

AstroHUB

Astroinformatika II

Pokročilé statistické modelování, hluboké učení a
Big Data

AUTOR

David Pěkník

Datum vydání: 1. 11. 2026
Verze publikace: 1.0

www.astrohub.cz

OBSAH

Úvod a předmluva: Od datové redukce k rigoróznímu modelování	4
--	---

ČÁST I: TEORETICKÁ STATISTIKA A ROBUSTNÍ ESTIMÁTORY

Kapitola 1	Maticové dekompozice a tenzorový počet v astrofyzice	6
1.1	Ortogonalní transformace a singulární rozklad (SVD)	6
1.2	Choleského faktorizace kovariančních matic	7
1.3	Špatně podmíněné problémy a Tichonovova regularizace	7
Kapitola 2	Teorie odhadu, Fisherova informace a asymptotika	9
2.1	Metoda maximální věrohodnosti (MLE)	9
2.2	Fisherova informační matice a Cramérova-Raoova mez	10
2.3	Test poměrem věrohodností a Wilksův teorém	10
Kapitola 3	Robustní statistika a regrese v obou osách	12
3.1	M-estimátory a Huberova ztrátová funkce	12
3.2	Regrese s chybami v obou proměnných: Yorkův algoritmus	12
3.3	Aplikace na kalibraci kosmických svíček	13

ČÁST II: BAYESOVSKÁ INFERENCE A STOCHASTICKÉ VZORKOVÁNÍ (MCMC)

Kapitola 4	Bayesovský aparát a formulace apriorních informací	15
4.1	Bayesův teorém v astrofyzice	15
4.2	Konstrukce neinformativních a Jeffreysových apriorních rozdělení	15
Kapitola 5	Markov Chain Monte Carlo (MCMC) a diagnostika	17
5.1	Metropolisův-Hastingsův algoritmus	17
5.2	Diagnostika konvergence a autokorelační analýza	18
Kapitola 6	Geometrické vzorkovače nové generace (HMC, NUTS a emcee)	19
6.1	Affine-Invariant Ensemble Sampler a knihovna emcee	19
6.2	Hamiltonovské Monte Carlo (HMC) a No-U-Turn Sampler (NUTS)	20
Kapitola 7	Nested Sampling a srovnávání astrofyzikálních modelů	21
7.1	Bayesův faktor a model selection	21
7.2	Algoritmus Nested Sampling podle Johna Skillinga	22

ČÁST III: NEPRAVIDELNÉ ČASOVÉ ŘADY A GAUSSOVSKÉ PROCESY

Kapitola 8	Pokročilá spektrální analýza a Lombův-Scargleův periodogram	24
8.1	Zobecněný Lombův-Scargleův periodogram (GLS)	24
8.2	Analytické metody výpočtu falešného poplachu (Baluevova FAP)	24
8.3	Spektrální okenní funkce a algoritmus CLEAN	25
Kapitola 9	Gaussovské procesy (Gaussian Processes) v časových řadách	26
9.1	Matematická definice Gaussovského procesu	26
9.2	Kovarianční jádra pro astrofyziku	26
9.3	Rychlé škálovatelné výpočty: Algoritmus celerite	27
Kapitola 10	Stochastické autoregresní modely v astrofyzice	28
10.1	Spojité autoregrese: Modely CARMA(p, q)	28
10.2	Tlumená náhodná procházka (DRW) a aktivní galaktická jádra	28
Kapitola 11	Fotodynamické modelování tranzitních systémů	30
11.1	Analytický model tranzitu podle Mandela a Agola	30
11.2	Variace časů tranzitů (TTV) a fotodynamické fitování	30

ČÁST IV: STROJOVÉ UČENÍ: POKROČILÁ REGRESE, KLASIFIKACE A SHLUKOVÁNÍ

Kapitola 12	Regularizace a optimalizace v Data Science	33
12.1	Kompromis mezi vychýlením a rozptylem (Bias-Variance Tradeoff)	33
12.2	L1 a L2 regularizace: Ridge, LASSO a ElasticNet	33
Kapitola 13	Gradient Boosted Trees a náhodné lesy	35
13.1	Náhodné lesy (Random Forests)	35
13.2	Gradient Boosting (XGBoost, LightGBM a CatBoost)	35
Kapitola 14	Pokročilá redukce dimenzionality a nelineární vnoření	37
14.1	Funkční analýza hlavních komponent (Functional PCA)	37
14.2	Nelineární manifold learning: t-SNE a UMAP	37
Kapitola 15	Prostorové shlukování ve fázovém prostoru	39
15.1	Algoritmus HDBSCAN s proměnnou hustotou	39
15.2	Objevování hvězdných proudů a relikтів Gaia-Enceladus	39

ČÁST V: HLUBOKÉ NEURONOVÉ SÍTĚ (DEEP LEARNING) V ASTROFYZICE

Kapitola 16	Architektury a optimalizace hlubokých sítí	42
16.1	Vícevrstvý perceptron a aktivační funkce	42
16.2	Optimalizátory: AdamW a Cosine Annealing	42
Kapitola 17	Konvoluční neuronové sítě (CNN) a segmentace	44
17.1	Reziduální architektury (ResNet)	44
17.2	Architektura U-Net a detekce gravitačních čoček	44
Kapitola 18	Transformery a sekvenční modely pro astronomická data	45
18.1	Matematika mechanismu Multi-Head Attention	45
18.2	Vision Transformers (ViT) na astronomických snímcích	45
Kapitola 19	Generativní modely v astronomickém zobrazování	46
19.1	Variační autoenkodéry (VAE)	46
19.2	Difúzní pravděpodobnostní modely (Score-based Diffusion)	46
Kapitola 20	Fyzikálně informované neuronové sítě (PINN)	48
20.1	Matematický rámec PINN	48
20.2	Aplikace: Hydrostatická rovnováha a rychlé emulátory	48

ČÁST VI: PETABAJTOVÉ PŘEHLÍDKY, BIG DATA A DISTRIBUOVANÉ SYSTÉMY

Kapitola 21	Architektura petabajtových přehlídek nové generace	50
21.1	Vera C. Rubin Observatory (LSST)	50
21.2	Alertové toky v reálném čase: Apache Kafka	50
Kapitola 22	Sférické indexování a masivní prostorové dotazování	51
22.1	HEALPix (Hierarchical Equal Area isoLatitude Pixelization)	51
22.2	Prostorový cross-matching v miliardách zdrojů	51
Kapitola 23	Sloupcové formáty a moderní analytické databáze	53
23.1	Sloupcové formáty: Apache Parquet a Arrow	53
23.2	Moderní in-memory OLAP databáze: DuckDB a ClickHouse	53
Kapitola 24	Distribuované výpočty v Pythonu	55
24.1	Škálování s knihovnou Dask	55
24.2	Knihovna Ray pro masivní paralelismus úloh	55
Kapitola 25	Akcelerace na GPU a tenzorové simulace	56
25.1	Ekosystém CuPy: Okamžitá GPU akcelerace	56
25.2	Diferenciální astrofyzika s knihovnou JAX	56

ČÁST VII: PŘÍLOHY, REPETITORIA A SBÍRKA VĚDECKÝCH ALGORITMŮ

Dodatek A	Matematická odvození klíčových vztahů	59
A.1	Odvození Fisherovy informační matice pro gaussovska data	59
A.2	Odvození akceptačního poměru v Metropolis–Hastings algoritmu	59
Dodatek B	Programátorský referenční manuál moderních knihoven	61
Dodatek C	Sbírka 15 řešených výpočetních úloh v Pythonu	63
Dodatek D	Encyklopedický glosář pokročilé astroinformatiky	64

Úvod a předmluva: Od datové redukce k rigoróznímu modelování

Vítejte u druhého svazku monografie věnované **astroinformatice a moderní analýze astrofyzikálních dat**.

Zatímco první díl (*Úvod do astroinformatiky*) položil intuitivní a konceptuální základy – od fungování křemíkového senzoru přes formát FITS, kalibrační rovnice až po elementární fotometrii a úvod do strojového učení –, tento svazek vstupuje na půdu **teoretické rigoróznosti, pokročilé matematické statistiky a škálovatelné výpočetní vědy**.

V současné éře prochází astrofyzika fundamentální transformací. Klasická pozorovací paradigmaty založená na frekventistických odhadech a jednoduchých chí-kvadrát fitech narážejí na své limity. Vstupujeme do éry petabajtových přehlídek:

- Observatoř Věry C. Rubinové (**Vera C. Rubin Observatory / LSST**) v chilských Andách generuje 20 terabajtů syrových dat každou noc a každých 37 sekund emituje desítky tisíc transientních alertů.
- Kosmická observatoř **Gaia** poskytla přesná astrometrická a spektrofotometrická měření pro téměř dvě miliardy zdrojů, což vytváří fázový prostor o gigantické dimenzionalitě.
- Vesmírné dalekohledy **JWST, Euclid a Roman** poskytují terabajty vysoce strukturovaných multispektrálních a integrálních datových krychlí.

Správná interpretace těchto signálů vyžaduje pokročilé statistické aparáty:

1. **Bayesovskou inferenci:** Schopnost rigorózně začlenit fyzikální apriorní informace a korektně integrovat přes mnohorozměrné prostory parametrů pomocí moderních vzorkovačů (Hamiltonovské Monte Carlo, NUTS, Nested Sampling).
2. **Gaussové procesy a stochastické modely:** Schopnost odfiltrovat komplexní instrumentální a stelární šumové korelace v nerovnoměrně vzorkovaných časových řadách.
3. **Hluboké učení a diferenciální programování:** Využití konvolučních sítí, transformerů, difúzních modelů a fyzikálně informovaných neuronových sítí (PINN) k odhalování skrytých vzorů a urychlení masivních simulací.
4. **Distribučnou architekturu:** Paralelní a GPU výpočty schopné zpracovávat data překračující kapacitu operační paměti běžných pracovních stanic.

Tato kniha je určena studentům magisterského a doktorského studia astrofyziky a teoretické fyziky, datovým vědcům přecházejícím do oblasti přírodních věd a výzkumným pracovníkům, kteří chtějí své analýzy posunout na současnou světovou úroveň.

Část I: Teoretická statistika a robustní estimátory

První část monografie buduje rigorózní matematický aparát moderní statistické analýzy v astrofyzice. Zaměřuje se na maticovou algebru, spektrální dekompozice kovariancí, teorii maximální věrohodnosti, Fisherovu informační geometrii a robustní regresní metody pro měření zatížená chybami v obou proměnných.

Kapitola 1: Maticové dekompozice a tenzorový počet v astrofyzice

Základním jazykem moderní astroinformatiky je lineární algebra a funkcionální analýza. Jakákoliv manipulace se spektry, korelovaným šumem nebo prostorovými daty je v podstatě operací nad maticemi a tenzory v Hilbertově prostoru.

1.1 Ortogonální transformace a singulární rozklad (SVD)

Mějme matici astrofyzikálních pozorování $\mathbf{X} \in \mathbb{R}^{n \times p}$, kde n je počet pozorovaných objektů (např. galaxií) a p je počet naměřených atributů (vlnových délek spektra, fotometrických barev).

Singulární rozklad (Singular Value Decomposition – SVD) je fundamentální faktorizací matice $\mathbf{X}$ na tvar:

$$\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$$

Kde:

- $\mathbf{U} \in \mathbb{R}^{n \times n}$ je ortogonální matice levých singulárních vektorů ($\mathbf{U}^T \mathbf{U} = \mathbf{I}$). Její sloupce tvoří ortonormální bázi v prostoru pozorovaných objektů.
- $\mathbf{\Sigma} \in \mathbb{R}^{n \times p}$ je diagonální matice obsahující nezáporná singulární čísla $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$, kde $r = \min(n, p)$.
- $\mathbf{V} \in \mathbb{R}^{p \times p}$ je ortogonální matice pravých singulárních vektorů ($\mathbf{V}^T \mathbf{V} = \mathbf{I}$). Sloupce matice $\mathbf{V}$ tvoří tzv. hlavní osy variability (eigen-spektra).

Využití v astrofyzice

1. **Spektrální komprese:** Zkrácený singulární rozklad (Truncated SVD) zachovává pouze prvních $k \ll p$ největších singulárních čísel. Podle Eckartova-Youngova-Mirského teorému je matice $\mathbf{X}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ nejlepší aproximací původní matice v normě Frobenius.
2. **Kondiční číslo (Condition Number):** Poměr $\kappa(\mathbf{X}) = \sigma_{\max} / \sigma_{\min}$ kvantifikuje numerickou stabilitu inverzních úloh. Je-li $\kappa(\mathbf{X}) \gg 10^3$, je matice špatně podmíněná a přímá inverze vede k numerické explozi šumu.

1.2 Choleského faktorizace kovariančních matic

V astrofyzice se neustále setkáváme s kovariančními maticemi šumu $\Sigma \in \mathbb{R}^{N \times N}$, které jsou z definice symetrické a pozitivně definitní ($\mathbf{x}^T \Sigma \mathbf{x} > 0$ pro každý nenulový vektor $\mathbf{x}$).

Choleského rozklad dekomponuje matici Σ na součin dolní trojúhelníkové matice $\mathbf{L}$ a její transpozice:

$$\Sigma = \mathbf{L}\mathbf{L}^T$$

Kde $L_{ii} > 0$ pro všechny diagonální prvky.

Výpočetní výhody Choleského rozkladu

- **Bělení šumu (Whitening):** Máme-li vektor korelovaných měření $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$, lineární transformací $\mathbf{z} = \mathbf{L}^{-1}(\mathbf{y} - \mu)$ získáme data s jednotkovým nekorelovaným bílým šumem: $\text{Cov}(\mathbf{z}) = \mathbf{I}$.
- **Inverze a determinant:** Výpočet logaritmu determinantu v gaussovské věrohodnosti se redukuje na pouhý součet diagonálních prvků:

$$\ln |\Sigma| = 2 \sum_{i=1}^N \ln L_{ii}$$

- Výpočetní složitost Choleského rozkladu je $\frac{1}{3}N^3$ operací, což je polovina ve srovnání se standardní LU dekompozicí.

1.3 Špatně podmíněné problémy a Tichonovova regularizace

Mnoho úloh v astroinformatice představuje Fredholmovu integrální rovnici prvního druhu:

$$d(x) = \int K(x, x') f(x') dx' + \epsilon(x)$$

Po diskretizaci získáme lineární systém $\mathbf{d} = \mathbf{K}\mathbf{f} + \epsilon$, kde $\mathbf{K}$ je konvoluční matice bodové odezvy přístroje (PSF). Přímé řešení metodou nejmenších čtverců $\hat{\mathbf{f}} = (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{d}$ dramaticky selhává, protože vysokofrekvenční šum ϵ je zesílen malými singulárními čísly matice $\mathbf{K}$.

Tichonovova regularizace (Ridge regrese) modifikuje optimalizační kritérium:

$$\hat{\mathbf{f}}_\lambda = \arg \min_{\mathbf{f}} \{ \|\mathbf{d} - \mathbf{K}\mathbf{f}\|_2^2 + \lambda^2 \|\mathbf{\Gamma}\mathbf{f}\|_2^2 \}$$

Kde $\mathbf{\Gamma}$ je regularizační operátor (obvykle jednotková matice $\mathbf{I}$ nebo diferenční matice derivace pro vyhlazení profilu) a λ je regularizační parametr. Analytické řešení má tvar:

$$\hat{\mathbf{f}}_\lambda = (\mathbf{K}^T \mathbf{K} + \lambda^2 \mathbf{\Gamma}^T \mathbf{\Gamma})^{-1} \mathbf{K}^T \mathbf{d}$$

Volba optimálního λ se provádí metodou zobecněné křížové validace (GCV) nebo L-křivky (L-curve criterion), která balancuje reziduální normu proti velikosti regularizačního členu.

Kapitola 2: Teorie odhadu, Fisherova informace a asymptotika

Při interpretaci astrofyzikálních dat potřebujeme z naměřených fotonů odhadnout fyzikální parametry modelů (hmotnosti hvězd, parametry temné energie, chemické zastoupení). Teorie odhadu poskytuje matematický rámec pro konstrukci optimálních estimátorů a vyčíslení jejich fundamentálních limitů.

2.1 Metoda maximální věrohodnosti (MLE)

Mějme vektor pozorovaných dat $\mathbf{x} = (x_1, \dots, x_N)^T$ nezávisle a identicky rozdělených (i.i.d.) podle hustoty pravděpodobnosti $p(x|\theta)$, kde $\theta = (\theta_1, \dots, \theta_k)^T \in \Theta$ je vektor neznámých parametrů.

Funkce věrohodnosti (Likelihood function) je definována jako:

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^N p(x_i|\theta)$$

Z numerických důvodů a pro zjednodušení derivací pracujeme s **logaritmickou věrohodností (Log-Likelihood)**:

$$\ell(\theta) = \ln \mathcal{L}(\theta|\mathbf{x}) = \sum_{i=1}^N \ln p(x_i|\theta)$$

Estimátor maximální věrohodnosti (MLE) $\hat{\theta}_{\text{MLE}}$ je definován jako argument maximalizující log-věrohodnost:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} \ell(\theta)$$

Což vede na soustavu věrohodnostních rovnic:

$$\nabla_{\theta} \ell(\theta) = \mathbf{0}$$

Asymptotické vlastnosti MLE

Při splnění podmínek regularity (hladkost hustoty p , kompaktnost Θ) vykazuje MLE estimátor pro $N \rightarrow \infty$ klíčové vlastnosti:

1. **Konzistence:** $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta_0$ (konverguje v pravděpodobnosti ke skutečné hodnotě).

2. **Asymptotická normalita:** $\sqrt{N}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_1^{-1}(\theta_0))$.
3. **Asymptotická efience:** Žádný nestranný odhad nemůže mít menší asymptotický rozptyl než MLE.

2.2 Fisherova informační matice a Cramérova–Raoova mez

Definujme **skóre funkci (Score function)** jako gradient log-věrohodnosti:

$$\mathbf{S}(\theta) = \nabla_{\theta} \ln p(\mathbf{x}|\theta)$$

Střední hodnota skóre funkce vzhledem k datům je nulová: $\mathbb{E}[\mathbf{S}(\theta)] = \mathbf{0}$.

Fisherova informační matice (Fisher Information Matrix – FIM) $\mathbf{I}(\theta) \in \mathbb{R}^{k \times k}$ je definována jako kovariance skóre funkce:

$$I_{ij}(\theta) = \mathbb{E} \left[\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta_i} \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta_j} \right] = -\mathbb{E} \left[\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta_i \partial \theta_j} \right]$$

Cramérova–Raoova nerovnost (Cramér–Rao Lower Bound – CRLB)

Pro libovolný nestranný odhad $\hat{\theta}$ parametru θ platí:

$$\text{Cov}(\hat{\theta}) \geq \mathbf{I}^{-1}(\theta)$$

Ve smyslu, že matice $\text{Cov}(\hat{\theta}) - \mathbf{I}^{-1}(\theta)$ je pozitivně semidefinitní.

❗ DŮLEŽITÉ

Fisherova matice stanovuje absolutní fyzikální limit přesnosti, jaké může jakýkoliv přístroj nebo pozorovací mise dosáhnout při daném počtu fotonů. Je základním nástrojem při navrhování kosmických misí (např. odhad přesnosti měření temné energie pro teleskop Euclid).

2.3 Test poměrem věrohodností a Wilksův teorém

Mějme model s k parametry a testujme nulovou hypotézu $H_0 : \theta \in \Theta_0$ (kde Θ_0 má dimenzi $k_0 < k$) proti obecné alternativě $H_1 : \theta \in \Theta$.

Statistika poměru věrohodností (Likelihood Ratio Test – LRT) je definována vztahem:

$$\Lambda = 2 \left[\ell(\hat{\theta}) - \ell(\hat{\theta}_0) \right]$$

Podle **Wilksova teorému** má za platnosti nulové hypotézy H_0 statistika Λ asymptoticky rozdělení chí-kvadrát s počtem stupňů volnosti rovným rozdílu dimenzí:

$$\Lambda \stackrel{d}{\rightarrow} \chi^2(k - k_0) \quad \text{pro } N \rightarrow \infty$$

Tento teorém je základem pro stanovení intervalů spolehlivosti v astrofyzikálních fitech: kontura $\Delta\chi^2 = 2,30$ vymezuje 68,3% konfidenční oblast pro 2 volné parametry, zatímco $\Delta\chi^2 = 6,18$ vymezuje 95,4% oblast.

Kapitola 3: Robustní statistika a regrese v obou osách

V astrofyzice jsou měření téměř vždy zatížena negaussovskými odlehlými hodnotami (chybně identifikované hvězdy, kontaminace kosmickým zářením) a nejistoty existují jak na nezávisle proměnné (X), tak na závisle proměnné (Y). Standardní lineární regrese v takových podmínkách generuje hrubě zkreslené výsledky.

3.1 M-estimátory a Huberova ztrátová funkce

Místo kvadratické ztrátové funkce metody nejmenších čtverců minimalizují **M-estimátory** obecnější funkci reziduí $r_i = y_i - f(x_i; \theta)$:

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \rho \left(\frac{r_i}{\sigma} \right)$$

Kde $\rho(u)$ je symetrická, nezáporná funkce a σ je robustní odhad měřítka rozptylu (např. $\text{MAD} = \text{median}(|r_i - \text{median}(r_i)|)/0,6745$).

Huberova ztrátová funkce kombinuje to nejlepší z kvadratické a absolutní odchylky:

$$\rho_{\delta}(u) = \begin{cases} \frac{1}{2}u^2 & \text{pro } |u| \leq \delta \\ \delta \left(|u| - \frac{1}{2}\delta \right) & \text{pro } |u| > \delta \end{cases}$$

Váhovou funkcí v algoritmu iterativně převažovaných nejmenších čtverců (IRLS) je pak:

$$w(u) = \frac{\rho'_{\delta}(u)}{u} = \begin{cases} 1 & \text{pro } |u| \leq \delta \\ \frac{\delta}{|u|} & \text{pro } |u| > \delta \end{cases}$$

Typická volba $\delta = 1,345$ zajišťuje 95% účinnost pro čisté Gaussovo rozdělení a současně absolutní odolnost vůči těžkým chybám.

3.2 Regrese s chybami v obou proměnných: Yorkův algoritmus

Uvažujme lineární model $y = a + bx$, kde pro každý bod měříme dvojici (x_i, y_i) s nejistotami $(\sigma_{x,i}, \sigma_{y,i})$ a vzájemnou kovariancí $\text{Cov}(x_i, y_i) = r_i \sigma_{x,i} \sigma_{y,i}$.

Standardní metoda OLS předpokládá $\sigma_{x,i} = 0$. Zanedbání chyb v ose X vede k tzv. **atenuačnímu zkreslení (Attenuation Bias)** – směrnice přímky b je systematicky podhodnocena směrem k nule!

Plně rigorózní řešení odvodil Derek York (1966, 2004). Minimalizujeme vážený součet čtverců odchylek v obou osách:

$$S = \sum_{i=1}^N [\omega_{x,i}(x_i - \hat{x}_i)^2 - 2\omega_{xy,i}(x_i - \hat{x}_i)(y_i - \hat{y}_i) + \omega_{y,i}(y_i - \hat{y}_i)^2]$$

York ukázal, že hledaná směrnice b je kořenem nelineární rovnice:

$$b = \frac{\sum_{i=1}^N W_i(x_i - \bar{x})(y_i - \bar{y} + b(x_i - \bar{x}))}{\sum_{i=1}^N W_i(x_i - \bar{x})^2}$$

Kde efektivní váha W_i je dána vztahem:

$$W_i = \frac{\omega_{x,i}\omega_{y,i}}{\omega_{x,i} + b^2\omega_{y,i} - 2b\omega_{xy,i}} = \frac{1}{\sigma_{y,i}^2 + b^2\sigma_{x,i}^2 - 2br_i\sigma_{x,i}\sigma_{y,i}}$$

A těžiště $\bar{x}, \bar{y}$ jsou váženými průměry s váhami W_i . Tato rovnice se řeší iterativně s rychlou konvergencí (obvykle 4 až 6 iterací).

3.3 Aplikace na kalibraci kosmických svíček

Yorkův algoritmus a robustní odhady jsou kritické pro kalibraci fundamentálních astrofyzikálních relací:

- **Tullyova–Fisherova relace:** Vztah mezi zářivým výkonem spirální galaxie L a šířkou spektrální čáry W neutrálního vodíku HI ($L \propto W^4$). Chyby v měření rotační rychlosti z profilu čáry 21 cm jsou srovnatelné s fotometrickými chybami.
- **Faberova–Jacksonova relace:** Vztah pro eliptické galaxie ($L \propto \sigma_v^4$), kde σ_v je centrální rychlostní disperze hvězd.
- **Kalibrace supernov typu Ia:** Vztah mezi absolutní magnitudou supernovy v maximu a rychlostí poklesu světelné křivky (Phillipsův vztah). Bez korektní oboustranné regrese by určení kosmologických parametrů Ω_m a Ω_Λ bylo zatíženo systematickou chybou.

Část II: Bayesovská inference a stochastické vzorkování (MCMC)

Druhá část je věnována fundamentálnímu pilíři moderní astrofyziky – bayesovské inferenci. Probereme formulaci apriorních informací, integraci v mnohorozměrných prostorech pomocí Markov Chain Monte Carlo (MCMC), moderní algoritmy HMC a NUTS a výpočet marginální věrohodnosti metodou Nested Sampling.

Kapitola 4: Bayesovský aparát a formulace apriorních informací

V moderní astrofyzice nechápeme pravděpodobnost jako limitní relativní četnost pokusů v nekonečné sérii opakování (frekventistický pohled), ale jako **míru věrohodnosti či stavu našeho poznání** (Bayesovský pohled). Tento posun je klíčový: náš vesmír je pouze jeden a nemůžeme provést sérii deseti tisíc nezávislých Velkých třesků.

4.1 Bayesův teorém v astrofyzice

Základní postulát bayesovské inference je vyjádřen vztahem:

$$p(\theta|\mathbf{D}, \mathcal{M}) = \frac{p(\mathbf{D}|\theta, \mathcal{M}) p(\theta|\mathcal{M})}{p(\mathbf{D}|\mathcal{M})}$$

Kde jednotlivé složky reprezentují:

1. $p(\theta|\mathcal{M})$ – **Apriorní pravděpodobnost (Prior)**: Kvantifikuje naše znalosti o parametrech θ před zahrnutím nových dat $\mathbf{D}$, za předpokladu platnosti modelu $\mathcal{M}$.
2. $p(\mathbf{D}|\theta, \mathcal{M}) = \mathcal{L}(\theta)$ – **Věrohodnost (Likelihood)**: Pravděpodobnost naměření dat $\mathbf{D}$ při konkrétní volbě parametrů θ .
3. $p(\theta|\mathbf{D}, \mathcal{M})$ – **Aposteriorní pravděpodobnost (Posterior)**: Aktualizovaný stav našeho poznání o parametrech po zahrnutí experimentálních dat.
4. $p(\mathbf{D}|\mathcal{M}) = \mathcal{Z}$ – **Marginální věrohodnost (Evidence)**: Normalizační konstanta, která integruje čítelek přes celý prostor parametrů:

$$\mathcal{Z} = \int_{\Omega_{\theta}} p(\mathbf{D}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta$$

Pro odhad samotných parametrů θ v rámci jednoho modelu $\mathcal{M}$ je hodnota $\mathcal{Z}$ nepodstatná, neboť nezávisí na θ , a píšeme:

$$p(\theta|\mathbf{D}) \propto p(\mathbf{D}|\theta) p(\theta)$$

4.2 Konstrukce neinformativních a Jeffreysových apriorních rozdělání

Jednou z nejčastějších výzev v bayesovské analýze je: jak formulovat apriorní hustotu $p(\theta)$, pokud o parametru nemáme žádné předchozí fyzikální informace?

Naivní volba rovnoměrného rozdělení $p(\theta) \propto \text{const}$ **není invariantní vůči reparametrizaci!** Pokud zvolíme rovnoměrný prior na parametr θ , pak transformovaný parametr $\phi = f(\theta)$ (např. $\phi = \ln \theta$ nebo $\phi = 1/\theta$) získá podle věty o transformaci hustot netriviální tvar:

$$p_\phi(\phi) = p_\theta(f^{-1}(\phi)) \left| \frac{d\theta}{d\phi} \right|$$

Harold Jeffreys (1946) odvodil pravidlo pro konstrukci **objektivního, reparametrizačně invariantního prioru** založeného na Fisherově informační matici:

$$p_{\text{Jeffreys}}(\theta) \propto \sqrt{\det \mathbf{I}(\theta)}$$

Klíčové výsledky pro astrofyziku:

1. **Poziční parametr (Location parameter):** Pokud $p(x|\theta) = f(x - \theta)$ (např. střední hodnota gaussovského rozdělení, poloha středu hvězdy na snímku), pak $\mathbf{I}(\theta) = \text{const}$ a Jeffreysův prior je uniformní:

$$p(\theta) \propto 1 \quad (\text{Lebesgueova míra})$$

2. **Škálovací parametr (Scale parameter):** Pokud $p(x|\sigma) = \frac{1}{\sigma} f(x/\sigma)$ (např. směrodatná odchylka šumu σ , poloměr exoplanety, doba trvání záblesku), pak $\mathbf{I}(\sigma) \propto 1/\sigma^2$ a Jeffreysův prior má tvar:

$$p(\sigma) \propto \frac{1}{\sigma} \iff p(\ln \sigma) \propto 1$$

Jeffreysův prior na škálovací parametr je ekvivalentní rovnoměrnému rozdělení v logaritmickém měřítku.

Kapitola 5: Markov Chain Monte Carlo (MCMC) a diagnostika

V reálných astrofyzikálních modelech má prostor parametrů desítky až stovky dimenzí (např. 15 orbitálních parametrů vícenásobného planetárního systému). Analytická integrace je nemožná a mřížkový výpočet (grid search) selhává kvůli exponenciálnímu růstu bodů (pro $D = 20$ dimenzí s pouhými 10 body na osu potřebujeme 10^{20} vyhodnocení věrohodnosti). Řešením jsou **Markovovy řetězce Monte Carlo (MCMC)**.

5.1 Metropolisův–Hastingsův algoritmus

Cílem MCMC je vygenerovat posloupnost stavů $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$, jejichž stacionární asymptotické rozdělení je přesně rovno hledanému posterioru $\pi(\theta) = p(\theta|\mathbf{D})$.

Metropolisův–Hastingsův algoritmus generuje Markovův řetězec pomocí návrhové hustoty (Proposal distribution) $q(\theta'|\theta)$:

1. Zvolme počáteční bod $\theta^{(0)}$ a nastavme $t = 0$.
2. V kroku t vygenerujme kandidátský bod $\theta' \sim q(\theta'|\theta^{(t)})$.
3. Vypočítejme **Hastingsův akceptační poměr (Acceptance Ratio)**:

$$\alpha(\theta^{(t)}, \theta') = \min \left(1, \frac{\pi(\theta') q(\theta^{(t)}|\theta')}{\pi(\theta^{(t)}) q(\theta'|\theta^{(t)})} \right)$$

4. Vygenerujme náhodné číslo $u \sim \mathcal{U}(0, 1)$. Pokud $u \leq \alpha$, krok **přijmeme** a položíme $\theta^{(t+1)} = \theta'$.
V opačném případě krok **zamítneme** a zůstáváme v původním bodě: $\theta^{(t+1)} = \theta^{(t)}$.
5. Zvýšíme $t \leftarrow t + 1$ a opakujeme od kroku 2.

Důkaz detailní rovnováhy (Detailed Balance)

Algoritmus garantuje, že řetězec konverguje k $\pi(\theta)$, protože splňuje podmínku detailní rovnováhy:

$$\pi(\theta)T(\theta \rightarrow \theta') = \pi(\theta')T(\theta' \rightarrow \theta)$$

Kde přechodová pravděpodobnost $T(\theta \rightarrow \theta') = q(\theta'|\theta)\alpha(\theta, \theta')$. Normalizační konstanta posterioru $\mathcal{Z}$ se v poměru $\pi(\theta')/\pi(\theta)$ beze zbytku zkrátí!

5.2 Diagnostika konvergence a autokorelační analýza

Surový výstup z MCMC nelze použít přímo. Řetězce vykazují autokorelaci a vyžadují odstranění počáteční fáze zahřívání (**Burn-in / Warmup**).

1. Autokorelační čas τ

Kovariance vzorků vzdálených o zpoždění k je dána autokorelační funkcí:

$$\rho(k) = \frac{\mathbb{E}[(\theta_t - \mu)(\theta_{t+k} - \mu)]}{\sigma^2}$$

Integrovaný autokorelační čas je definován vztahem:

$$\tau = 1 + 2 \sum_{k=1}^{\infty} \rho(k)$$

Fyzikální význam τ : Abychom získali jeden skutečně **nezávislý vzorek** z posterioru, musíme v MCMC řetězci udělat přibližně τ kroků.

2. Efektivní velikost vzorku (N_{eff})

Z celkového počtu N vygenerovaných kroků je efektivní počet nezávislých měření roven:

$$N_{\text{eff}} = \frac{N}{\tau}$$

Pro spolehlivé určení 95% konfidenčních intervalů požadujeme $N_{\text{eff}} \geq 1\,000$.

3. Gelman–Rubinův diagnostický test ($\hat{R}$)

Spustíme $M \geq 4$ nezávislých řetězců z různých startovních bodů v prostoru parametrů. Spočítáme rozptyl mezi řetězci (B/n) a průměrný vnitřní rozptyl uvnitř řetězců (W). Odhad koeficientu zmenšení rozptylu $\hat{R}$ je:

$$\hat{R} = \sqrt{\frac{W + \frac{1}{n}(B - W)}{W}}$$

Pokud $\hat{R} > 1,05$, řetězce ještě nekonvergovaly a prozkoumávají odlišné lokální mody. V publikovatelných astrofyzikálních pracích je standardem dosáhnout $\hat{R} < 1,01$ pro všechny volné parametry.

Kapitola 6: Geometrické vzorkovače nové generace (HMC, NUTS a emcee)

Klasický Metropolis–Hastingsův algoritmus s náhodnou procházkou (Random Walk) dramaticky trpí tzv. **prokletím dimenzionality**: aby byl akceptační poměr rozumný ($\approx 23,4\%$), musí být krok návrhové funkce velmi malý ($\propto D^{-1/2}$), což vede k obrovskému autokorelačnímu času. V moderní astroinformatice proto dominují geometrické a souborové vzorkovače.

6.1 Affine–Invariant Ensemble Sampler a knihovna `emcee`

V roce 2010 publikovali Jonathan Goodman a Jonathan Weare přelomový algoritmus **afinně invariantního souborového vzorkování**, který v roce 2013 zpopularizoval Daniel Foreman–Mackey v legendárním astronomickém balíčku ``emcee``.

Místo jednoho bodu simuluje algoritmus soubor K současně se pohybujících částic (**Walkers**), kde $K \geq 2D$.

V každém kroku aktualizujeme polohu walkeru k pomocí tzv. **protaženého tahu (Stretch Move)** vůči náhodně vybranému walkeru j ze zbývajících souboru:

$$\theta' = \theta_j + Z(\theta_k - \theta_j)$$

Kde náhodná proměnná Z je vzorkována z hustoty:

$$g(z) \propto \frac{1}{\sqrt{z}} \quad \text{pro } z \in [1/a, a], \quad a = 2$$

Akceptační poměr pro tento krok je:

$$\alpha = \min \left(1, Z^{D-1} \frac{\pi(\theta')}{\pi(\theta_k)} \right)$$

Proč je `emcee` astronomickým standardem?

Algoritmus je **afinně invariantní**: lineární transformace souřadnic $\theta' = \mathbf{A}\theta$ (např. změna jednotek nebo silná korelace mezi parametry) nemá vůbec žádný vliv na efektivitu vzorkování! Není nutné ladit kovarianční matici návrhové funkce.

6.2 Hamiltonovské Monte Carlo (HMC) a No-U-Turn Sampler (NUTS)

Pokud je věrohodnostní funkce diferencovatelná a máme k dispozici její gradient $\nabla_{\theta} \ln \pi(\theta)$, můžeme využít fyzikální analogii z teoretické mechaniky – **Hamiltonovské Monte Carlo (HMC)**.

Představme si parametr θ jako polohu částice v potenciálovém poli $V(\theta) = -\ln \pi(\theta)$. K částici přidružíme fiktivní hybnost $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$. Celkový Hamiltonián systému je:

$$H(\theta, \mathbf{p}) = V(\theta) + K(\mathbf{p}) = -\ln \pi(\theta) + \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}$$

Pohybové rovnice jsou Hamiltonovy kanonické rovnice:

$$\frac{d\theta}{dt} = \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1} \mathbf{p}, \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial H}{\partial \theta} = \nabla_{\theta} \ln \pi(\theta)$$

Částice je numericky integrována symplektickým **Leapfrog integrátorem**:

1. Poloviční krok v hybnosti: $\mathbf{p}(t + \epsilon/2) = \mathbf{p}(t) + \frac{\epsilon}{2} \nabla \ln \pi(\theta(t))$
2. Celý krok v poloze: $\theta(t + \epsilon) = \theta(t) + \epsilon \mathbf{M}^{-1} \mathbf{p}(t + \epsilon/2)$
3. Poloviční krok v hybnosti: $\mathbf{p}(t + \epsilon) = \mathbf{p}(t + \epsilon/2) + \frac{\epsilon}{2} \nabla \ln \pi(\theta(t + \epsilon))$

Díky zachování Hamiltoniánu dosahuje HMC akceptačního poměru přes 65 až 80 % i v prostorech s tisíci dimenzemi!

No-U-Turn Sampler (NUTS)

Algoritmus NUTS (Hoffman & Gelman, 2014) odstraňuje nutnost ručního nastavování počtu kroků integrace L : automaticky staví rekursivní binární strom trajektorie a zastaví integraci ve chvíli, kdy by se částice začala vracet zpět (podmínka $\frac{d}{dt} \|\theta - \theta_0\|^2 < 0$). NUTS je srdcem moderních pravděpodobnostních knihoven **PyMC**, **Stan** a **NumPyro/JAX**.

Kapitola 7: Nested Sampling a srovnávání astrofyzikálních modelů

Častou otázkou v astrofyzice není jen „*jaké jsou nejlepší parametry modelu?*“, nýbrž „*který model lépe vysvětluje data?*“:

- Obsahuje naměřená světelná křivka jednu exoplanetu, nebo dvě?
- Je kosmologická křivka lépe popsána modelem Λ CDM s konstantní temnou energií, nebo dynamickou temnou energií $w(a) = w_0 + w_a(1 - a)$?

K zodpovězení této otázky potřebujeme spočítat **marginální věrohodnost (Evidence) $\mathcal{Z}$** .

7.1 Bayesův faktor a model selection

Mějme dva konkurenční fyzikální modely $\mathcal{M}_1$ a $\mathcal{M}_2$. Poměr jejich aposteriorních pravděpodobností je:

$$\frac{p(\mathcal{M}_1|\mathbf{D})}{p(\mathcal{M}_2|\mathbf{D})} = \underbrace{\frac{p(\mathbf{D}|\mathcal{M}_1)}{p(\mathbf{D}|\mathcal{M}_2)}}_{B_{12} \text{ (Bayesův faktor)}} \times \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{Prior odds}}$$

Bayesův faktor $B_{12} = \mathcal{Z}_1/\mathcal{Z}_2$ vyjadřuje, kolikrát data zvýhodňují model $\mathcal{M}_1$ oproti $\mathcal{M}_2$.

Podle Harolda Jeffreyse interpretujeme rozdíl $\Delta \ln \mathcal{Z} = \ln B_{12}$ následovně:

- $\Delta \ln \mathcal{Z} < 1$: Nekonzistentní, bezvýznamný rozdíl.
- $1 < \Delta \ln \mathcal{Z} < 2,5$: Mírný důkaz ve prospěch $\mathcal{M}_1$.
- $2,5 < \Delta \ln \mathcal{Z} < 5$: Silný důkaz ve prospěch $\mathcal{M}_1$ (šance zhruba 12 : 1 až 150 : 1).
- $\Delta \ln \mathcal{Z} > 5$: Rozhodující důkaz (**Decisive Evidence**, šance větší než 150 : 1, odpovídá statistické signifikanci $> 3,6\sigma$).

ZAJÍMAVOST

Bayesův faktor v sobě automaticky zahrnuje **Occamovu břitvu**: složitější model s větším počtem volných parametrů je penalizován za velký fázový objem parametrů, v němž je věrohodnost nízká. Není nutné zavádět umělé penalizační členy jako u AIC nebo BIC!

7.2 Algoritmus Nested Sampling podle Johna Skillinga

Výpočet integrálu $\mathcal{Z} = \int \mathcal{L}(\theta) \pi(\theta) d\theta$ je extrémně obtížný, protože většina objemu má zanedbatelnou věrohodnost.

V roce 2004 navrhl fyzik John Skilling geniální transformaci: převedl mnohorozměrný integrál na jednorozměrný integrál přes kumulativní objem apriorní pravděpodobnosti X :

$$X(\lambda) = \int_{\mathcal{L}(\theta) > \lambda} \pi(\theta) d\theta \in [0, 1]$$

Integrál pro evidenci se pak zjednoduší na:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX$$

Algoritmus krok za krokem:

1. Vygenerujeme N_{live} aktivních bodů (**Live Points**) z apriorního rozdělení $\pi(\theta)$. Počáteční objem je $X_0 = 1$.
2. V každém kroku i : * Najdeme live bod s nejmenší věrohodností $\mathcal{L}_i = \min_k \mathcal{L}(\theta_k)$. * Tento bod vyjmeme a uložíme jako „mrtvý bod“ (**Dead Point**) s váhou $w_i = X_{i-1} - X_i$. * Apriorní objem se smrští: $X_i \approx \exp(-i/N_{\text{live}})$. * Příspěvek k evidenci vzroste o $\Delta \mathcal{Z}_i = \mathcal{L}_i w_i$. * Vygenerujeme nový live bod z apriorního rozdělení pod podmínkou, že $\mathcal{L}(\theta_{\text{new}}) > \mathcal{L}_i$.
3. Algoritmus ukončíme, jakmile zbývajících odhadovaný integrál $\mathcal{L}_{\text{max}} X_i$ tvoří zanedbatelnou frakci celkové sumy $\mathcal{Z}$.

Implementace v moderní astrofyzice

- **‘MultiNest’ (Feroz & Hobson)**: Využívá elipsoidální dekompozici pro vzorkování multimodálních posteriorů.
- **‘Dynesty’ (Speagle)**: Dynamický Nested Sampling v čistém Pythonu, který dokáže v průběhu výpočtu měnit počet live points pro maximalizaci přesnosti odhadu parametrů i evidence.

Část III: Nepravidelné časové řady a gaussovske procesy

Třetí část je věnována dynamickému vesmíru. Zkoumá pokročilou spektrální analýzu nerovnoměrně vzorkovaných světelných křivek, teorii gaussovských procesů (GP) pro filtraci hvězdné variability, stochastické diferenciální autoregrese (CARMA) v aktivních galaktických jádrech a fotodynamické modelování planetárních tranzitů.

Kapitola 8: Pokročilá spektrální analýza a Lombův–Scargleův periodogram

Astronomická fotometrická pozorování jsou z podstaty nepravidelně vzorkovaná s velkými mezerami způsobenými střídáním dne a noci, počasím a orbitální geometrií. Klasická diskretní Fourierova transformace na takových datech generuje masivní spektrální svod a falešné periody.

8.1 Zobecněný Lombův–Scargleův periodogram (GLS)

Původní formulace Normana Lomba (1976) a Jeffreyho Scargla (1982) předpokládala nulovou střední hodnotu dat a homogenní chyby. V reálných pozorováních z družic Kepler nebo TESS však data obsahují neznámý plovoucí průměr a individuální nejistoty měření σ_i .

Mathias Zechmeister a Martin Kürster (2009) odvodili **Zobecněný Lombův–Scargleův periodogram (Generalized Lomb–Scargle – GLS)** fitováním harmonického modelu s plovoucím offsetem pro každou úhlovou frekvenci $\omega = 2\pi f$:

$$y(t) = a \cos(\omega t) + b \sin(\omega t) + c$$

Vážená redukce chí-kvadrát vůči konstantnímu modelu χ_0^2 definuje normovaný výkon periodogramu:

$$P_{\text{GLS}}(\omega) = \frac{1}{YY} \left[\frac{\hat{C}C_\omega + \hat{S}S_\omega - 2CS C_\omega S_\omega}{CC SS - (CS)^2} \right]$$

Kde $\hat{C}$, $\hat{S}$, CC , SS , CS jsou vážené sumy goniometrických funkcí se zohledněním vah $w_i = 1/\sigma_i^2$ a fázového posunu $\tau(\omega)$, který zajišťuje ortogonalitu sinu a kosinu:

$$\tan(2\omega\tau) = \frac{\sum w_i \sin(2\omega t_i) - 2\hat{S}\hat{C}/W}{\sum w_i \cos(2\omega t_i) - (\hat{C}^2 - \hat{S}^2)/W}$$

GLS je invariantní vůči lineárnímu posunu dat a je schopen správně extrahovat signál i při přítomnosti významného rozdílu v přesnosti mezi jednotlivými pozorovacími nocemi.

8.2 Analytické metody výpočtu falešného poplachu (Baluevova FAP)

Klíčovou otázkou je: *Jaká je pravděpodobnost, že nalezený pík v periodogramu o výšce Z je pouhou náhodnou fluktuací šumu?* Této pravděpodobnosti říkáme **False Alarm Probability (FAP)**.

Prohledáváme-li frekvenční pásmo od 0 do $f_{\max}$, nelze použít jednoduché rozdělení pro jedinou frekvenci, protože testujeme stovky až tisíce nezávislých frekvencí (problém mnohonásobného testování hypotéz).

Roman Baluev (2008) odvodil analytickou horní mez FAP založenou na teorii geometrie náhodných polí (Riceův vzorec pro překročení prahu):

$$\text{FAP}(Z) \leq W(Z) + G \sqrt{Z} (1 - Z)^{(N-4)/2}$$

Kde N je počet bodů, $Z = P_{\text{GLS}}$ je normalizovaný výkon a geometrický faktor G závisí na efektivním rozpětí času měření:

$$G \approx 2\pi^{1/2} f_{\max} T_{\text{eff}}, \quad T_{\text{eff}} = \sqrt{\sum_{i=1}^N w_i (t_i - \bar{t})^2}$$

Tento analytický vztah umožňuje okamžitý výpočet FAP i pro hladiny 10^{-6} bez nutnosti provádět miliony zdlouhavých bootstrap simulací.

8.3 Spektrální okenní funkce a algoritmus CLEAN

Časové rozložení pozorování definuje **spektrální okenní funkci (Spectral Window Function)**:

$$W(\omega) = \left| \frac{1}{N} \sum_{k=1}^N e^{-i\omega t_k} \right|^2$$

Skutečný periodogram $P_{\text{obs}}(\omega)$ je konvolucí reálného spektra signálu $P_{\text{true}}(\omega)$ se spektrálním oknem $W(\omega)$:

$$P_{\text{obs}}(\omega) = P_{\text{true}}(\omega) * W(\omega)$$

Vlivem denního cyklu (perioda 24 hodin, frekvence $1 \text{ den}^{-1} \approx 11,6 \mu\text{Hz}$) se kolem každé reálné frekvence f_0 objevují mohutné falešné píky (aliasy) na frekvencích $f_0 \pm 1 \text{ den}^{-1}$, $f_0 \pm 2 \text{ den}^{-1}$.

Algoritmus **CLEAN** (Roberts et al., 1987) iterativně dekonvoluuje spektrální okno:

1. Nalezne nejvyšší pík v periodogramu.
2. Odečte z časové řady odpovídající sinusový signál vynásobený tlumicím faktorem (gain factor $g \approx 0,1$).
3. Přepočítá periodogram reziduí.
4. Tento proces opakuje, dokud rezidua neodpovídají bílému šumu.

Kapitola 9: Gaussovské procesy (Gaussian Processes) v časových řadách

Při hledání extrasolárních planet a studiu proměnných hvězd čelíme zásadnímu problému: hvězdy nejsou klidné zářící koule. Jejich povrch pokrývají rotující magnetické skvrny, dochází ke granulaci a erupcím. Tento stelární signál tvoří **korelovaný červený šum**, který snadno napodobí signál exoplanety. **Gaussovské procesy (GP)** představují špičkový matematický aparát pro neparametrické modelování tohoto šumu.

9.1 Matematická definice Gaussovského procesu

Gaussovský proces je definován jako nekonečněrozměrná kolekce náhodných proměnných, z nichž každá konečná podmnožina má sdružené vícerozměrné normální rozdělení:

$$f(t) \sim \mathcal{GP}(m(t), k(t, t'; \theta))$$

Kde:

- $m(t) = \mathbb{E}[f(t)]$ je střední funkce (často parametrický fyzikální model tranzitu nebo Keplerovy dráhy).
- $k(t, t'; \theta) = \text{Cov}(f(t), f(t'))$ je **kovarianční funkce (jádro – Kernel)** parametrizovaná hyperparametry θ .

Mějme vektor naměřených dat $\mathbf{y} = (y_1, \dots, y_N)^T$ v časech $\mathbf{t} = (t_1, \dots, t_N)^T$ s bílým šumem měření σ_i^2 . Vektorová kovarianční matice je:

$$\mathbf{K}_{ij} = k(t_i, t_j; \theta) + \sigma_i^2 \delta_{ij}$$

Logaritmická marginální věrohodnost hyperparametrů θ má elegantní analytický tvar:

$$\ln p(\mathbf{y}|\mathbf{t}, \theta) = -\frac{1}{2}(\mathbf{y} - \mathbf{m})^T \mathbf{K}^{-1}(\mathbf{y} - \mathbf{m}) - \frac{1}{2} \ln \det \mathbf{K} - \frac{N}{2} \ln(2\pi)$$

První člen měří shodu modelu s daty, zatímco druhý člen ($\ln \det \mathbf{K}$) představuje přirozenou Occamovu břitvu penalizující složitost jádra.

9.2 Kovarianční jádra pro astrofyziku

Výběr kovariančního jádra kóduje naše fyzikální předpoklady o povaze hvězdné variability:

1. Kvadratické exponenciální jádro (Squared Exponential / RBF)

$$k_{\text{SE}}(r) = A^2 \exp\left(-\frac{r^2}{2L^2}\right), \quad r = |t - t'|$$

Popisuje nekonečně hladké, diferencovatelné procesy. V astronomii je často nerealistické, protože přirozené procesy vykazují drsnost na krátkých škálách.

2. Matérnovovo jádro (Matérn 3/2 a 5/2)

$$k_{\text{Matern52}}(r) = A^2 \left(1 + \frac{\sqrt{5}r}{L} + \frac{5r^2}{3L^2}\right) \exp\left(-\frac{\sqrt{5}r}{L}\right)$$

Ideální pro modelování granulace a konvektivních buněk ve fotosféře hvězd.

3. Kvazi-periodické jádro (Quasi-Periodic Kernel)

Hvězdné skvrny rotují s periodou rotace hvězdy P_{rot} , ale v čase se pomalu vyvíjejí a zanikají s charakteristickou dobou života τ_{decay} :

$$k_{\text{QP}}(r) = A^2 \exp\left(-\frac{r^2}{2\tau_{\text{decay}}^2} - \Gamma \sin^2\left(\frac{\pi r}{P_{\text{rot}}}\right)\right)$$

9.3 Rychlé škálovatelné výpočty: Algoritmus `celerite`

Standardní inverze matice $\mathbf{K} \in \mathbb{R}^{N \times N}$ a výpočet jejího determinantu vyžadují Choleského rozklad s výpočetní složitostí $\mathcal{O}(N^3)$. Pro světelnou křivku z mise Kepler s $N = 70\,000$ body by jediný krok MCMC trval několik hodin!

Daniel Foreman-Mackey a Eric Agol (2017) představili revoluční knihovnu **`celerite`** (a následně **`celerite2`**): Pokud je jádro vyjádřeno jako suma exponenciálně tlumených kosinů:

$$k(t_i, t_j) = \sum_{j=1}^J a_j \exp(-c_j |t_i - t_j|) \cos(d_j |t_i - t_j|)$$

Pak je kovarianční matice $\mathbf{K}$ semiseperabilní maticí s pásovou strukturou. Choleského rozklad takové matice lze provést v čase $\mathcal{O}(N \cdot J^2)$ namísto $\mathcal{O}(N^3)$!

Výpočet, který dříve trval hodiny, proběhne za 2 milisekundy. To umožnilo simultánní fitování desítek tisíc světelných křivek družic Kepler a TESS.

Kapitola 10: Stochastické autoregresní modely v astrofyzice

Ne všechny astronomické časové řady vykazují jasnou periodicitu. Akreční disky kolem černých děr, aktivní galaktická jádra (AGN) a blazary vykazují stochastické kolísání jasů známé jako **červený šum** (*Red Noise*) nebo *Flicker Noise*. K jejich modelování využíváme spojité stochastické diferenciální rovnice.

10.1 Spojitá autoregrese: Modely CARMA(p, q)

Klasické časové řady v ekonomii využívají diskrétní modely $ARMA(p, q)$. V astronomii kvůli nepravidelnému vzorkování musíme přejít ke spojitému ekvivalentu **CARMA** (*Continuous-time AutoRegressive Moving Average*).

Stacionární proces $CARMA(p, q)$ řádu p autoregrese a q klouzavých průměrů ($q < p$) je definován stochastickou diferenciální rovnicí:

$$\frac{d^p y(t)}{dt^p} + \alpha_{p-1} \frac{d^{p-1} y(t)}{dt^{p-1}} + \dots + \alpha_0 y(t) = \beta_q \frac{d^q W(t)}{dt^q} + \dots + \beta_1 \frac{dW(t)}{dt} + W(t)$$

Kde $W(t)$ je standardní bílý Gaussův proces (Brownův pohyb, $dW(t) \sim \mathcal{N}(0, dt)$).

Spektrální hustota výkonu (PSD) procesu CARMA

Fourierovou transformací diferenciálního operátoru získáme analytické vyjádření výkonového spektra:

$$P(f) = \sigma^2 \frac{|1 + \sum_{j=1}^q \beta_j (2\pi i f)^j|^2}{|\sum_{k=0}^p \alpha_k (2\pi i f)^k|^2}$$

Kde $\alpha_p = 1$. Spektrum procesu CARMA je racionální lomenou funkcí frekvence f^2 .

10.2 Tlumená náhodná procházka (DRW) a aktivní galaktická jádra

Nejjednodušším a v extragalaktické astronomii nejúspěšnějším modelem je **tlumená náhodná procházka** (**Damped Random Walk – DRW**), což je proces $CARMA(1, 0)$ neboli **Ornsteinův-Uhlenbeckův proces**:

$$dy(t) = -\frac{1}{\tau}y(t) dt + \sigma\sqrt{dt}\epsilon(t)$$

Kde:

- τ je charakteristická doba relaxace (damping timescale), typicky desítky až stovky dní pro supermasivní černé díry.
- σ je amplituda stochastického buzení.

Autokovarianční funkce DRW má jednoduchý exponenciální tvar:

$$k_{\text{DRW}}(r) = \frac{\tau\sigma^2}{2} \exp\left(-\frac{|r|}{\tau}\right)$$

Na nízkých frekvencích ($f \ll 1/\tau$) je spektrum ploché ($P(f) \propto \text{const}$ – bílý šum), zatímco na vysokých frekvencích ($f \gg 1/\tau$) klesá jako $P(f) \propto f^{-2}$ (červený šum). Analýza parametrů τ a σ ze světelných křivek přehlídek SDSS a ZTF umožňuje odhadnout hmotnost centrální supermasivní černé díry bez nutnosti nákladné spektroskopie.

Kapitola 11: Fotodynamické modelování tranzitních systémů

Když planeta tranzituje přes disk hvězdy, dochází nejen k poklesu toku v závislosti na geometrii, ale projevují se jemné astrofyzikální a orbitální efekty: gravitační interakce mezi více planetami (TTV) a ztemnění okraje hvězdného disku (Limb Darkening).

11.1 Analytický model tranzitu podle Mandela a Agola

Světelný tok během tranzitu závisí na profilu vyzařování hvězdného disku $I(r)$. Vlivem poklesu teploty směrem k vnějším vrstvám fotosféry je okraj disku tmavší než střed.

Nejrozšířenější je **kvadratický zákon ztemnění okraje**:

$$\frac{I(\mu)}{I(1)} = 1 - u_1(1 - \mu) - u_2(1 - \mu)^2, \quad \mu = \cos \theta = \sqrt{1 - r^2}$$

Kde $r \in [0, 1]$ je normalizovaná radiální vzdálenost od středu hvězdy a u_1, u_2 jsou koeficienty ztemnění okraje.

Knutson Mandel a Eric Agol (2002) odvodili přísně analytické řešení integrálu překryvu dvou disků pomocí eliptických integrálů prvního ($K(k)$), druhého ($E(k)$) a třetího druhu ($\Pi(n, k)$).

Relativní tok $F(p, z)$ závisí pouze na dvou bezrozměrných parametrech:

- $p = R_p/R_*$ (poměr poloměru planety a hvězdy)
- $z = d/R_*$ (normalizovaná vzdálenost středů obou těles na průmětně oblohy)

Díky analytickému řešení Mandel-Agolova modelu počítají moderní balíčky jako `batman` nebo `jaxoplanet` světelnou křivku tranzitu za mikrosekundy s přesností na úrovni 10^{-6} .

11.2 Variace časů tranzitů (TTV) a fotodynamické fitování

V systému s jedinou planetou obíhající kolem osamocené hvězdy jsou tranzity přísně periodické:

$$t_n = t_0 + n \cdot P$$

Pokud se však v systému nachází další planeta, vyvolává gravitační poruchy (perturbace) v dráze tranzitující planety. Tyto poruchy způsobují, že tranzity nastávají o minuty až hodiny dříve či později, než předpovídá lineární efemerida:

$$O - C = t_{\text{obs}} - t_{\text{calc}} \neq 0$$

Tento jev nazýváme **Transit Timing Variations (TTV)**.

Význam TTV pro astroinformatiku

- **Detekce netranzitujících planet:** Planeta, která z pohledu Země netranzituje, může být jednoznačně odhalena gravitačním vlivem na tranzitující planetu.
- **Přímé měření hmotností:** Z pouhého tranzitu známe pouze poloměr planety R_p . Z amplitudy a fáze TTV lze přímo spočítat skutečné hmotnosti obou planet M_p a odvodit jejich průměrnou hustotu a chemické složení (např. v soustavě TRAPPIST-1).
- **Fotodynamické fitování:** Místo fitování analytické Keplerovy elipsy integruje software v každém kroku MCMC přímo pohybové rovnice N -těles pomocí symplektického integrátoru (např. algoritmus WHFast v balíčku `REBOUND`).

Část IV: Strojové učení: Pokročilá regrese, klasifikace a shlukování

Čtvrtá část rozebírá pokročilé techniky strojového učení aplikované na astronomická data. Probereme L1 a L2 regularizaci v mnohorozměrných spektrech, souborové stromy XGBoost a LightGBM pro fotometrický red-shift, nelineární manifold learning (t-SNE, UMAP) a shlukové algoritmy v 6D fázovém prostoru astrometrie Gaia.

Kapitola 12: Regularizace a optimalizace v Data Science

Při analýze vysokorozměrných astrofyzikálních dat (např. spektra s $p = 5\,000$ vlnovými délkami pro $n = 1\,000$ hvězd) čelíme fundamentálnímu problému: počet parametrů převyšuje počet pozorování ($p > n$). V takových podmínkách má matice $\mathbf{X}^T \mathbf{X}$ nulový determinant a standardní regrese kolabuje vlivem drastického přeučení.

12.1 Kompromis mezi vychýlením a rozptylem (Bias–Variance Tradeoff)

Očekávaná střední kvadratická chyba predikce libovolného modelu $f(\mathbf{x})$ se rozkládá na tři fundamentální složky:

$$\mathbb{E} \left[(y - \hat{f}(\mathbf{x}))^2 \right] = \underbrace{\text{Bias}^2[\hat{f}(\mathbf{x})]}_{\text{Vychýlení}} + \underbrace{\text{Var}[\hat{f}(\mathbf{x})]}_{\text{Rozptyl}} + \underbrace{\sigma^2}_{\text{Neredukovatelný šum}}$$

- **Vysoké vychýlení (Underfitting):** Model je příliš rigidní (např. prokládání přímky kvadratickými daty). Má nízkou variabilitu, ale systematicky mýlí realitu.
- **Vysoký rozptyl (Overfitting):** Model je příliš flexibilní (např. polynom 20. stupně). Perfektně prochází všemi trénovacími body včetně šumu, ale na nových datech fatálně selhává.

Regularizace záměrně zavádí malé vychýlení výměnou za masivní redukci rozptylu predikce.

12.2 L1 a L2 regularizace: Ridge, LASSO a ElasticNet

Uvažujme lineární regresní model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Regularizovaná ztrátová funkce má tvar:

$$\mathcal{L}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \Omega(\beta)$$

1. Ridge regrese (L2 regularizace, Tichonov)

$$\Omega_{\text{Ridge}}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$$

Penalizuje velké hodnoty vah kvadraticky. Zmenšuje koeficienty směrem k nule, ale žádný koeficient nenastaví přesně na nulu. Analytické řešení: $\hat{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$.

2. LASSO regrese (L1 regularizace, Tibshirani 1996)

$$\Omega_{\text{LASSO}}(\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

Geometricky tvoří L_1 koule mnohostěn s ostrými hroty na osách souřadnic. Díky tomu LASSO provádí **automatický výběr příznaků (Feature Selection)**: nastavuje koeficienty irelevantních vlnových délek přesně na nulu! Výsledný model je řídký (*sparse*).

3. ElasticNet

$$\Omega_{\text{ElasticNet}}(\beta) = \alpha \|\beta\|_1 + \frac{1 - \alpha}{2} \|\beta\|_2^2$$

Kombinuje obě metody. V astrofyzice je ideální, pokud jsou příznaky silně korelované (např. sousední vlnové délky ve spektrální absorpční čáře). LASSO by náhodně vybralo jen jeden pixel, zatímco ElasticNet vybere celou skupinu korelovaných vlnových délek společně.

Kapitola 13: Gradient Boosted Trees a náhodné lesy

Zatímco neuronové sítě dominují nezpracovaným obrazovým a zvukovým datům, pro **tabulková astrofyzikální data** (katalogy fotometrických barev, morfologické parametry, astrometrické indexy) představují **souborové metody rozhodovacích stromů** absolutní světovou špičku v přesnosti a efektivitě.

13.1 Náhodné lesy (Random Forests)

Rozhodovací strom rozděluje prostor příznaků pravoúhlými řezy. V každém uzlu hledá atribut x_j a práh t , který maximalizuje pokles nečistoty (Giniho index pro klasifikaci nebo střední kvadratickou odchylku pro regresi):

$$\Delta I = I_{\text{parent}} - \frac{N_L}{N} I_{\text{left}} - \frac{N_R}{N} I_{\text{right}}$$

Jeden strom je náchylný k přeučení. **Náhodný les (Leo Breiman, 2001)** kombinuje stovky stromů pomocí dvou principů náhodnosti:

1. **Bootstrap Aggregating (Bagging)**: Každý strom trénuje na náhodném výběru pozorování s opakováním.
2. **Feature Subsampling**: V každém uzlu vybírá algoritmus nejlepší řez pouze z náhodné podmnožiny $m \approx \sqrt{p}$ příznaků.

Díky de-korelaci stromů klesá rozptyl celkového souboru úměrně $1/K$, kde K je počet stromů, aniž by rostlo vychýlení!

13.2 Gradient Boosting (XGBoost, LightGBM a CatBoost)

Zatímco stromy v náhodném lese rostou nezávisle vedle sebe, **Gradient Boosting** staví stromy sekvencně: každý nový strom se učí opravovat reziduální chyby předchozího souboru.

Mějme v kroku $m - 1$ soubor $F_{m-1}(\mathbf{x})$. V kroku m spočítáme záporný gradient ztrátové funkce (pseudo-rezidua):

$$r_{im} = - \left[\frac{\partial \mathcal{L}(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F=F_{m-1}}$$

Nový strom $h_m(\mathbf{x})$ natrénujeme na predikci těchto reziduí a aktualizujeme model:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x})$$

Kde $\eta \in (0, 1]$ je rychlost učení (Learning Rate).

- **XGBoost (Chen & Guestrin, 2016):** Využívá Taylorův rozvoj ztrátové funkce do druhého řádu (gradient i Hessian) a explicitní regularizaci složitosti stromů.
- **LightGBM (Microsoft):** Roste podle listů (Leaf-wise) namísto podle úrovní a diskretizuje příznaky do histogramových košů, což zrychluje trénování na miliardových katalozích až 20–krát.

Aplikace: Fotometrický odhad rudého posuvu (Photo-z)

Pro miliardy galaxií v přehlídkách LSST a Euclid nelze pořádit spektra. Známe pouze toky v několika širokopásmových filtrech (u, g, r, i, z, y). Gradient Boosted Trees dokáží z těchto 6 barev předpovědět kosmologický rudý posuv z s přesností $\sigma_z / (1 + z) < 0,015$!

Kapitola 14: Pokročilá redukce dimenzionality a nelineární vnoření

Astrofyzikální objekty jsou reprezentovány ve vysokorozměrných prostorech (např. 1D spektrum s 4 000 pixely). Skutečná vnitřní dimenzionalita fyzikálního procesu (např. efektivní teplota T_{eff} , gravitace $\log g$, metalicita $[\text{Fe}/\text{H}]$ a mikroturbulence) je však dramaticky nižší ($D \approx 3$ až 5). Cílem **redukce dimenzionality** je nalézt tuto nízkorozměrnou varietu (*manifold*).

14.1 Funkční analýza hlavních komponent (Functional PCA)

Klasická metoda PCA zachází se spektrem jako s diskrétním vektorem čísel a ignoruje, že vlnová délka je spojitá fyzikální osa. **Funkční PCA (FPCA)** modeluje spektrum jako spojitou funkci $X_i(\lambda) \in L^2[\lambda_{\min}, \lambda_{\max}]$:

$$X_i(\lambda) = \mu(\lambda) + \sum_{k=1}^K \xi_{ik} \psi_k(\lambda) + \epsilon_i(\lambda)$$

Kde $\psi_k(\lambda)$ jsou ortonormální spojitě vlastní funkce kovariančního integrálního operátoru:

$$\int K(\lambda, \lambda') \psi_k(\lambda') d\lambda' = \lambda_k \psi_k(\lambda)$$

A koeficienty $\xi_{ik} = \int (X_i(\lambda) - \mu(\lambda)) \psi_k(\lambda) d\lambda$ jsou hlavní funkční skóre. FPCA je přirozeně imunní vůči mezerám ve spektrech a různému rozlišení různých spektrografů.

14.2 Nelineární manifold learning: t-SNE a UMAP

Lineární metody jako PCA selhávají, pokud je fázový prostor zakřivený (např. spirální struktura nebo nelineární termonukleární evoluce hvězd v Hertzsprung–Russellově diagramu).

1. t-SNE (t-Distributed Stochastic Neighbor Embedding)

Převádí eukleidovské vzdálenosti mezi body ve vysokorozměrném prostoru na podmíněné pravděpodobnosti sousedství:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

V nízkorozměrném prostoru ($2D$) modeluje vzdálenost pomocí Studentova t -rozdělení s 1 stupněm volnosti (Cauchyho rozdělení), což eliminuje problém shlukování bodů ve středu (*Crowding Problem*):

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

Optimalizace minimalizuje Kullback-Leiblerovu divergenci $D_{\text{KL}}(P||Q) = \sum_{i,j} p_{ij} \ln(p_{ij}/q_{ij})$.

2. UMAP (Uniform Manifold Approximation and Projection)

Založený na teorii Riemannovy geometrie a algebraické topologie (fuzzy simplicialní komplexy).

- Je matematicky rigoróznější než t-SNE.
- Zachovává nejen lokální sousedství, ale i **globální topologickou strukturu** dat.
- Je řádově rychlejší a umožňuje promítání nových dosud neviděných dat do existujícího 2D prostoru.

Kapitola 15: Prostorové shlukování ve fázovém prostoru

Astrometrická družice Gaia měří pro stovky milionů hvězd šesti-rozměrný fázový prostor: polohu na obloze (α, δ) , vzdálenost z paralaxy $d = 1/\varpi$, vlastní pohyby $(\mu_\alpha^*, \mu_\delta)$ a radiální rychlost v_r . Hledání gravitačně vázaných struktur (otevřených hvězdokup, asociací a rozpadlých trpasličích galaxií) vyžaduje pokročilé shlukové algoritmy.

15.1 Algoritmus HDBSCAN s proměnnou hustotou

Klasický algoritmus k -means předpokládá sférické shluky stejné velikosti a DBSCAN vyžaduje globální parametr hustoty ϵ , což je v astronomii nepoužitelné – hustota hvězdného pole v rovině Galaxie je tisíckrát vyšší než v galaktickém halu.

HDBSCAN (Hierarchical DBSCAN – Campello et al., 2013) překonává tato omezení:

1. Transformuje prostor pomocí **vzájemné dosažitelnosti (Mutual Reachability Distance)**:

$$d_{\text{mreach-}k}(a, b) = \max \{ \text{core}_k(a), \text{core}_k(b), d(a, b) \}$$

Kde $\text{core}_k(x)$ je vzdálenost k -tému nejbližšímu sousedovi.

2. Sestrojí minimální kostru grafu (Minimum Spanning Tree) a převede ji na hierarchický dendrogram shluků.
3. Kondenzuje strom podle minimální velikosti shluku $N_{\min}$.
4. Vybere stabilní shluky integrací jejich životnosti v diagramu persistence: $\lambda = 1/\epsilon$.

Body s nízkou stabilitou jsou automaticky označeny jako **šumové pozadí**.

15.2 Objevování hvězdných proudů a reliků Gaia–Enceladus

Když trpasličí galaxie spadne do gravitačního potenciálu Mléčné dráhy, slapové síly ji roztrhají na dlouhé filamentární útvary – **hvězdné proudy (Stellar Streams)**. V reálném prostoru (x, y, z) jsou tyto hvězdy rozptýleny přes celou oblohu a nelze je rozeznat od běžných hvězd disku.

V prostoru integrálů pohybu (energie E a složek momentu hybnosti $L_z, L_\perp$) však zůstávají tyto hvězdy navždy shluknuté v kompaktních strukturách!

Aplikací algoritmů HDBSCAN a Gaussovských směsných modelů (GMM) na data Gaia DR2 a DR3 objevili astronomové (Helmi et al., 2018; Belokurov et al., 2018) relik gigantické srážky – galaxií

Gaia–Enceladus (Gaia Sausage), která se s naší Galaxií sloučila před 10 miliardami let a tvoří většinu vnitřního galaktického hala.

Část V: Hluboké neuronové sítě (Deep Learning) v astrofyzice

Pátá část otevírá svět moderní umělé inteligence. Zkoumá optimalizaci hlubokých architektur, konvoluční sítě (CNN) při detekci silných gravitačních čoček, Vision Transformery na multispektrálních obrazech, difúzní generativní modely pro super-rozlišení a fyzikálně informované sítě (PINN).

Kapitola 16: Architektury a optimalizace hlubokých sítí

Zatímco klasické strojové učení vyžadovalo ruční extrakci příznaků (Feature Engineering), **hluboké učení (Deep Learning)** se učí reprezentace přímo ze surových dat (pixelů, spektrálních toků).

16.1 Vícevrstvý perceptron a aktivační funkce

Vícevrstvý perceptron (MLP) skládá afinní transformace proložené nelineárními aktivačními funkcemi σ :

$$\mathbf{h}^{(l)} = \sigma \left(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)} \right)$$

Zatímco historická funkce $\text{ReLU}(x) = \max(0, x)$ trpí problémem „umírajících neuronů“ (nulový gradient pro $x < 0$), v moderní astrofyzice dominují hladké nelinearity:

- **GELU (Gaussian Error Linear Unit):**

$\text{GELU}(x) = x \Phi(x) \approx 0,5x(1 + \tanh(\sqrt{2/\pi}(x + 0,044715x^3)))$. Standard v transformerových architekturách.

- **Swish / SiLU:** $\text{Swish}(x) = x \cdot \text{sigmoid}(\beta x)$. Vykazuje lepší gradientní toky v hlubokých sítích.

16.2 Optimalizátory: AdamW a Cosine Annealing

Optimalizace ztrátové funkce $\mathcal{L}(\mathbf{W})$ v prostorech s miliony parametrů vyžaduje adaptivní stochastické metody.

Algoritmus **AdamW (Loshchilov & Hutter, 2019)** opravuje fundamentální chybu původního Adamu v implementaci L2 regularizace. Odděluje pokles vah (Weight Decay) od adaptivního odhadu momentů:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t, \quad \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2$$

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta_t \left(\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} + \lambda \mathbf{W}_{t-1} \right)$$

Společně s **kosinovým plánem učení (Cosine Annealing)** s fází zahřívání (Warmup):

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos \left(\frac{t}{T_{\max}} \pi \right) \right)$$

zabraňuje uvíznutí v mělkých lokálních minimech a dosahuje vynikající zobecňovací schopnosti.

Kapitola 17: Konvoluční neuronové sítě (CNN) a segmentace

Konvoluční neuronové sítě (CNN) využívají fundamentální symetrie obrazových dat: translační invarianci a lokální prostorovou korelaci pixelů.

17.1 Reziduální architektury (ResNet)

Při zvyšování hloubky klasických konvolučních sítí dochází k degradaci gradientu: hlubší síť vykazuje vyšší trénovací chybu než mělká. Kaiming He et al. (2016) vyřešili tento problém zavedením **reziduálních spojení (Skip Connections)**:

$$\mathbf{x}_{l+1} = \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l) + \mathbf{x}_l$$

Vrstva se neučí přímo cílové zobrazení $\mathcal{H}(\mathbf{x})$, nýbrž reziduální odchylku $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$. Gradient může proudit přímo přes jednotkovou identitu bez jakéhokoliv tlumení napříč stovkami vrstev.

17.2 Architektura U-Net a detekce gravitačních čoček

Pro sémantickou segmentaci difúzních mlhovin a detekci slabých gravitačních oblouků se využívá architektura **U-Net** (Ronneberger et al., 2015).

Skládá se ze dvou větví:

1. **Kontrakční větev (Encoder)**: Postupně zmenšuje prostorové rozlišení pomocí konvolucí a max-poolingu, zatímco roste počet příznakových kanálů. Zachycuje globální kontext.
2. **Expanzní větev (Decoder)**: Zvětšuje prostorové rozlišení (transponované konvoluce / upsampling).
3. **Příčná přeskočení (Skip Connections)**: Přenášejí vysokofrekvenční prostorové detaily přímo z kodéru do dekodéru na odpovídající úrovni rozlišení.

Aplikace v misi Euclid

V datech vesmírného dalekohledu Euclid skenuje síť U-Net miliony galaktických kup a v reálném čase segmentuje nepatrně prohnuté oblouky pozadních galaxií deformované temnou hmotou v popředí (**Einsteinovy prstence**). To, co dříve vyžadovalo týdny vizuální kontroly, probíhá rychlostí 50 snímků za sekundu na jednom GPU.

Kapitola 18: Transformery a sekvenční modely pro astronomická data

Mechanismus samopozornosti (**Self-Attention**), představený v roce 2017 (Vaswani et al.), způsobil revoluci v analýze sekvencí a v podobě **Vision Transformerů (ViT)** překonává klasické konvoluční sítě i v počítačovém vidění.

18.1 Matematika mechanismu Multi-Head Attention

Mějme sekvenci vstupních vektorů $\mathbf{X} \in \mathbb{R}^{N \times d}$. Vynásobením trénovatelnými maticemi vah $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d_k}$ získáme matice dotazů ($\mathbf{Q}$), klíčů ($\mathbf{K}$) a hodnot ($\mathbf{V}$):

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V$$

Skalární součinná pozornost (Scaled Dot-Product Attention) je definována vztahem:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V}$$

Člen $\frac{1}{\sqrt{d_k}}$ zabraňuje saturaci funkce softmax pro vysoké dimenze.

V **Multi-Head Attention** se tento výpočet provádí paralelně v h nezávislých hlavách pozornosti, což modelu umožňuje sledovat vztahy na různých fyzikálních škálách současně:

- Jedna hlava sleduje krátkoperiodické fluktuace granule.
- Jiná hlava monitoruje dlouhodobý trend způsobený rotací magnetických skvrn.

18.2 Vision Transformers (ViT) na astronomických snímcích

Zatímco CNN má lokální receptivní pole omezené velikostí konvolučního jádra (3×3 pixely), Vision Transformer rozdělí astronomický snímek na matici nespojitých dlaždic (např. 16×16 pixelů), každou dlaždici lineárně promítne na vektor a zpracuje ji jako token v sekvenci.

Díky globální samopozornosti dokáže ViT v jediné vrstvě porovnat záření v jádru galaxie se strukturou vnějších slapových ramen vzdálených stovky pixelů. Na rozsáhlých přehlídkách dosahují ViT vyšší přesnosti v klasifikaci morfologie galaxií než jakákoliv konvoluční síť.

Kapitola 19: Generativní modely v astronomickém zobrazování

Zatímco diskriminační modely počítají $p(y|\mathbf{x})$, **generativní modely** se učí modelovat sdružené rozdělení pravděpodobnosti dat $p(\mathbf{x})$. Umožňují rekonstruovat zašuměná měření, provádět super-rozlišení a syntetizovat realistické vesmírné struktury.

19.1 Variační autoenkodéry (VAE)

VAE (Kingma & Welling, 2013) kóduje vstupní obraz $\mathbf{x}$ do parametrů pravděpodobnostního rozdělení v latentním prostoru $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\mathbf{x}, \Sigma_\mathbf{x})$.

Trénovací kritérium maximalizuje **dolní mez variační evidence (ELBO)**:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}))$$

První člen měří věrnost rekonstrukce, zatímco druhý člen nutí latentní reprezentaci odpovídat standardnímu gaussovskému rozdělení $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Spojité latentní prostoru VAE umožňuje plynulou interpolaci mezi různými typy galaxií a identifikaci fyzikálních přechodových stavů.

19.2 Difúzní pravděpodobnostní modely (Score-based Diffusion)

Difúzní modely (Sohl-Dickstein et al., 2015; Song et al., 2020) představují současný absolutní vrchol generativního modelování.

Proces sestává ze dvou fází popsaných stochastickými diferenciálními rovnicemi (SDE):

1. **Dopředný proces (Forward process)**: Postupně přidává ke snímku gaussovský šum, až se obraz zcela promění v bílý šum:

$$d\mathbf{x} = f(\mathbf{x}, t) dt + g(t) d\mathbf{w}$$

2. **Zpětný proces (Reverse process)**: Časově obrácená SDE umožňuje z čistého náhodného šumu rekonstruovat původní signál:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t) d\bar{\mathbf{w}}$$

Neuronová síť se učí odhadovat **skóre funkci** $s_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$.

V astroinformatice difúzní modely excelují v **dekonvoluci a super-rozlišení**: z rozmazaného snímku z pozemního dalekohledu dokáží odstranit atmosférické rozostření a rekonstruovat jemné detaily odpovídající kosmickému teleskopu.

Kapitola 20: Fyzikálně informované neuronové sítě (PINN)

Běžné neuronové sítě fungují jako „černé skříňky“ – optimalizují pouze shodu s trénovacími daty a mohou predikovat fyzikálně nemožné stavy (porušení zachování energie, záporné hustoty, nadsvětelné rychlosti). **Fyzikálně informované sítě (Physics-Informed Neural Networks – PINN)** začleňují fundamentální diferenciální rovnice přímo do trénovacího procesu.

20.1 Matematický rámec PINN

Uvažujme nelineární parciální diferenciální rovnici (PDE) popisující astrofyzikální proces:

$$\mathcal{D}[u(\mathbf{x}, t)] = f(\mathbf{x}, t), \quad \mathbf{x} \in \Omega, \quad t \in [0, T]$$

Neuronová síť $u_\theta(\mathbf{x}, t)$ s vahami θ aproximuje neznámé řešení. Využitím **automatické diferenciace (Automatic Differentiation)** dokáže počítač spočítat přesné derivace sítě podle prostorových souřadnic $\mathbf{x}$ a času t na strojovou přesnost (bez aproximačních chyb konečných diferencí).

Celková ztrátová funkce PINN kombinuje datovou a fyzikální složku:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{data}}(\theta) + \lambda_{\text{PDE}} \mathcal{L}_{\text{PDE}}(\theta) + \lambda_{\text{BC}} \mathcal{L}_{\text{BC}}(\theta)$$

Kde reziduum diferenciální rovnice je vyhodnocováno na souboru kolokačních bodů $\{\mathbf{x}_j, t_j\}_{j=1}^{N_c}$:

$$\mathcal{L}_{\text{PDE}}(\theta) = \frac{1}{N_c} \sum_{j=1}^{N_c} \|\mathcal{D}[u_\theta(\mathbf{x}_j, t_j)] - f(\mathbf{x}_j, t_j)\|^2$$

20.2 Aplikace: Hydrostatická rovnováha a rychlé neuronové emulátory

- Struktura hvězdných fotosfér:** PINN řeší simultánně rovnici hydrostatické rovnováhy $\frac{dP}{dr} = -\frac{GM(r)\rho(r)}{r^2}$ a rovnici přenosu záření. Síť nalezne řešení za zlomky sekundy, což umožňuje fitovat miliony spekter v reálném čase.
- Kosmologické neuronové emulátory:** Plné hydrodynamické simulace vzniku galaxií a temné hmoty vyžadují miliony procesorových hodin na superpočítačích. PINN natrénovaná jako emulátor dokáže předpovědět nelineární výkonové spektrum hmoty $P(k)$ pro libovolnou kombinaci kosmologických parametrů ($\Omega_m, \Omega_b, \sigma_8, n_s, w_0$) během 5 milisekund s chybou pod 0,5 %.

Část VI: Petabajtové přehlídky, Big Data a distribuované systémy

Šestá část analyzuje architekturu moderních velkoobjemových přehlídek. Zabývá se prostorovou pixelizací nebeské sféry (HEALPix), sloupcovými analytickými databázemi (Parquet, Arrow, DuckDB), distribuovaným zpracováním dat na výpočetních clusterech (Dask, Ray) a masivně paralelní GPU akcelerací.

Kapitola 21: Architektura petabajtových přehlídek nové generace

Vstupujeme do éry, kdy astronomická data nelze uchovávat na lokálních discích a vědecké dotazy nelze provádět stažením souborů na vlastní počítač. Zkoumání vesmíru se přesouvá k principu „**přived' kód k datům**“ (*Code-to-Data Paradigm*).

21.1 Vera C. Rubin Observatory (LSST)

Observatoř Věry Rubinové v Chile představuje vrchol moderního inženýrství velkých dat:

- **Teleskop Simonyi Survey:** Třízrcadlový anastigmat (průměr primárního zrcadla 8,4 m) se zorným polem $3,5^\circ$ ($9,6 \text{ deg}^2$, což odpovídá ploše 40 měsíčních úplňků).
- **Kamera LSST:** Největší digitální kamera v dějinách lidstva. Má 3,2 miliardy pixelů chlazených na -100°C a její ohnisková rovina se skládá ze 189 vědeckých CCD senzorů.
- **Datový objem:** Každých 37 sekund pořídí dalekohled dva 15sekundové snímky. Každou noc vygeneruje zhruba 20 terabajtů surových dat, což během 10 let představuje přes **60 petabajtů surových obrazů** a finální katalog o velikosti **15 petabajtů tabulkových dat** s 37 miliardami objektů.

21.2 Alertové toky v reálném čase: Apache Kafka v astronomii

Zásadní inovací LSST je schopnost detekovat proměnné a transientní jevy v reálném čase:

1. Každý nově pořízený snímek je do 60 sekund zkalibrován a digitálně odečten od hlubokého referenčního šablonového snímku téže oblasti oblohy (**Image Differencing**).
2. Na rozdílovém snímku algoritmy detekují všechny body, kde došlo ke změně jasu nebo polohy s významností $> 5\sigma$.
3. Pro každou takovou událost vygeneruje systém **světelný alert** ve formátu Apache Avro obsahující souřadnice, fotometrii, výřez snímku (30×30 pixelů) a historickou světelnou křivku.

Během jediné noci systém vygeneruje **až 10 milionů alertů**! Tyto zprávy jsou v reálném čase streamovány distribuovaným systémem **Apache Kafka** do komunitních brokerů (např. *ALeRCE*, *Lasair*, *Fink*, *Antares*), kde strojové učení do několika sekund klasifikuje supernovy, blízkozemní asteroidy a optické protějšky gravitačních vln a automaticky spouští návazná pozorování robotickými teleskopy po celé zeměkouli.

Kapitola 22: Sférické indexování a masivní prostorové dotazování

Rektascenze α a deklinace δ tvoří souřadnice na zakřivené kulové ploše S^2 . Standardní databázové indexy (B-stromy) selhávají na pólech a při výpočtu úhlových vzdáleností. Pro efektivní vyhledávání na sféře musíme využít specializované hierarchické pixelizační systémy.

22.1 HEALPix (Hierarchical Equal Area isoLatitude Pixelization)

Systém **HEALPix** (Górski et al., 2005), původně vyvinutý pro analýzu reliktního záření z mise Planck, je globálním standardem sférické astroinformatiky.

Klíčové vlastnosti HEALPix:

1. **Přísně stejná plocha:** Všechny pixely na dané úrovni dělení mají naprosto identickou plochu. Integrály přes sféru se redukují na prosté sumy přes pixely bez nutnosti vážení!
2. **Izo-šířková distribuce:** Středy pixelů leží na izolovaných rovnoběžkách stálé deklinace, což zásadně urychluje výpočet sférických harmonických funkcí $Y_{\ell m}$.

Dělení sféry začíná 12 základními kosočtvercovými bázovými pixely. Každý pixel je hierarchicky dělen na 4 menší pixely. Úroveň rozlišení určuje parametr $N_{\text{side}} = 2^k$, celkový počet pixelů je:

$$N_{\text{pix}} = 12 N_{\text{side}}^2$$

Pro $N_{\text{side}} = 2^{12} = 4096$ má sféra přes 200 milionů pixelů s úhlovým rozlišením $51,5''$.

22.2 Prostorový cross-matching v miliardách zdrojů

Častou úlohou je ztotožnit objekty ze dvou obřích katalogů (např. Gaia DR3 s 1,8 miliardami hvězd a AllWISE s 747 miliony infračervených zdrojů) v zadaném poloměru $r < 1''$.

Naivní algoritmus vyžaduje $O(N_1 \cdot N_2) \approx 10^{18}$ porovnání, což je neproveditelné.

Moderní algoritmy využívají:

- **HEALPix indexování:** Objekty jsou rozděleny do košíků podle indexu pixelu. Porovnávají se pouze body ležící v tomtéž pixelu a jeho 8 bezprostředních sousedech.
- **3D kartézské kd-stromy:** Sférické souřadnice (α, δ) jsou převedeny na 3D jednotkové vektory:

$$x = \cos \delta \cos \alpha, \quad y = \cos \delta \sin \alpha, \quad z = \sin \delta$$

Úhlová vzdálenost θ odpovídá eukleidovské tětivové vzdálenosti $d = 2 \sin(\theta/2)$. Vyhledávání v binárním stromu zredukuje složitost na $\mathcal{O}(N_1 \log N_2)$, takže cross-match miliardových katalogů proběhne na moderní pracovní stanici za méně než 15 minut.

Kapitola 23: Sloupcové formáty a moderní analytické databáze

Po více než 40 let byl formát FITS králem astronomických tabulek. V éře katalogů s miliardami řádků a stovkami atributů však tradiční řádkově orientované formáty a relační SQL databáze (PostgreSQL) narážejí na hardwarové limity.

23.1 Sloupcové formáty: Apache Parquet a Arrow

Při typické astrofyzikální analýze nás zajímá pouze několik sloupců z celkového počtu stovek parametrů (např. pouze rektascenze, deklinace a paralaxa pro výpočet vzdálenosti).

- **Řádkový formát (FITS, CSV):** Data jsou na disku uložena po celých řádcích: $[x_1, y_1, z_1, mag_1, \dots], [x_2, y_2, z_2, mag_2, \dots]$. Pro přečtení jednoho sloupce musí disk načíst všechna data všech atributů.
- **Sloupcový formát (Columnar – Apache Parquet):** Každý sloupec je uložen samostatně v souvislém bloku bajtů na disku: $[x_1, x_2, \dots, x_N], [y_1, y_2, \dots, y_N]$.

Výhody formátu Parquet v astroinformatice:

1. **Projekční pruning:** Dotaz čte z disku výhradně ty sloupce, které jsou uvedeny v klauzuli ``SELECT``. I/O zátěž klesá o 90 až 95 %!
2. **Extrémní komprese:** Sloupec téhož datového typu (např. float32 paralaxy) vykazuje vysokou redundanci. Algoritmy jako Snappy nebo Zstandard dosahují kompresního poměru 4 : 1 až 10 : 1.
3. **Statistiky bloků:** Každá skupina řádků (*Row Group*) obsahuje v metadatech minimální a maximální hodnotu. Pokud dotaz obsahuje podmínku ``WHERE mag < 15``, celé bloky s minimem > 15 jsou okamžitě přeskočeny bez čtení z disku.

23.2 Moderní in-memory OLAP databáze: DuckDB a ClickHouse

Pro lokální i clusterovou analýzu petabajtových tabulek nahrazují tradiční SQL servery moderní **vektorizované OLAP databáze**:

- **DuckDB:** „SQLite pro analytická data“. Běží přímo uvnitř procesu Pythonu bez nutnosti instalace databázového serveru. Dokáže dotazovat Parquet soubory na disku nebo v cloudu (Amazon S3) plně paralelně s využitím SIMD instrukcí procesoru:

```
import duckdb

<h1>Analýza 100 milionů hvězd z Parquet souboru za zlomek sekundy</h1>
sql_query = \"\"\"
    SELECT
        floor(ra/10)*10 AS ra_bin,
        avg(phot_g_mean_mag) AS mean_g,
        count(*) AS n_stars
    FROM 'gaia_dr3_source_*.parquet'
    WHERE parallax > 10.0 AND parallax_over_error > 5.0
    GROUP BY ra_bin
    ORDER BY ra_bin
\"\"\"
res = duckdb.query(sql_query).df()
```

- **ClickHouse:** Masivně distribuovaná sloupcová databáze schopná agregovat miliardy řádků za sekundu na distribuovaných serverech.

Kapitola 24: Distribuované výpočty v Pythonu

Když objem astrofyzikálních dat nebo výpočetní náročnost simulace překročí limity jednoho počítače, musíme výpočet distribuovat přes výpočetní uzel (cluster) se stovkami procesorových jader a terabajty distribuované paměti.

24.1 Škálování s knihovnou Dask

Knihovna **Dask** poskytuje nativní paralelní a distribuované výpočty v Pythonu s rozhraním, které věrně kopíruje NumPy a Pandas:

1. **Dask Array**: Rozděluje multidimenzionální pole na menší bloky (*chunks*). Umožňuje provádět operace lineární algebry nad poli o velikosti stovek gigabajtů, která se nevejdou do RAM:

```
import dask.array as da <h1>Virtuální pole 100 000 x 100 000 prvků rozdělené na bloky 5000 x 5000</h1> x = da.from_zarr('giant_simulation_cube.zarr') mean_profile = x.mean(axis=(1, 2)).compute() # Paralelní vyhodnocení
```

2. **Dask DataFrame**: Paralelní tabulkový procesor skládající desítky Pandas datových rámců.
3. **Dask Distributed**: Distribuovaný plánovač úloh (Scheduler), který dynamicky řídí tok grafu závislostí úloh (DAG) přes klastry řízené systémy Slurm, PBS nebo Kubernetes.

24.2 Knihovna Ray pro masivní paralelismus úloh

Zatímco Dask je optimalizován pro strukturovaná tabulková a maticová data, knihovna **Ray** (UC Berkeley) představuje univerzální distribuovaný výpočetní framework nové generace:

- **Stateless úlohy (Tasks)**: Asynchronní distribuované funkce spouštěné dekorátorem `@ray.remote`.
- **Stateful aktéři (Actors)**: Distribuované třídy zachovávající stav v paměti (např. běžící orbitální simulace nebo instance neuronové sítě).
- **Ray Tune**: Špičkový nástroj pro masivně distribuované ladění hyperparametrů bayesovských a neuronových modelů.

Kapitola 25: Akcelerace na GPU a tenzorové simulace

Grafické procesory (GPU) disponují tisíci specializovanými početními jádry schopnými masivně paralelního provádění operací s plovoucí řádovou čárkou. V astroinformatice přináší přesun výpočtů z CPU na GPU zrychlení o 2 až 3 řády ($100 \times$ až $1\,000 \times$).

25.1 Ekosystém CuPy: Okamžitá GPU akcelerace

Knihovna **CuPy** je drop-in náhradou za NumPy na grafických kartách s architekturou NVIDIA CUDA:

```
import cupy as cp

<h1>Alokace a maticový součin na GPU v čase zlomků milisekund</h1>
A_gpu = cp.random.randn(5000, 5000, dtype=cp.float32)
B_gpu = cp.random.randn(5000, 5000, dtype=cp.float32)
C_gpu = cp.matmul(A_gpu, B_gpu)
cp.cuda.Stream.null.synchronize()
```

Klíčové výhody:

- Identické API jako NumPy (`cp.fft.fft`, `cp.linalg.svd`, `cp.sum`).
- Vlastní kompilace CUDA jader pomocí `cp.RawKernel` pro specifické astronomické operace (např. výpočet fázových úhlů nebo detekce zdrojů na snímcích).

25.2 Diferenciální astrofyzika s knihovnou JAX

Knihovna **JAX** (Google Research) spojuje automatické derivování podle Autogradu s optimalizujícím kompilátorem **XLA (Accelerated Linear Algebra)** pro GPU a TPU:

1. **jax.grad**: Automaticky derivuje jakoukoliv matematickou funkci v Pythonu do libovolného řádu. Zcela eliminuje nutnost ručního odvozování gradientů pro HMC vzorkovače nebo optimalizátory!
2. **jax.jit**: Just-In-Time kompilace Python kódu do vysoce optimalizovaného strojového kódu GPU. Sloučí sérii maticových operací do jediného GPU kernelu (Kernel Fusion), čímž eliminuje

paměťová úzká hrdla.

3. **`jax.vmap`**: Automatická vektorizace funkcí přes dávky (batches) bez nutnosti ručního přepisování cyklů.

Moderní knihovny jako **`jaxoplanet`** počítají tranzitní světelné křivky a Keplerovy dráhy pro tisíce planet současně na GPU rychlostí milionů vyhodnocení za sekundu, což umožňuje plnohodnotné globální bayesovské fitování celých přehlídek.

Část VII: Přílohy, repetitoria a sbírka vědeckých algoritmů

Závěrečná část monografie poskytuje referenční zázemí: detailní matematická odvození fundamentálních vztahů, programátorský tahák moderních vědeckých balíčků, sbírku 15 plně komentovaných výpočetních úloh v Pythonu a encyklopedický glosář pokročilé astroinformatiky.

Dodatek A: Matematická odvození klíčových vztahů

V tomto dodatku uvádíme přísná matematická odvození tří fundamentálních vztahů používaných napříč pokročilou astroinformatikou.

A.1 Odvození Fisherovy informační matice pro gaussovská data

Uvažujme vektor měření $\mathbf{y} \sim \mathcal{N}(\mu(\theta), \Sigma(\theta))$. Logaritmická věrohodnost má tvar:

$$\ell(\theta) = -\frac{1}{2} \ln \det \Sigma - \frac{1}{2} (\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu) - \frac{N}{2} \ln(2\pi)$$

Aplikací operátoru střední hodnoty na druhou derivaci $\mathbb{E}[-\partial^2 \ell / \partial \theta_i \partial \theta_j]$ odvodil S. S. Sitter slavnou **Whittleovu formuli**:

$$I_{ij}(\theta) = \frac{\partial \mu^T}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right)$$

První člen odpovídá informaci obsažené ve změně střední hodnoty (např. tvaru spektrální čáry), druhý člen odpovídá informaci obsažené ve struktuře kovariancí šumu.

A.2 Odvození akceptačního poměru v Metropolisově–Hastingsově algoritmu

Nechť $\pi(\theta)$ je cílová pravděpodobnost a $q(\theta'|\theta)$ návrhová hustota. Přejížděcí pravděpodobnost je:

$$P(\theta \rightarrow \theta') = q(\theta'|\theta) \alpha(\theta, \theta')$$

Podmínka detailní rovnováhy vyžaduje:

$$\pi(\theta) q(\theta'|\theta) \alpha(\theta, \theta') = \pi(\theta') q(\theta|\theta') \alpha(\theta', \theta)$$

Úpravou poměru získáme:

$$\frac{\alpha(\theta, \theta')}{\alpha(\theta', \theta)} = \frac{\pi(\theta') q(\theta|\theta')}{\pi(\theta) q(\theta'|\theta)}$$

Aby byla maximalizována pravděpodobnost přechodu, volíme $\alpha \leq 1$:

$$\alpha(\theta, \theta') = \min \left(1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)} \right)$$

Což dokazuje platnost Hastingsa vztahu.

Dodatek B: Programátorský referenční manuál moderních knihoven

Stručný přehled syntaxe klíčových Python knihoven pokročilé astroinformatiky.

1. `emcee` – Affine-Invariant MCMC

```
import emcee
import numpy as np

def log_prob(theta, x, y, yerr):
    lp = log_prior(theta)
    if not np.isfinite(lp):
        return -np.inf
    return lp + log_likelihood(theta, x, y, yerr)

nwalkers, ndim = 32, 5
sampler = emcee.EnsembleSampler(nwalkers, ndim, log_prob, args=(x, y, yerr))
state = sampler.run_mcmc(initial_pos, 5000, progress=True)
samples = sampler.get_chain(discard=1000, thin=15, flat=True)
```

2. `celerite2` – Škálovatelné Gaussovské procesy

```
import celerite2
from celerite2 import terms

<h1>Kvazi-periodické SHOTerm jádro</h1>
kernel = terms.SHOTerm(S0=1.0, w0=2.0, Q=10.0) + terms.RealTerm(a=0.5, c=0.1)
gp = celerite2.GaussianProcess(kernel, mean=0.0)
gp.compute(t, yerr=sigma_y)
log_lik = gp.log_likelihood(y)
mu, var = gp.predict(y, t=t_pred, return_var=True)
```

3. `dynesty` – Dynamický Nested Sampling

```
import dynesty

<h1>ptform mapuje jednotkovou hyperkrychli  $[0, 1]^D$  na prostor parametrů</h1>
dsampler = dynesty.DynamicNestedSampler(log_likelihood, prior_transform,
ndim=4)
dsampler.run_nested()
results = dsampler.results
log_evidence = results.logz[-1]
```

Dodatek C: Sbírka 15 řešených výpočetních úloh v Pythonu

V tomto dodatku uvádíme přehled 15 komplexních řešených astrofyzikálních úloh pokrývajících celou šíři pokročilé astroinformatiky:

1. **Úloha 1:** Výpočet SVD komprese a spektrální dekompozice 10 000 galaktických spekter z přehlídky SDSS.
2. **Úloha 2:** Konstrukce Fisherovy informační matice pro odhad kosmologických parametrů (Ω_m, w_0, w_a) z přehlídky supernov SNLS.
3. **Úloha 3:** Implementace Yorkova algoritmu oboustranné lineární regrese na kalibraci relace Tully-Fisher.
4. **Úloha 4:** Analytická konstrukce Jeffreypsova apriorního rozdělení pro parametry exoplanetárního tranzitu.
5. **Úloha 5:** Vlastní implementace Metropolis-Hastings vzorkovače s diagnostikou konvergence Gelman-Rubin $\hat{R}$.
6. **Úloha 6:** Paralelní MCMC analýza světelné křivky exoplanety Kepler-1654b pomocí knihovny ``emcee``.
7. **Úloha 7:** Výpočet marginální evidence $\mathcal{Z}$ a Bayesova faktoru mezi modely jedné a dvou planet pomocí knihovny ``dynesty``.
8. **Úloha 8:** Detekce rotace hvězdy pomocí zobecněného Lomb-Scargleova periodogramu a Baluevovy FAP z dat TESS.
9. **Úloha 9:** Filtrace stelární granulace a hvězdných skvrn pomocí gaussovských procesů s balíčkem ``celerite2``.
10. **Úloha 10:** Modelování stochastického červeného šumu v kvasaru 3C 273 pomocí spojitého procesu CARMA(1, 0).
11. **Úloha 11:** Fotodynamické fitování TTV poruch a odhad hmotnosti dvou planet v rezonanci 2:1.
12. **Úloha 12:** Fotometrický odhad rudého posuvu (Photo-z) galaxií pomocí Gradient Boosted Trees (``LightGBM``).
13. **Úloha 13:** Vizualizace 6D fázového prostoru hvězdného okolí Slunce z dat Gaia DR3 pomocí algoritmu UMAP.
14. **Úloha 14:** Segmentace slabých gravitačních oblouků v galaktických kupách pomocí neuronové sítě U-Net v PyTorch.
15. **Úloha 15:** Masivně paralelní prostorový cross-matching miliardového katalogu Gaia DR3 s katalogem AllWISE pomocí HEALPix a DuckDB.

Dodatek D: Encyklopedický glosář pokročilé astroinformatiky

- **AdamW**: Optimalizační algoritmus pro hluboké neuronové sítě, který správně odděluje pokles vah (Weight Decay) od adaptivního odhadu momentů gradientu.
- **Affine Invariance (Afinní invariance)**: Vlastnost statistického algoritmu (např. vzorkovače `emcee`), díky níž je chování a efektivita metody nezávislá na lineárních transformacích a měřících prostoru parametrů.
- **Autokorelační čas (τ)**: Počet kroků v Markovově řetězci nutný k tomu, aby se vygeneroval jeden statisticky nezávislý vzorek z cílového posterioru.
- **Bayesův faktor (B_{12})**: Poměr marginálních věrohodností (evidencí) dvou konkurenčních modelů $\mathcal{Z}_1/\mathcal{Z}_2$, vyjadřující míru preference modelu $\mathcal{M}_1$ oproti $\mathcal{M}_2$ na základě naměřených dat.
- **CARMA**: Continuous-time AutoRegressive Moving Average; spojité stochastické diferenciální modely používané k analýze nepravidelně vzorkovaného červeného šumu.
- **Choleského faktorizace**: Dekompozice pozitivně definitní kovarianční matice $\Sigma = \mathbf{L}\mathbf{L}^T$, nezbytná pro efektivní bělení šumu a vyhodnocování gaussovských procesů.
- **Cramérova–Raoova dolní mez (CRLB)**: Teoretická mez, která stanovuje minimální možný rozptyl jakéhokoliv nestranného odhadu parametrů jako inverzi Fisherovy informační matice.
- **Dynesty**: Specializovaná Python knihovna pro dynamický Nested Sampling umožňující simultánní odhad parametrů a výpočet marginální věrohodnosti.
- **Evidence ($\mathcal{Z}$)**: Marginální věrohodnost modelu získaná integrací věrohodnosti přes celý apriorní prostor parametrů.
- **Fisherova informační matice (FIM)**: Míra množství informace, kterou pozorovaná data nesou o neznámých parametrech modelu; je střední hodnotou druhé derivace logaritmické věrohodnosti.
- **Gaussovský proces (GP)**: Nekonečněrozměrný stochastický proces definovaný střední funkcí a kovariančním jádrem, sloužící k neparametrickému modelování korelovaného šumu.
- **GELU**: Gaussian Error Linear Unit; hladká nelineární aktivační funkce využívaná v transformerových architekturách.
- **Gelman–Rubinův diagnostický test ($\hat{R}$)**: Míra konvergence paralelních MCMC řetězců porovnávající rozptyl mezi řetězci s průměrným vnitřním rozptylem řetězců.
- **Hamiltonovské Monte Carlo (HMC)**: Fyzikálně motivovaný vzorkovací algoritmus propojující parametry s fiktivní hybností a integrující pohybové rovnice po trajektoriích zachovávajících Hamiltonián.

- **HDBSCAN:** Hierarchický shlukový algoritmus založený na hustotě, schopný identifikovat shluky proměnlivé hustoty a izolovat šumové pozadí.
- **HEALPix:** Hierarchical Equal Area isoLatitude Pixelization; matematický standard pixelizace kulové plochy se stejnou plochou všech pixelů.
- **Lombův–Scargleův periodogram:** Metoda odhadu výkonového spektra pro nepravidelně vzorkované časové řady.
- **Nested Sampling:** Algoritmus Johna Skillinga transformující mnohorozměrný integrál evidence na jednorozměrný integrál přes apriorní kumulativní objem.
- **No-U-Turn Sampler (NUTS):** Pokročilé rozšíření HMC, které automaticky určuje optimální délku integračního kroku rekursivním stavěním binárních stromů.
- **PINN:** Physics-Informed Neural Networks; neuronové sítě integrující parciální diferenciální rovnice přímo do své optimalizační ztrátové funkce.
- **UMAP:** Uniform Manifold Approximation and Projection; nelineární technika redukce dimenzionality zachovávající lokální i globální topologické struktury.
- **Vision Transformer (ViT):** Architektura hlubokého učení aplikující mechanismus samopozornosti na obrazová data rozdělená do dlaždic.
- **Yorkův algoritmus:** Přísné řešení lineární regrese minimalizující odchylky se zohledněním nezávislých i korelovaných chyb v obou osách.