posterior distribution:

$$\pi(\beta, \sigma) \propto \text{prior}(\beta, \sigma) \prod_{i=1}^{n} f(y_i|x_i'\beta, \sigma, \tau).$$

Normal prior(m_0, V_0) on β, InverseGamma(n_0, s_0) prior on σ. `data("CPSSWEducation")` from Stock and Watson (2007), 2,950 observations, earnings vs education and gender.

Example: Bayesian quantile regression

Gibbs sampler of Kozumi & Kobayashi (2011), package `bayesQR` of Benoit & Van den Poel (2017).

Asymmetric Laplace distribution can be represented as follows:

$$y_i = x_i'\beta + \epsilon_i, \quad \text{with} \quad \epsilon_i = \theta\nu_i + \omega\sqrt{\sigma\nu_i}u_i,$$

where $\nu_i = \sigma z_i$, z_i is Exponential(1), $\omega = (1-2\tau)/\tau(1-\tau)$, $\omega^2 = 2/\tau(1-\tau)$, u_i is Normal$(0,1)$.

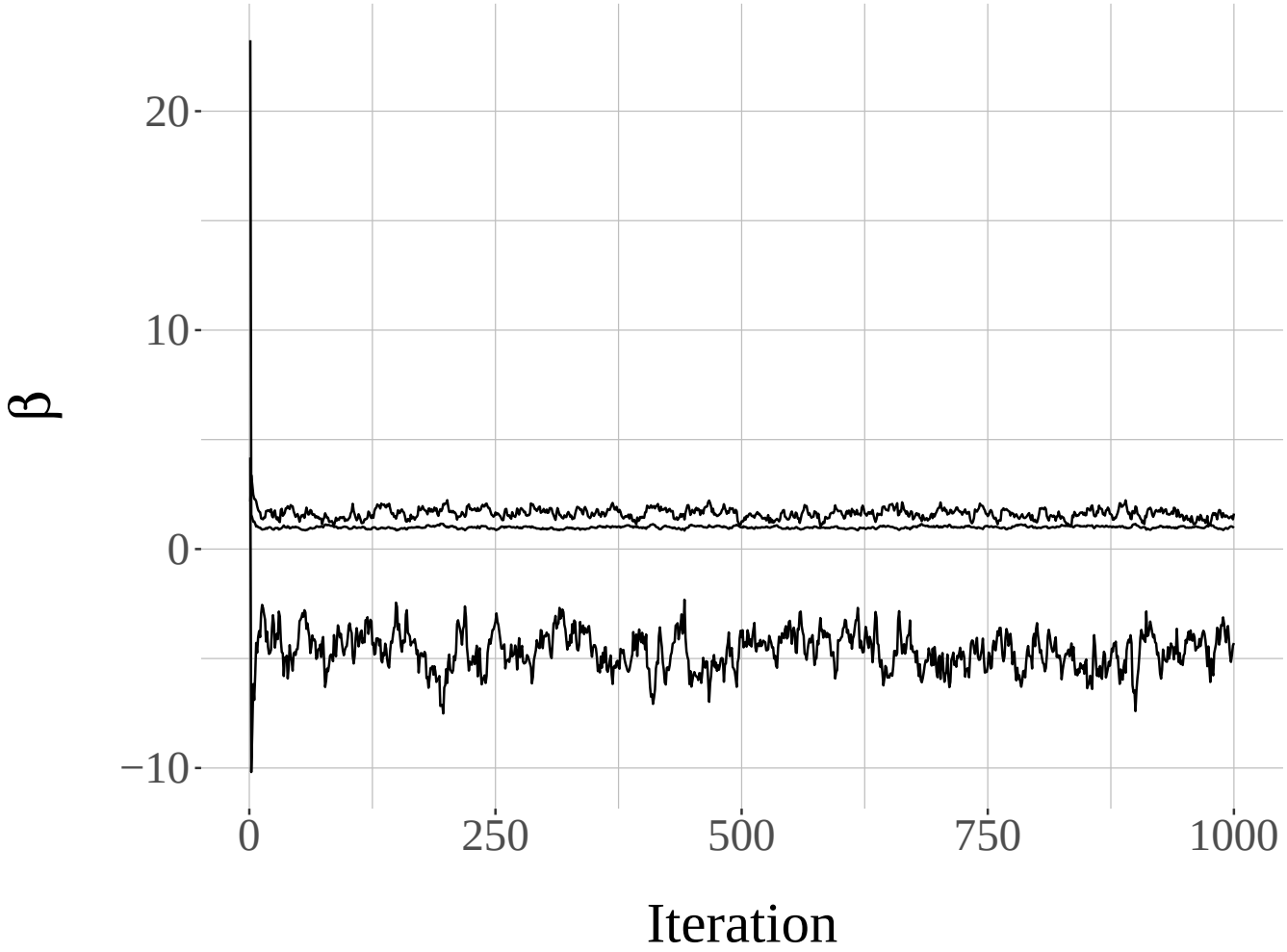
This leads to a sampler alternating updates of β, σ and (ν_i).

Example: Bayesian quantile regression

- Update of β: Normal$(m_1(\sigma, \nu), V_1(\sigma, \nu))$.
- Update of σ: InverseGamma$(n_1, s_1(\beta, \nu))$.
- Update of each ν_i: GeneralizedInverseGaussian$(1/2, \delta(\beta, \sigma), \gamma(\sigma))$.

In our example, $\beta \in \mathbb{R}^3$, $\sigma \in \mathbb{R}_+$, $\nu \in \mathbb{R}^{2950}$.

Example: Bayesian quantile regression



How long should we run this?

Example: Bayesian quantile regression

- Update of β: $\text{Normal}(m_1(\sigma, \nu), V_1(\sigma, \nu))$.
- Update of σ: $\text{InverseGamma}(n_1, s_1(\beta, \nu))$.
- Update of each ν_i: $\text{GeneralizedInverseGaussian}(1/2, \delta(\beta, \sigma), \gamma(\sigma))$.

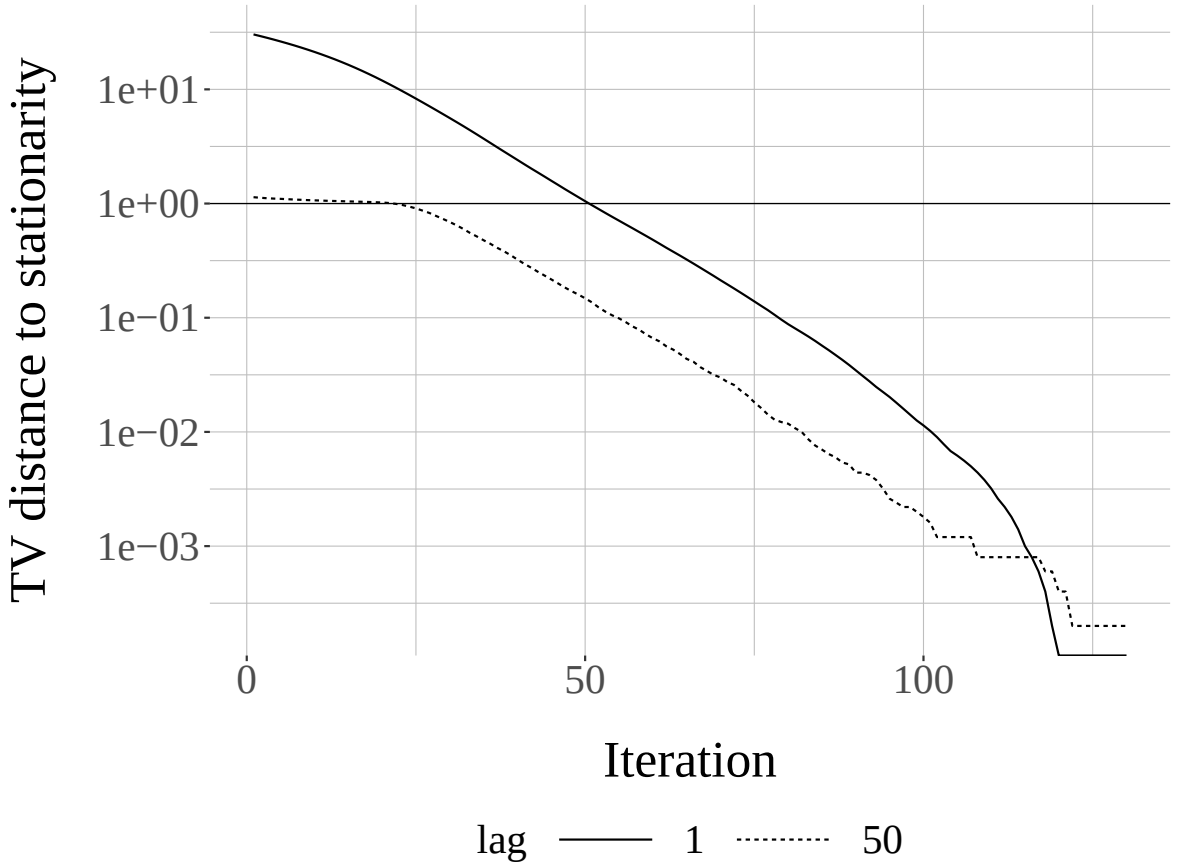
How to couple this Markov transition?

Example: Bayesian quantile regression

- Update of β: Normal$(m_1(\sigma, \nu), V_1(\sigma, \nu))$.
- Update of σ: InverseGamma$(n_1, s_1(\beta, \nu))$.
- Update of each ν_i:
  GeneralizedInverseGaussian$(1/2, \delta(\beta, \sigma), \gamma(\sigma))$.

How to couple this Markov transition?
Let's try: common random numbers on β, maximal coupling on σ and on each ν_i.

Example: Bayesian quantile regression



Maybe ~ 100 steps is enough? Confidence bands?

Outline

Metropolis–Rosenbluth–Teller–Hastings (kernel P)

At each iteration t, Markov chain at state X_t,

1. propose $X^\star \sim q(X_t, \cdot)$,
2. sample $U \sim \text{Uniform}(0, 1)$,
3. if
$$U \leq \frac{\pi(X^\star) q(X^\star, X_t)}{\pi(X_t) q(X_t, X^\star)},$$
set $X_{t+1} = X^\star$, otherwise set $X_{t+1} = X_t$.

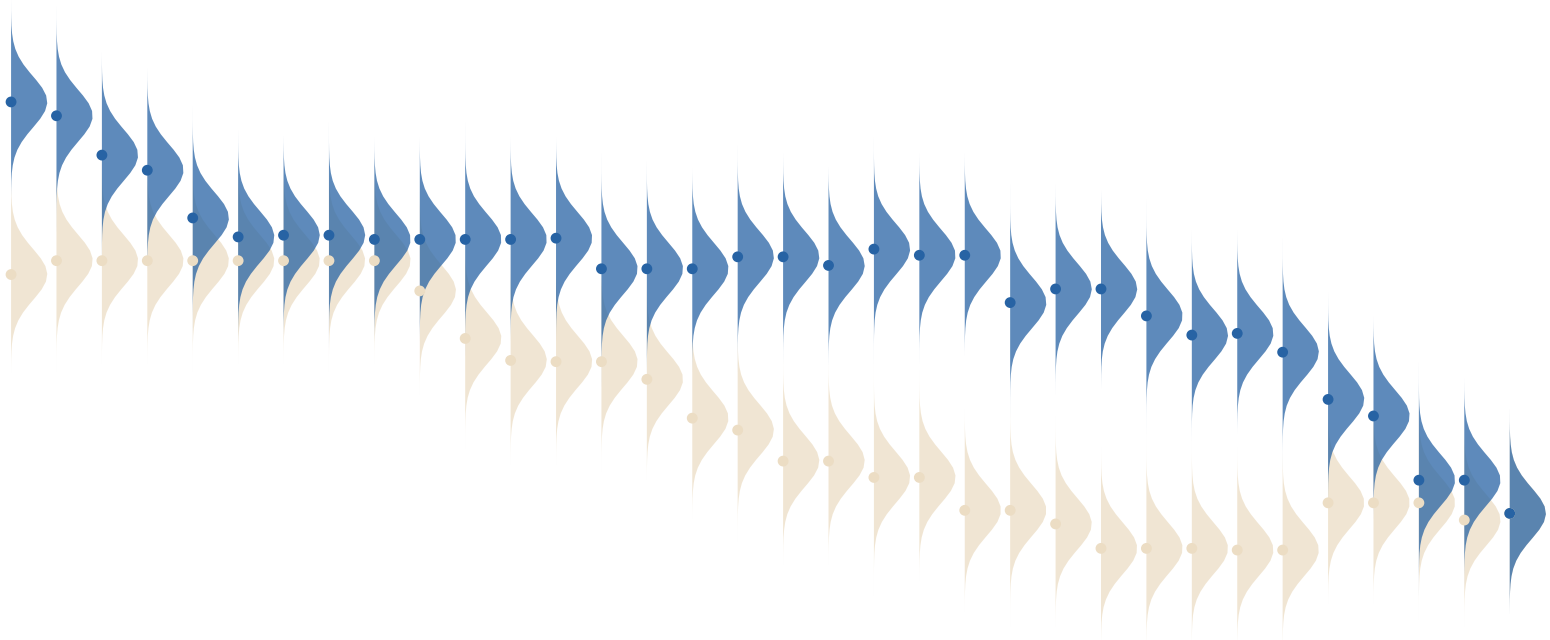
How to propagate two MRTH chains from states X_t and Y_t such that $\{X_{t+1} = Y_{t+1}\}$ can happen?

Metropolis–Rosenbluth–Teller–Hastings (kernel $\bar{P}$)

At each iteration t, two Markov chains at states X_t, Y_t,

1. propose $(X^\star, Y^\star)$ from max coupling of $q(X_t, \cdot)$, $q(Y_t, \cdot)$,
2. sample $U \sim \text{Uniform}(0, 1)$,
3. if
$$U \leq \frac{\pi(X^\star) q(X^\star, X_t)}{\pi(X_t) q(X_t, X^\star)},$$
set $X_{t+1} = X^\star$, otherwise set $X_{t+1} = X_t$,
if
$$U \leq \frac{\pi(Y^\star) q(Y^\star, Y_t)}{\pi(Y_t) q(Y_t, Y^\star)},$$
set $Y_{t+1} = Y^\star$, otherwise set $Y_{t+1} = Y_t$.

Proposal distributions of two MRTH chains



Metropolis–Rosenbluth–Teller–Hastings

Maximal coupling of proposals + common uniform for acceptance: does not result in a maximal coupling of the MRTH transition kernels.

Wang, O'Leary, Jacob, 2021, *Maximal Couplings of the Metropolis–Hastings Algorithm.*

Reflection-maximal couplings of Normal proposals often result in good performance, but better performance is obtained with *gradient common random number* couplings.

Papp & Sherlock, 2022, *A new and asymptotically optimally contracting coupling for the random walk Metropolis.*

Contraction before meeting

If X_t and Y_t are far, $|q(X_t, \cdot) - q(Y_t, \cdot)|_{\mathrm{TV}}$ is close to one, so $\{X^\star = Y^\star\}$ can only occur with small probability.

It can be useful to implement two-scale strategies:

- If $|X - Y| >$ threshold, propagate in a contractive way.
- If $|X - Y| <$ threshold, attempt to obtain a meeting.

Employed in Biswas, Bhattacharya, Jacob & Johndrow, 2022, *Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors*

Contractive coupling

No general recipe, but some useful devices:

- Common random numbers.

  Example: same uniform in $F_X^{-1}(U)$ and $F_Y^{-1}(U)$.

  Glasserman & Yao, 1992, *Some guidelines and guarantees for common random numbers.*

- Reflection couplings.

  Example: Brownian motion.

  Eberle, 2011, *Reflection coupling and Wasserstein contractivity without convexity.*

Concluding remarks

Monte Carlo methods form a toolbox to approximate probabilities, integrals, expectations, volumes, etc.

Useful for many tasks, not just for Bayesian inference.

Fruitful connections between Monte Carlo and optimization methods, and also some distinctive features.

Coupling techniques well-known as theoretical devices, but also possibly useful in practice...?

Thank you for your attention!