Polya-Gamma Verteilung und Data Augmentation

Anwendung in der Erdbeben-Vorhersage

Toni Luhdo

March 2, 2025

Inhalt

Einleitung

- Motivation: Warum Erdbeben-Vorhersage?
- Herausforderung: Komplexität und Unsicherheit in seismischen Daten
- Ziel: Einsatz der Polya-Gamma Verteilung zur Datenaugmentation

Was ist die Polya-Gamma Verteilung?

- Einführung durch Polson, Scott und Windle (2013)
- Wird verwendet zur Vereinfachung von logistischen Modellen
- Eigenschaften:
  - Symmetrische Verteilung
  - Verwendbar in Gibbs-Sampling Algorithmen

Mathematische Definition

Eine Zufallsvariable X folgt einer Polya-Gamma Verteilung $PG(b, c)$, wenn ihre Dichte gegeben ist durch:

$$X \sim PG(b, c)$$

$$X = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{G_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$

wobei G_k unabhängige Gamma-Verteilungen sind.

Herleitung der Polya-Gamma Verteilung

- Ursprünglich entwickelt zur Behandlung von logistischen Likelihoods
- Verwendung der Identität:

$$\frac{\exp(\psi)}{1+\exp(\psi)} = \int_0^\infty \exp\left(-\omega\frac{\psi^2}{2}\right) p(\omega) d\omega$$

- Einführung der Polya-Gamma Verteilung zur Darstellung von $p(\omega)$

Mathematische Transformation

- Transformation des logistischen Terms:

$$\frac{1}{(1+\exp(\psi))^b} = 2^{-b} \exp\left(-\frac{b\psi}{2}\right) \int_0^\infty \exp\left(-\omega\frac{\psi^2}{2}\right) p(\omega) d\omega$$

- Durch diese Transformation wird die Abhängigkeit von ψ linearisiert
- Dadurch vereinfachte Berechnung im Rahmen von Gibbs-Sampling

Data Augmentation: Motivation und Idee

- Ziel: Vereinfachung der posterioren Verteilungen
- Idee: Einführung latenter Variablen (ω) zur linearen Darstellung
- Vorteile:
  - Reduktion der Korrelation zwischen Parametern
  - Beschleunigung der Konvergenz von Gibbs-Sampling

Mathematische Herleitung der Data Augmentation

- Logistische Likelihood:

$$p(y_i|\eta_i) = \frac{\exp(y_i\eta_i)}{1 + \exp(\eta_i)}$$

- Einführung der latenten Variablen ω_i:

$$p(y_i|\eta_i) = \int_0^\infty \exp\left(-\omega_i \frac{\eta_i^2}{2}\right) p(\omega_i) d\omega_i$$

- Dabei folgt $\omega_i \sim PG(1, \eta_i)$

Warum funktioniert Data Augmentation?

- Polya-Gamma Variablen ermöglichen conjugate priors für logistische Modelle
- Führt zu bedingten Normalverteilungen für die Parameter
- Verbessert numerische Stabilität und Effizienz

Erdbeben-Vorhersage mit Polya-Gamma

- Seismische Daten: Hohe Unsicherheit und Nichtlinearität
- Modellierung der Auftretenswahrscheinlichkeit von Erdbeben
- Verwendung von logistischen Regressionsmodellen mit Polya-Gamma Data Augmentation

Vorteile und Herausforderungen

- Vorteile:
  - Stabilere Schätzung der Modellparameter
  - Effizientere Sampling-Algorithmen
- Herausforderungen:
  - Hohe Rechenkosten bei großen Datensätzen
  - Komplexität der Modellvalidierung

. . .

. . .

Gutenberg-Richter-Gesetz

- Beschreibt die Häufigkeit von Erdbeben in Abhängigkeit von der Magnitude M
- Mathematische Darstellung:

$$\log_{10} N(M) = a - bM$$

- $N(M)$: Anzahl der Erdbeben mit Magnitude $\geq M$
- Parameter:
  - a: Maß für die seismische Aktivität einer Region
  - b: Skalenfaktor, der die Häufigkeit großer vs. kleiner Beben beschreibt

Bedeutung der Parameter a und b

- **Parameter a**:
  - Bestimmt die absolute Anzahl der Erdbeben in einer Region.
  - Höhere Werte von a bedeuten mehr seismische Aktivität.
  - Beeinflusst durch geologische Gegebenheiten.
- **Parameter b**:
  - Bestimmt das Verhältnis zwischen kleinen und großen Erdbeben.
  - Typischerweise liegt b zwischen 0.8 und 1.2.
  - Niedrige Werte von b deuten auf eine höhere Wahrscheinlichkeit großer Erdbeben hin.

Visualisierung der Parameter a und b

gutenberg_richter_plot.png

Schätzung der Parameter mit Polya-Gamma

- Modellierung der Erdbebenhäufigkeit als Poisson-Prozess:

$$N(M) \sim \text{Poisson}(\lambda(M))$$

- Durch Log-Transformation entsteht ein logistisches Modell:

$$\log N(M) = a - bM + \epsilon$$

- ϵ ist ein Fehlerterm, der oft als Normalverteilung modelliert wird
- Die Polya-Gamma Verteilung kann zur Data Augmentation genutzt werden, um Gibbs-Sampling zu ermöglichen.

...

...

Fazit und Diskussion

- Polya-Gamma Verteilung bietet elegante Lösung für logistische Modelle
- Vielversprechend in der Erdbeben-Vorhersage, aber herausfordernd
- Zukünftige Forschung:
  - Optimierung der Rechenleistung
  - Erweiterung auf komplexere seismische Modelle

Definition der PG(b, c)-Verteilung

Die Pólya-Gamma Verteilung PG(b, c) ist eine exponentiell geneigte Version der PG(b,0)-Verteilung:

$\omega \sim PG(b, c)$ wenn die Dichte gegeben ist durch

$p(\omega|b, c) = \frac{\exp(-c^2\omega/2)p(\omega|b,0)}{\mathbb{E}[\exp(-c^2\omega/2)]}$

wobei $p(\omega|b, 0)$ die Dichte der Standard-Pólya-Gamma-Verteilung PG(b, 0) ist.

Fragen und Diskussion

Vielen Dank!

Fragen?

Einführung

- Bayesianische Inferenz für logistische Modelle ist rechnerisch anspruchsvoll.
- Die Pólya-Gamma (PG) Verteilung ermöglicht eine effiziente Daten-Augmentierung.
- Ziel: Mathematische Grundlagen und Herleitungen für PG(b, c) verstehen.

Definition der Pólya-Gamma Verteilung

Eine Zufallsvariable ω folgt einer **Pólya-Gamma Verteilung** $\omega \sim PG(b, c)$, wenn sie die Dichte

$$p(\omega|b, c) = \frac{\exp(-c^2\omega/2)p(\omega|b, 0)}{\mathbb{E}[\exp(-c^2\omega/2)]} \tag{1}$$

besitzt, wobei $p(\omega|b, 0)$ die Standard-PG-Verteilung ist.

Laplace-Transformation der PG(b, c)-Verteilung

Die Laplace-Transformierte ist gegeben durch:

$$\mathbb{E}[e^{-t\omega}] = \frac{\cosh^{-b}(\sqrt{t}/2)}{\cosh^{-b}(c/2)} \tag{2}$$

Dies zeigt, dass die PG-Verteilung als **unendliches Produkt von Gamma-Verteilungen** geschrieben werden kann.

Darstellung als unendliche Summe von Gammas

$$\omega \sim \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2\pi^2 + c^2/4}, \quad g_k \sim \text{Gamma}(b, 1) \tag{3}$$

Dies zeigt, dass die PG-Verteilung eine **unendliche Mischung von Gamma-Verteilungen** ist.

Erwartungswert und Varianz

$$\mathbb{E}[\omega] = \frac{b}{2c} \tanh(c/2) \tag{4}$$

$$\mathrm{Var}(\omega) = \frac{b}{4c^3}(\sinh(c) - c) \tag{5}$$

Bedingte Verteilung für Gibbs-Sampling

$$(\omega|\psi) \sim PG(b, \psi) \tag{6}$$

Die Posterior-Verteilung der Koeffizienten ergibt sich als Normalverteilung:

$$(\beta|\omega, y) \sim N(\mu_\omega, \Sigma_\omega) \tag{7}$$

mit:

$$\Sigma_\omega = (X^T \Omega X + B^{-1})^{-1}, \quad \mu_\omega = \Sigma_\omega (X^T \kappa + B^{-1} m) \tag{8}$$

Fazit

- Die Pólya-Gamma Verteilung ermöglicht effiziente Gibbs-Sampling-Algorithmen.
- Sie ist unter Addition geschlossen, was für hierarchische Modelle nützlich ist.
- Erwartungswerte und Varianz sind explizit berechenbar, was EM-Algorithmen erleichtert.