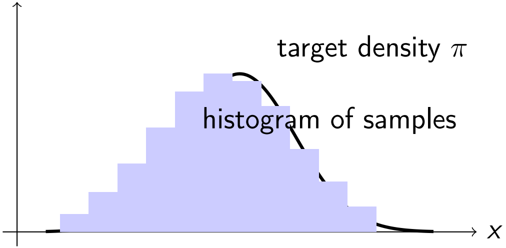
Sampling algorithms: from tossing a coin to a matrix-valued case

Yiye Jiang

Statify

What is sampling?

- **Sampling** from a distribution $\pi(x)$, $x \in \mathcal{X}$:
  generating values $x_1, x_2, \dots, x_n \in \mathcal{X}$ such that



Examples :

- **Bernoulli**$(1/2)$, $\mathcal{X} = \{0, 1\}$: toss a fair coin, we generate $\{0, 0, 1, 0, 1 \dots\}$.
- **Uniform on** $[0, 1/2]$, $\mathcal{X} = [0, 1/2]$: choose a point at random on a segment.

Why we need sampling ?

In computational statistics:

- Estimate expectations: $\int f(x)\pi(x)dx \approx \sum_{i=1}^{n} f(x_i)$, and other quantities of π whose explicit formulas are not possible.

Why we need sampling ?

In computational statistics:

- Estimate expectations: $\int f(x)\pi(x)dx \approx \sum_{i=1}^{n} f(x_i)$, and other quantities of π whose explicit formulas are not possible.
- In Bayesian inference $\rightarrow$ *our context*, we proposed a new bayesien model for inferring covariance matrix $\mathbf{\Sigma}$:

$$z_i, \dots, z_n | \mathbf{\Sigma} \overset{iid}{\sim} \mathcal{N}(0, \mathbf{\Sigma})$$
$$\mathbf{\Sigma} \sim \mathcal{SIW}(\Psi, \nu),$$

Bayesian inference of $\mathbf{\Sigma}$ requires to derive:

- $\pi(\Sigma) := \Sigma | z_1, \dots, z_n$, $\rightarrow$ Explicit formula exists $\pi(\Sigma) = \mathcal{SIW}(\Psi + \sum z_i z_i^\top, \nu + n/2)$.
- $\mathbb{E}_\pi(f(\Sigma)) := \int f(\Sigma)\pi(\Sigma)d\Sigma$. $\rightarrow$ No explicit formula $\rightarrow$ Sampling algo.

Idea 1: decomposition by independence

Uniform on $[0, 1/2]$: choose a point at random on a segment.

Uniform on square $[0, 1/2] \times [0, 1/2]$.
A sample takes 2 variables (X, Y), with $X \in [0, 1/2]$, $Y \in [0, 1/2]$.

Idea 1: decomposition by independence

Uniform on $[0, 1/2]$: choose a point at random on a segment.

Uniform on square $[0, 1/2] \times [0, 1/2]$.
A sample takes 2 variables (X, Y), with $X \in [0, 1/2]$, $Y \in [0, 1/2]$.
Note that

$$\pi(X, Y) = 4 = 2 \times 2 = \pi(X)\pi(Y)$$

$\iff$ X is independent of Y, then we can draw

$$X \sim \mathrm{Unif}(0, 1/2), \qquad Y \sim \mathrm{Unif}(0, 1/2),$$

and return the pair (X, Y).

Idea 1: decomposition by independence

Suppose that we want to sample a pair (X, Y) with density

$$\pi(x, y) = f(x)g(y).$$

Then:

- first sample $X \sim f$,
- then independently sample $Y \sim g$.

Conclusion: when a multidimensional distribution factorizes, we can sample it coordinate by coordinate.

Idea 2: decomposition by conditioning

More often (X, Y) are not independent, we can try decompose by conditioning

$$\pi(x, y) = \pi(y \mid x)\, \pi_X(x).$$

When $\pi_X(x)$ and $\pi(y \mid x)$ are easy to sample, then :

1. sample $X \sim \pi_X$,
2. then sample $Y \mid X \sim \pi(\cdot \mid X)$.

Idea 2: decomposition by conditioning

More often (X, Y) are not independent, we can try decompose by conditioning

$$\pi(x, y) = \pi(y \mid x)\, \pi_X(x).$$

When $\pi_X(x)$ and $\pi(y \mid x)$ are easy to sample, then :

1. sample $X \sim \pi_X$,
2. then sample $Y \mid X \sim \pi(\cdot \mid X)$.

Example: Sampling from density ?

$$f_{X,Y}(x, y) = 1_{[0,1]}(x)\, \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right).$$

A sample takes 2 variables (X, Y), with $X \in [0, 1]$, $Y \in \mathbb{R}$.

Idea 2: decomposition by conditioning

More often (X, Y) are not independent, we can try decompose by conditioning

$$\pi(x, y) = \pi(y \mid x)\, \pi_X(x).$$

When $\pi_X(x)$ and $\pi(y \mid x)$ are easy to sample, then :

1. sample $X \sim \pi_X$,
2. then sample $Y \mid X \sim \pi(\cdot \mid X)$.

Example: Sampling from density ?

$$f_{X,Y}(x, y) = 1_{[0,1]}(x) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right).$$

A sample takes 2 variables (X, Y), with $X \in [0, 1]$, $Y \in \mathbb{R}$.

$$X \sim \mathrm{Unif}(0, 1), \qquad Y \mid X = x \sim \mathcal{N}(x, \sigma^2).$$

The actual distribution in the work

Let $\Sigma = \Gamma\Lambda\Gamma$ be the eigendecomposition, SIW with $b = 1$ writes as:

$$\pi(\Lambda, \Gamma \mid \nu, \Psi) \; = c \; \frac{\exp\left(\operatorname{tr}\left(-\frac{1}{2}\Gamma\Lambda^{-1}\Gamma^{\top}\Psi\right)\right)}{|\Lambda|^{\nu}} 1_{O_K}(\Gamma),$$

where c is some constant, and O_K is the set of all orthonormal matrices of size K.

The actual distribution in the work

Let $\Sigma = \Gamma\Lambda\Gamma$ be the eigendecomposition, SIW with $b = 1$ writes as:

$$\pi(\Lambda, \Gamma \mid \nu, \Psi) \ = c \ \frac{\exp\left(\mathrm{tr}\left(-\frac{1}{2}\Gamma\Lambda^{-1}\Gamma^\top\Psi\right)\right)}{|\Lambda|^{\nu}} 1_{O_K}(\Gamma),$$

where c is some constant, and O_K is the set of all orthonormal matrices of size K.

Decomposition into simpler distributions.

- **By conditioning:** $\pi(\Lambda, \Gamma|\nu, \Psi) = \pi(\Lambda|\Gamma, \nu, \Psi)\ \pi(\Gamma|\nu, \Psi)$, where $\pi(\Gamma|\nu, \Psi) = 1_{O_K}(\Gamma)$, *and efficient sampling has been proposed.*
- **By indpendence:** Given Γ fixed,

$$\pi(\Lambda|\Gamma, \nu, \Psi) = c \ \frac{\exp\left(\mathrm{tr}\left(-\frac{1}{2}\Gamma\Lambda^{-1}\Gamma^\top\Psi\right)\right)}{|\Lambda|^{\nu}} = c\prod_{i=1}^{K} k_{\mathcal{IG}}(\lambda_i|\nu - 1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i),$$

which are K independent IG distributions. IG classical ID distribution, easy to sample.

Details : bias from Step 2 - correct by importance sampling

Let $\Sigma = \Gamma\Lambda\Gamma$ be the eigendecomposition, SIW with $b = 1$ writes as:

$$\pi(\Lambda, \Gamma \mid \nu, \Psi) = c\,\frac{\exp\left(\mathrm{tr}\left(-\frac{1}{2}\Gamma\Lambda^{-1}\Gamma^\top\Psi\right)\right)}{|\Lambda|^\nu} 1_{O_K}(\Gamma) = c\prod_{i=1}^{K} k_{\mathcal{IG}}(\lambda_i|\nu - 1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i) 1_{O_K}(\Gamma).$$

Two-step sampling strategy following $\pi(\Lambda, \Gamma|\nu, \Psi) = \pi(\Lambda|\Gamma, \nu, \Psi)\ \pi(\Gamma|\nu, \Psi)$:

1. $\Gamma \sim$ Unif over orthonormal matrices of $K \times K$
2. $\Lambda|\Gamma \sim \prod_{i=1}^{K} k_{\mathcal{IG}}(\lambda_i|\nu - 1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i)$

Details : bias from Step 2 - correct by importance sampling

Let $\Sigma = \Gamma\Lambda\Gamma$ be the eigendecomposition, SIW with $b = 1$ writes as:

$$\pi(\Lambda, \Gamma \mid \nu, \Psi) = c\,\frac{\exp\left(\operatorname{tr}\left(-\frac{1}{2}\Gamma\Lambda^{-1}\Gamma^\top\Psi\right)\right)}{|\Lambda|^\nu} 1_{O_K}(\Gamma) = c\prod_{i=1}^{K} k_{\mathcal{IG}}(\lambda_i|\nu - 1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i) 1_{O_K}(\Gamma).$$

Two-step sampling strategy following $\pi(\Lambda, \Gamma|\nu, \Psi) = \pi(\Lambda|\Gamma, \nu, \Psi)\ \pi(\Gamma|\nu, \Psi)$:

1. $\Gamma \sim$ Unif over orthonormal matrices of $K \times K$
2. $\Lambda|\Gamma \sim \prod_{i=1}^{K} k_{\mathcal{IG}}(\lambda_i|\nu - 1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i)$
    - $\lambda_i|\Gamma \sim \mathcal{IG}(\lambda_i|\nu - 1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i)$?

Details : bias from Step 2 - correct by importance sampling

Let $\Sigma = \Gamma\Lambda\Gamma$ be the eigendecomposition, SIW with $b = 1$ writes as:

$$\pi(\Lambda, \Gamma \mid \nu, \Psi) = c \frac{\exp\left(\mathrm{tr}\left(-\frac{1}{2}\Gamma\Lambda^{-1}\Gamma^\top\Psi\right)\right)}{|\Lambda|^\nu} 1_{O_K}(\Gamma) = c\prod_{i=1}^{K} k_{\mathcal{IG}}(\lambda_i|\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i) 1_{O_K}(\Gamma).$$

Two-step sampling strategy following $\pi(\Lambda, \Gamma|\nu, \Psi) = \pi(\Lambda|\Gamma, \nu, \Psi)\ \pi(\Gamma|\nu, \Psi)$:

1. $\Gamma \sim$ Unif over orthonormal matrices of $K \times K$
2. $\Lambda|\Gamma \sim \prod_{i=1}^K k_{\mathcal{IG}}(\lambda_i|\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i)$
   - $\lambda_i|\Gamma \sim \mathcal{IG}(\lambda_i|\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i)$? NO ! Because

$$\pi(\lambda_i|\Gamma) = c_{\mathcal{IG}}(\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i) k_{\mathcal{IG}}(\lambda_i|\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i)$$

$$\rightarrow \pi(\Lambda, \Gamma \mid \nu, \Psi) = \prod_{i=1}^{K} c_{\mathcal{IG}}(\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i) k_{\mathcal{IG}}(\lambda_i|\nu-1, \frac{1}{2}\Gamma_i^\top\Psi\Gamma_i) 1_{O_K}(\Gamma)$$

Fast

Table 4: Mean running time over 10 simulations (in seconds) of Algorithm 2.

K, ν, Ψ \ M	500	2500	4500	6500	8500	10000
(10, 4, Case 1)	0.053	0.208	0.377	0.528	0.684	0.800
(**100**, 4, Case 1)	1.494	6.512	11.522	16.540	21.576	**25.361**
(10, 20, Case 1)	0.039	0.199	0.357	0.511	0.668	0.789
(**100**, 20, Case 1)	1.456	6.426	11.418	16.392	21.353	**25.081**

Berger et al. (2020, AOS) takes 0.26 seconds for $K = 100$ to generate 1 sample.

Conclusions

In this work, we propose a fast sampling algorithm for the Shrinkage Inverse Wishart distribution (SIW). This algorithm facilitates posterior inference in Bayesian models based on the SIW prior by substantially **reducing the computational cost**, and therefore makes **higher-dimensional problems** more tractable. Since SIW is a flexible prior for covariance matrices and is conjugate to the Gaussian likelihood, the proposed algorithm **broadens its practical applicability** and encourages its use in a wider range of Bayesian models.

Jiang, Y. (2025). New sampling approaches for Shrinkage Inverse-Wishart distribution. arXiv preprint arXiv:2511.03044. Revision, *Statistics and computing.*