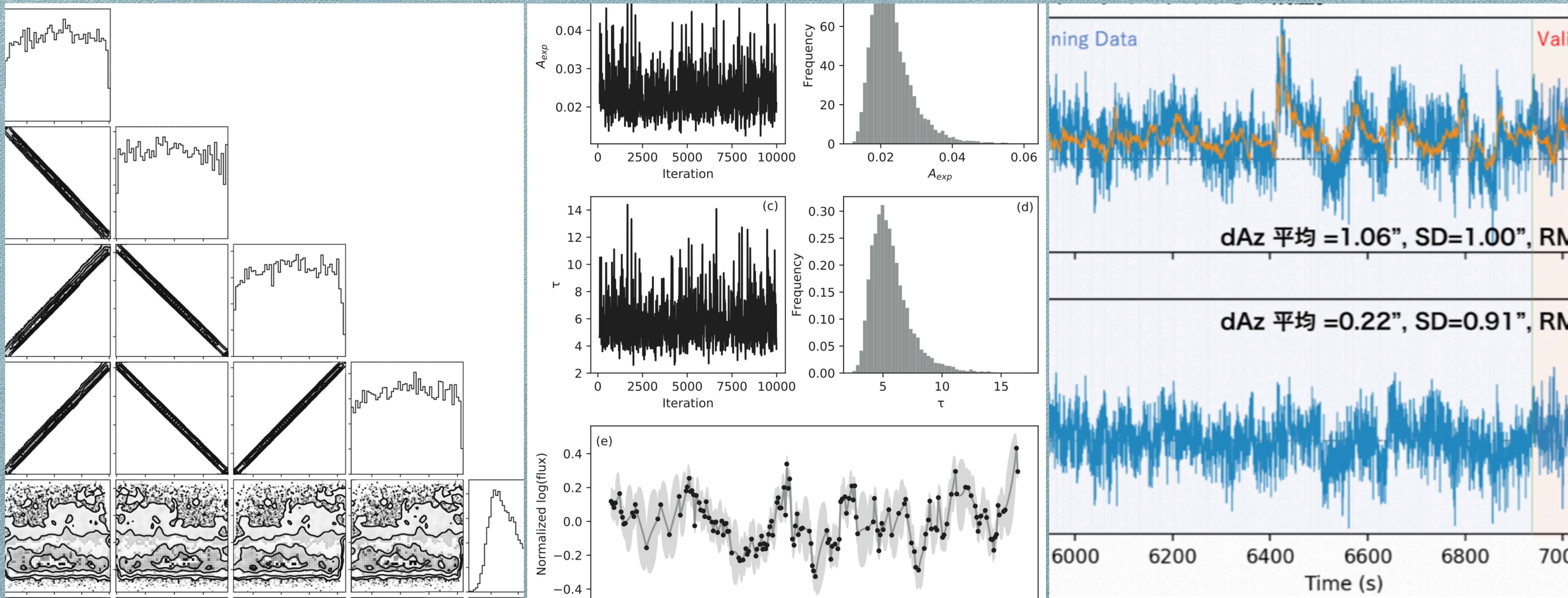




データ科学・機械学習的な手法の応用例

植村誠 (広島大学) @光赤天連シンポ2020

- ◆ MCMC

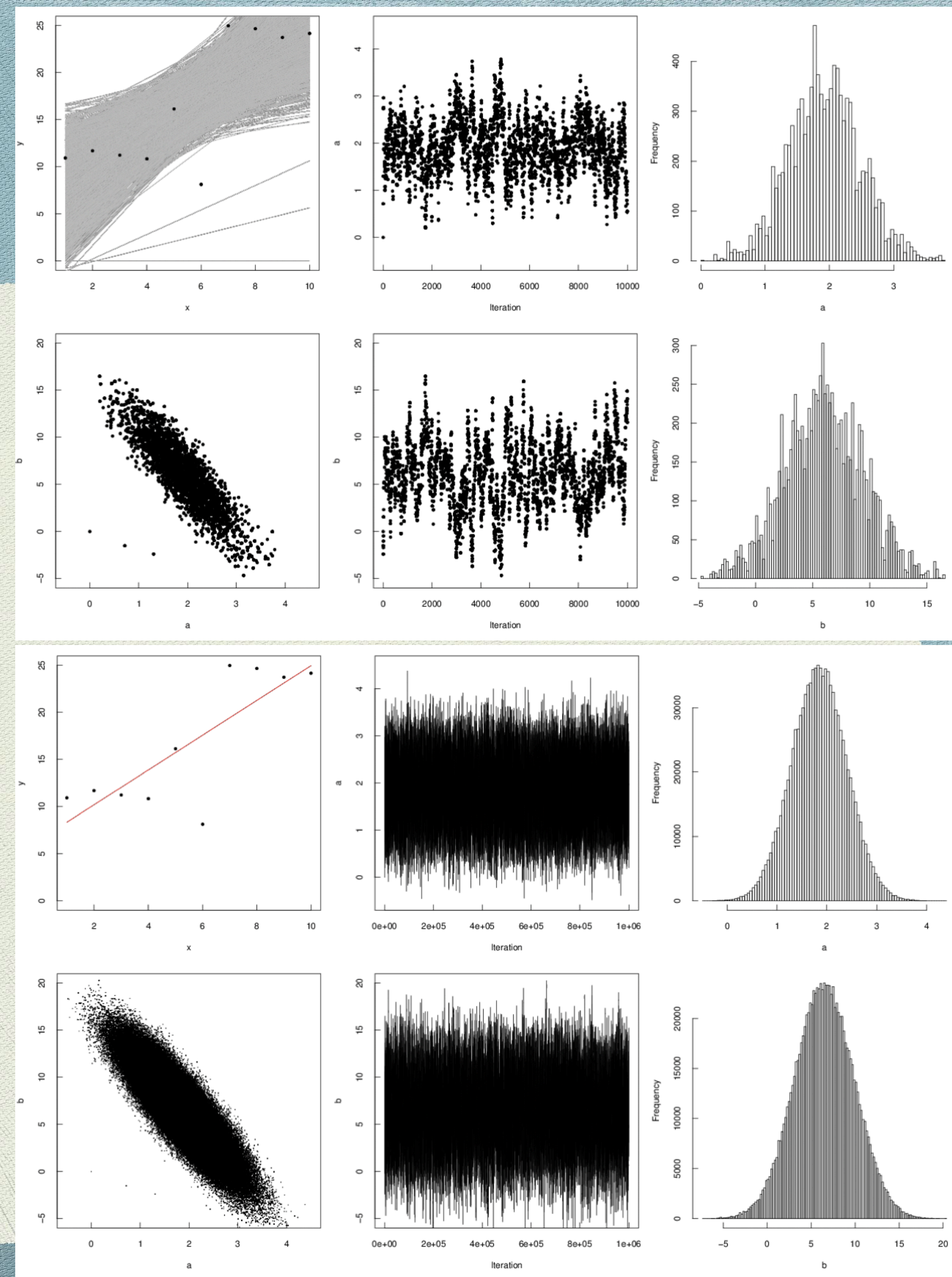
- ◆ ガウス過程

- ◆ 機械判別（スパースロジスティック回帰）

- ◆ ニューラルネットワーク

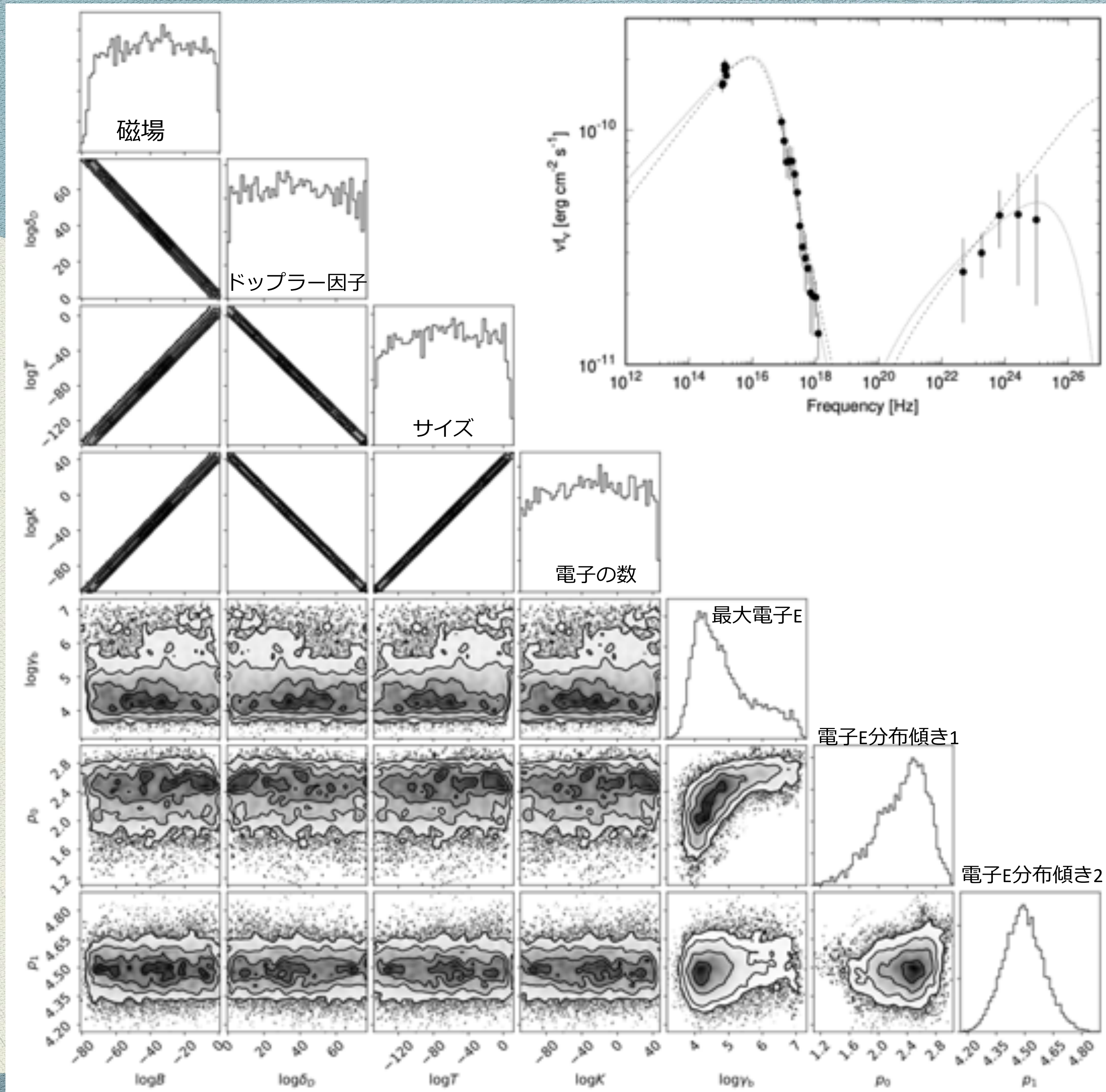
マルコフ連鎖モンテカルロ法 (MCMC)

- ◆ モデルパラメータの確率分布を推定する
- ◆ パラメータの数が3つくらいまでなら「カイ二乗マップ」を描く。それ以上の次元なら、効率の良いサンプリング手法が必要
- ◆ 実際には高次元パラメータ空間を適切に探索するのは難しい
- ◆ 様々なアルゴリズム：メトロポリス法、ハミルトニアンモンテカルロ、レプリカ交換法



ブレーザーのSEDから SSCのパラメータ推定

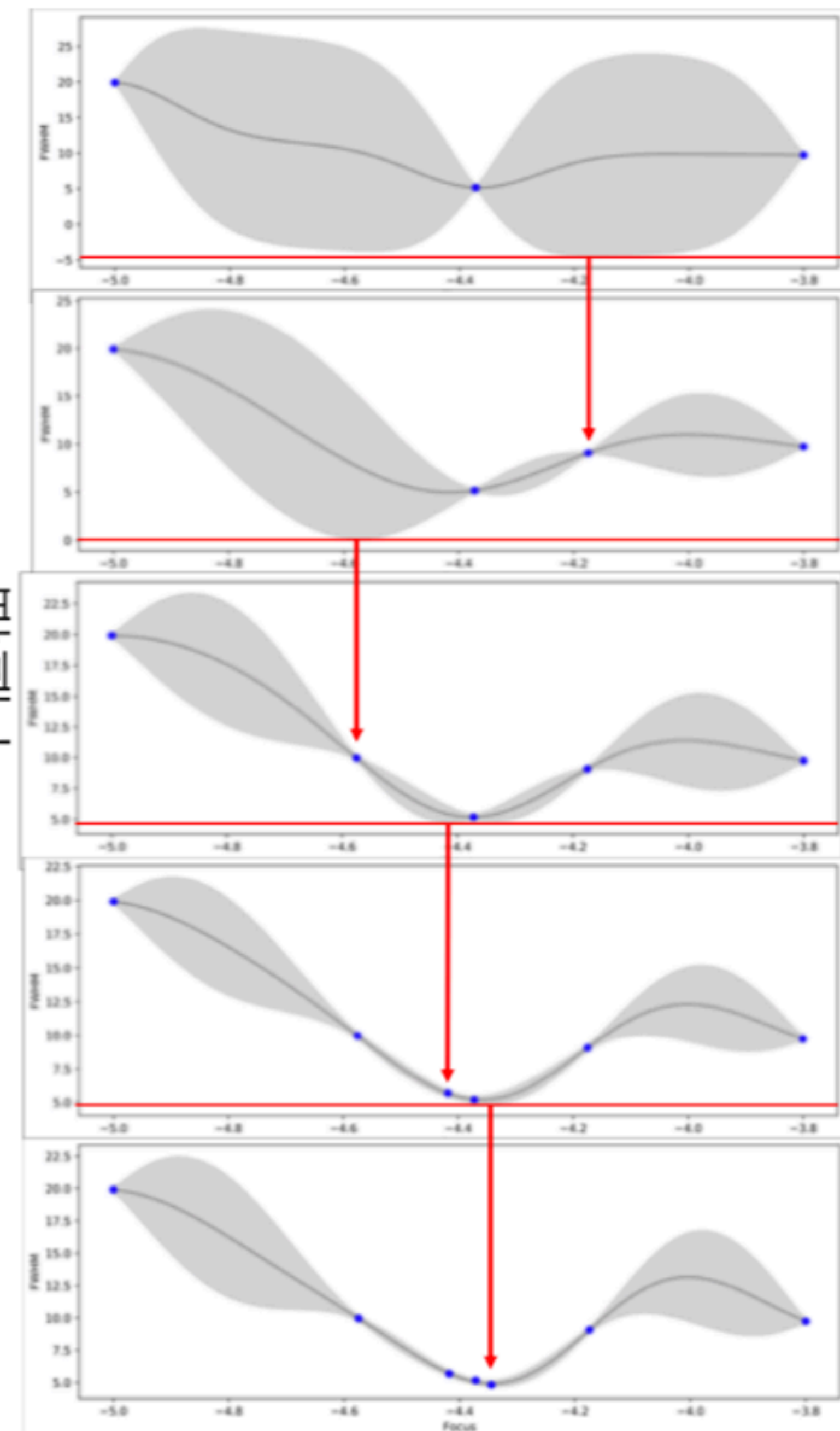
- ◆ Yamada, et al. (2020)
- ◆ ブレーザー：AGNジェットがこっちに向いてる天体
- ◆ SSC：シンクロトロン自己コンプトンモデル
- ◆ モデルが「縮退」して解が一意に決まらない
- ◆ 解の分布をMCMCでサンプリング
- ◆ 4つのパラメータが強く相関。1つだけは決めないといけない。電子のエネルギー分布に関する3つのパラメータは決まる。



ガウス過程 (Gaussian Process: GP)

- ◆ いくつかデータがあって内挿して欠損部分を推定したいけど、関数の形はわからない
- ◆ N 個のデータを N 次元正規分布からのサンプル、と考える
- ◆ $N \times N$ 共分散行列の要素にカーネル関数を使うことで、データからカーネルのパラメータを推定
- ◆ 信頼区間付きで推定されるので、実験計画に使える
- ◆ $N \times N$ 行列を解くので $N=10^4$ を越すとしんどくなってくる

半値幅

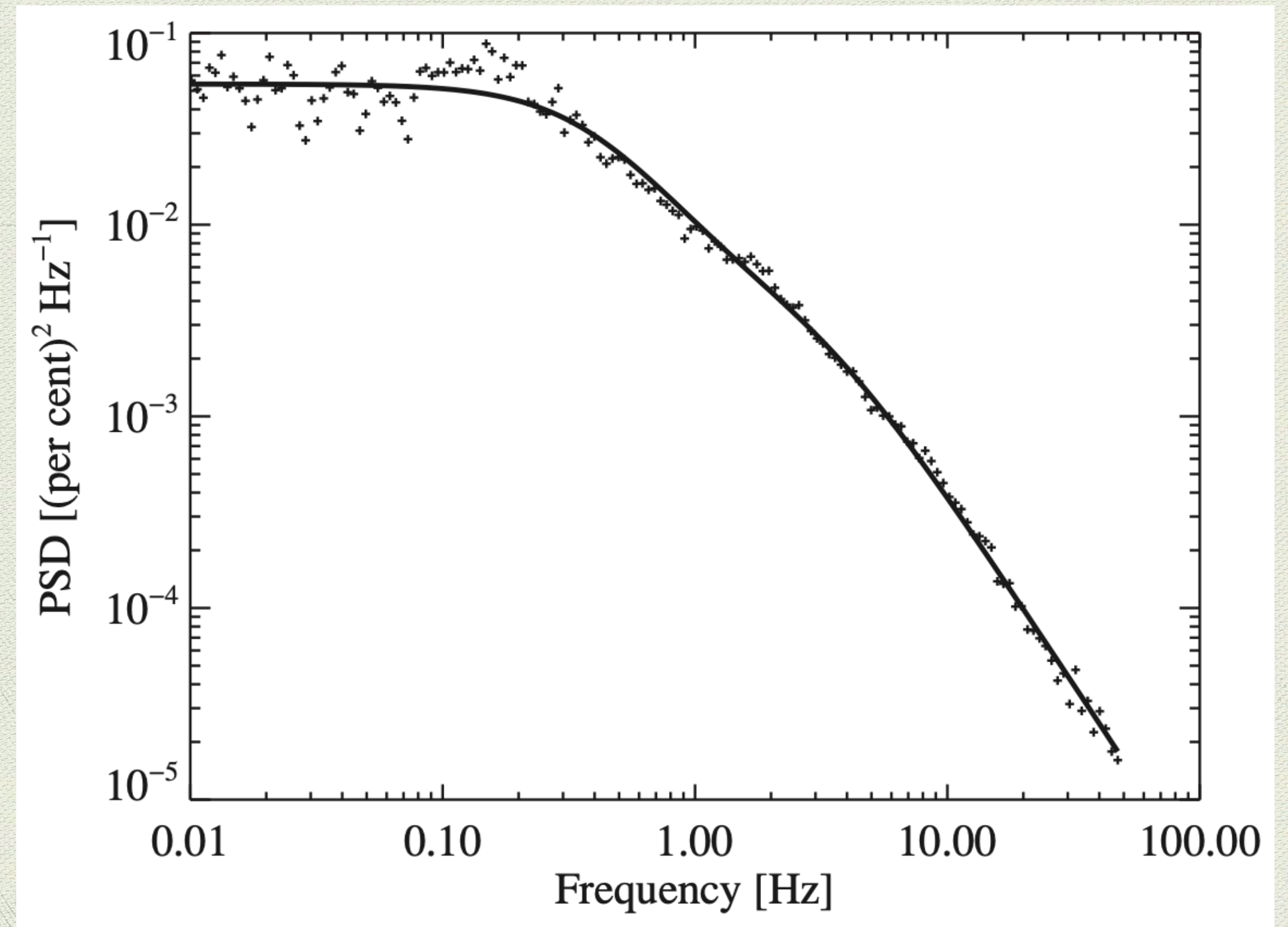


副鏡位置

Ornstein-Uhlenbeck過程 (OU process)

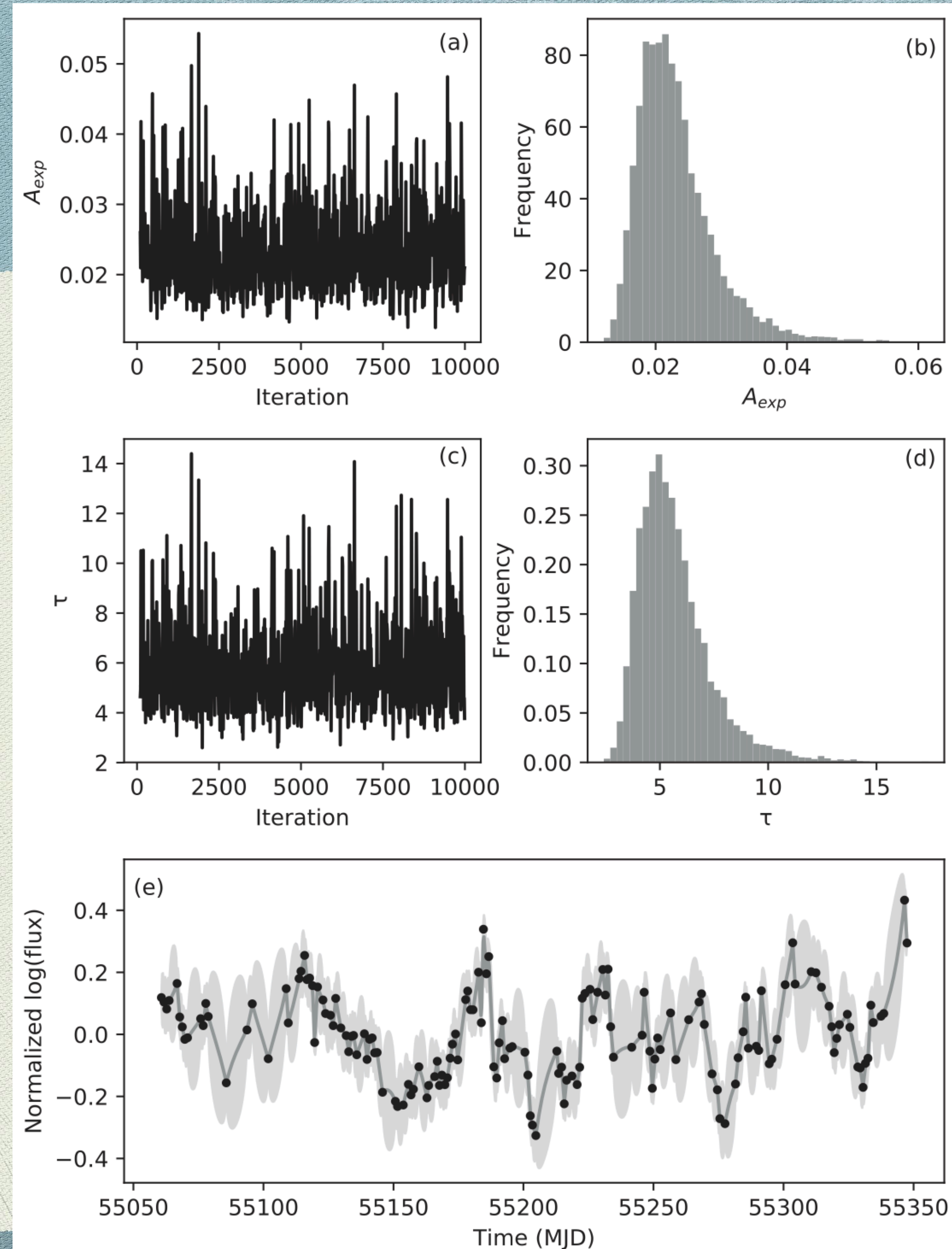
Kelly+11

- ◆ ガウス過程の共分散行列に指数カーネルを使ったものと等価
$$S_{ij} = A_{\text{exp}} \exp\left(-\frac{|t_i - t_j|}{\tau}\right)$$
- ◆ Damped Random Walk モデル、1次自己回帰モデルとも同じ
- ◆ パワースペクトルが低周波側で一定、高周波側で-2乗で落ちる
- ◆ AGNの時間変動で昔からよく使われてきたモデル



ブレーザーの光度曲線の特徴抽出

- ◆ Uemura, et al. (2020)
- ◆ ブレーザーの不規則な時間変動：どう特徴付ける？
- ◆ OU過程回帰をして、カーネルパラメータである「変動の大きさ（分散）」「タイムスケール」をMCMCで推定
- ◆ 一部を検証データとして隠してOU過程回帰、交差検証誤差を「非定常性」の指標とする。



機械判別

(スパースロジスティック回帰: SMLR)

- ◆ 多クラスのロジスティック回帰：基本的な機械判別手法

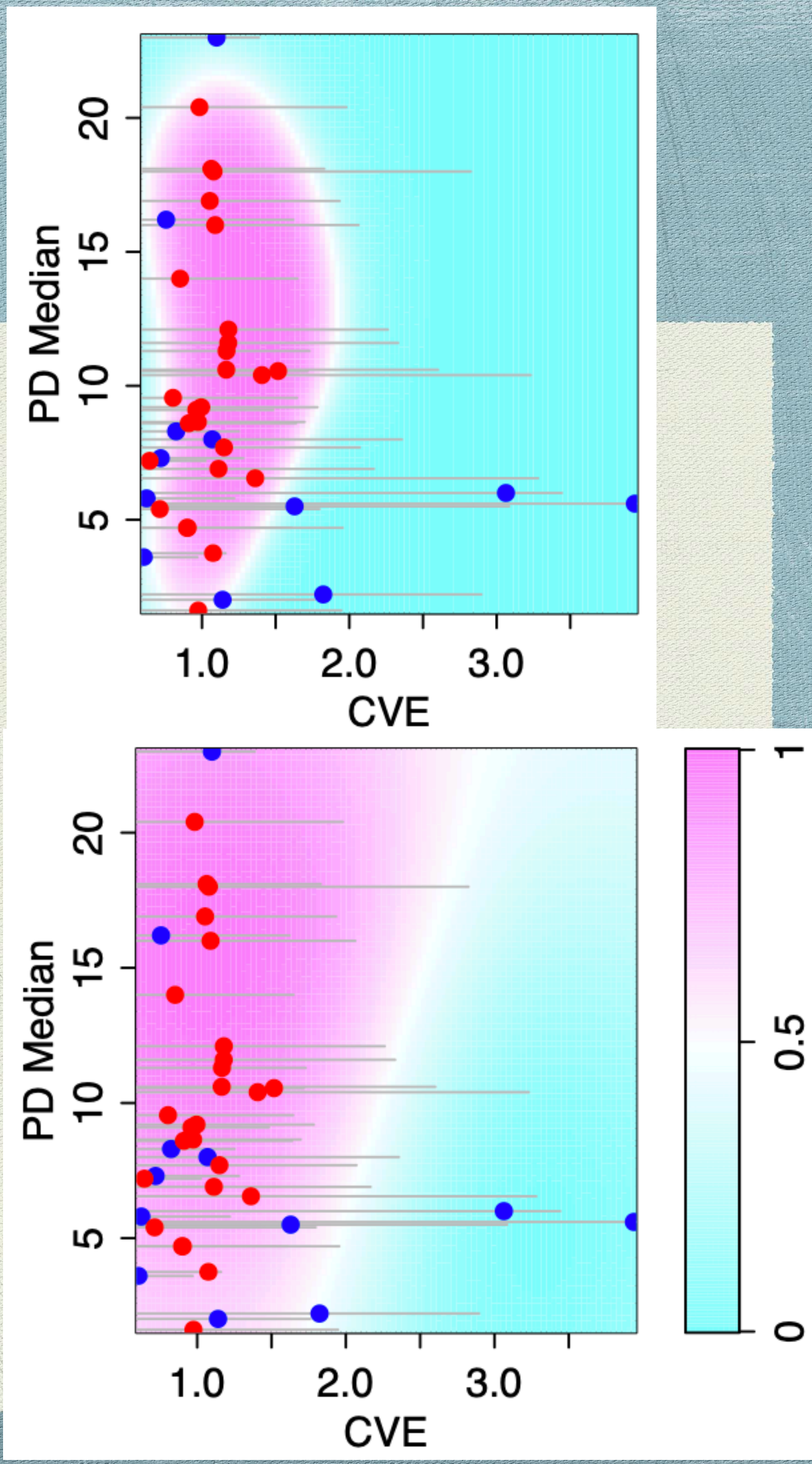
$$P(y^{(j)} = 1 \mid \boldsymbol{\theta}, \boldsymbol{w}) = \frac{\exp(\boldsymbol{w}^{(j)T} \boldsymbol{\theta})}{\sum_{j=1}^M \exp(\boldsymbol{w}^{(j)T} \boldsymbol{\theta})}$$

$$\ell(\boldsymbol{w}) = \sum_{i=1}^N \log P(y_i \mid \boldsymbol{\theta}_i, \boldsymbol{w})$$

- ◆ 特徴量をカーネル関数で非線形空間へ＝過適合の恐れ＝スパース正則化

$$\hat{\boldsymbol{w}} = \arg \max_{\boldsymbol{w}} \{\ell(\boldsymbol{w}) - \lambda \|\boldsymbol{w}\|_1\}.$$

- ◆ カーネル関数にガウス分布（RBFカーネル）を使うと、サポートベクターマシンのようなモデルに。



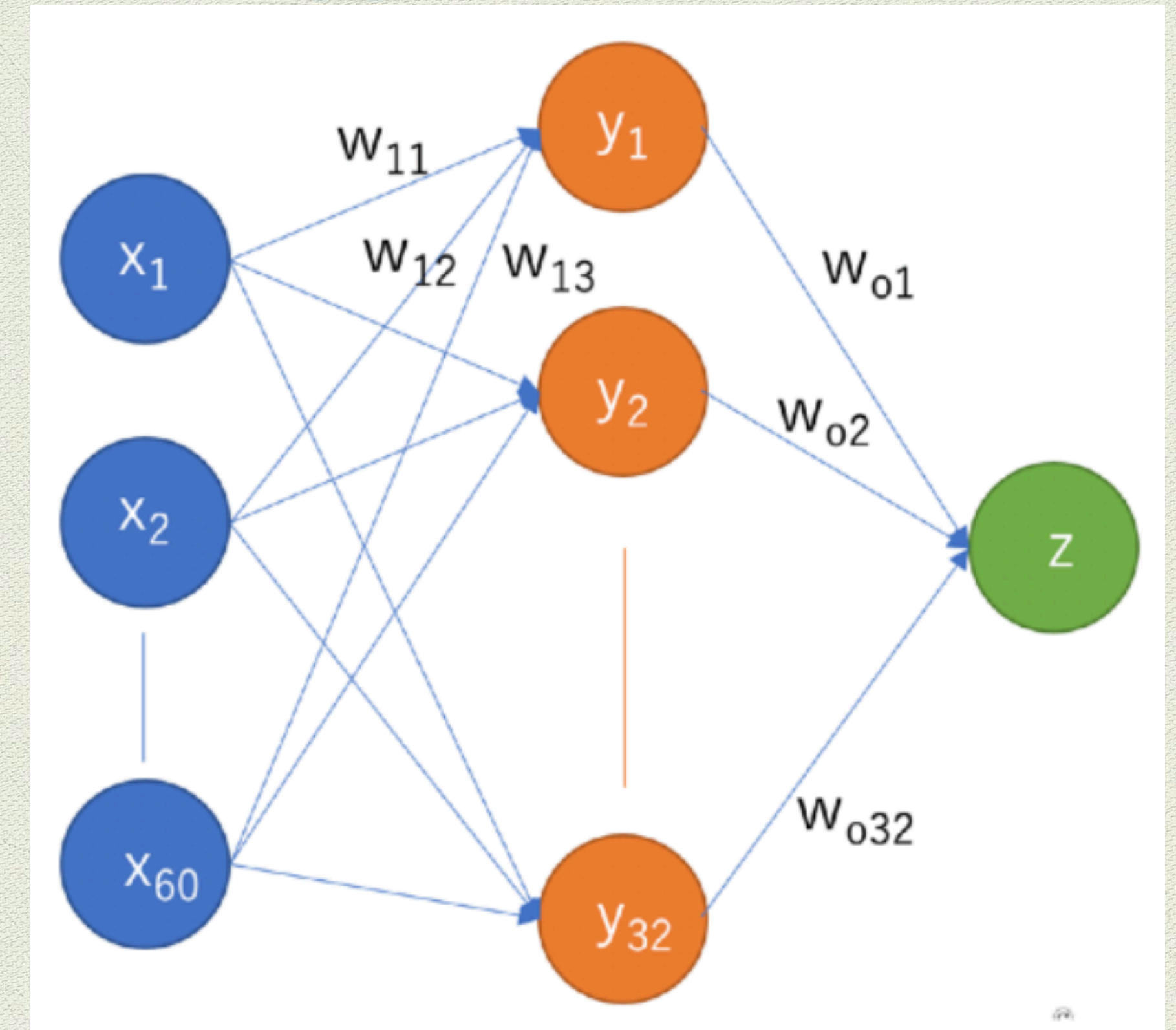
BL LacとFSRQを分ける特徴量

- ◆ ブレーザーの2つのサブタイプ：BL Lac型とFlat Spectrum Radio Quasar (FSRQ)型 → 何が違う？
- ◆ OU過程回帰でブレーザーの時系列データから特徴抽出（光度曲線＋偏光の変化）
- ◆ SMLRでBL Lac と FSRQの判別モデルを構築し、モデルの性能を評価。
- ◆ どの特徴量を使うと良い判別モデルになる？ → 非定常性、変動幅、タイムスケール、平均偏光度 が2つのサブタイプを分ける重要な特徴。

Light curve				Polarization degree					
CVE	Slope	A_{exp}	τ	Median	CVE	Slope	A_{exp}	τ	AUC
●	—	●	●	●	—	●	—	—	0.907
●	●	●	●	●	—	—	—	—	0.904
●	●	●	●	●	—	●	—	—	0.888
●	●	●	●	●	—	—	—	●	0.885
●	●	●	●	●	●	—	—	—	0.881
●	—	●	●	●	—	●	—	●	0.881
●	●	●	●	●	●	●	●	●	0.878
●	—	●	●	●	●	●	—	—	0.878
●	—	●	●	●	●	—	—	●	0.875

ニューラルネットワーク(NN)

- ◆ 入力データを「活性化関数」で変換。変換したものをさらに次の「層」で変換して・・・を何度かして、出力。出力結果とデータを比較して（最小二乗法など）最適な活性化関数のパラメータを決める。
- ◆ 時系列の予測に特化した Recurrent NN、画像判別に有効な Convolutional NN など、様々な「ネットワーク」の発展。
- ◆ 最新の深層学習(Deep NN)ではパラメータ数=1000億個。データはそれ以上が必要。



風データからASTEの指向方向を推定

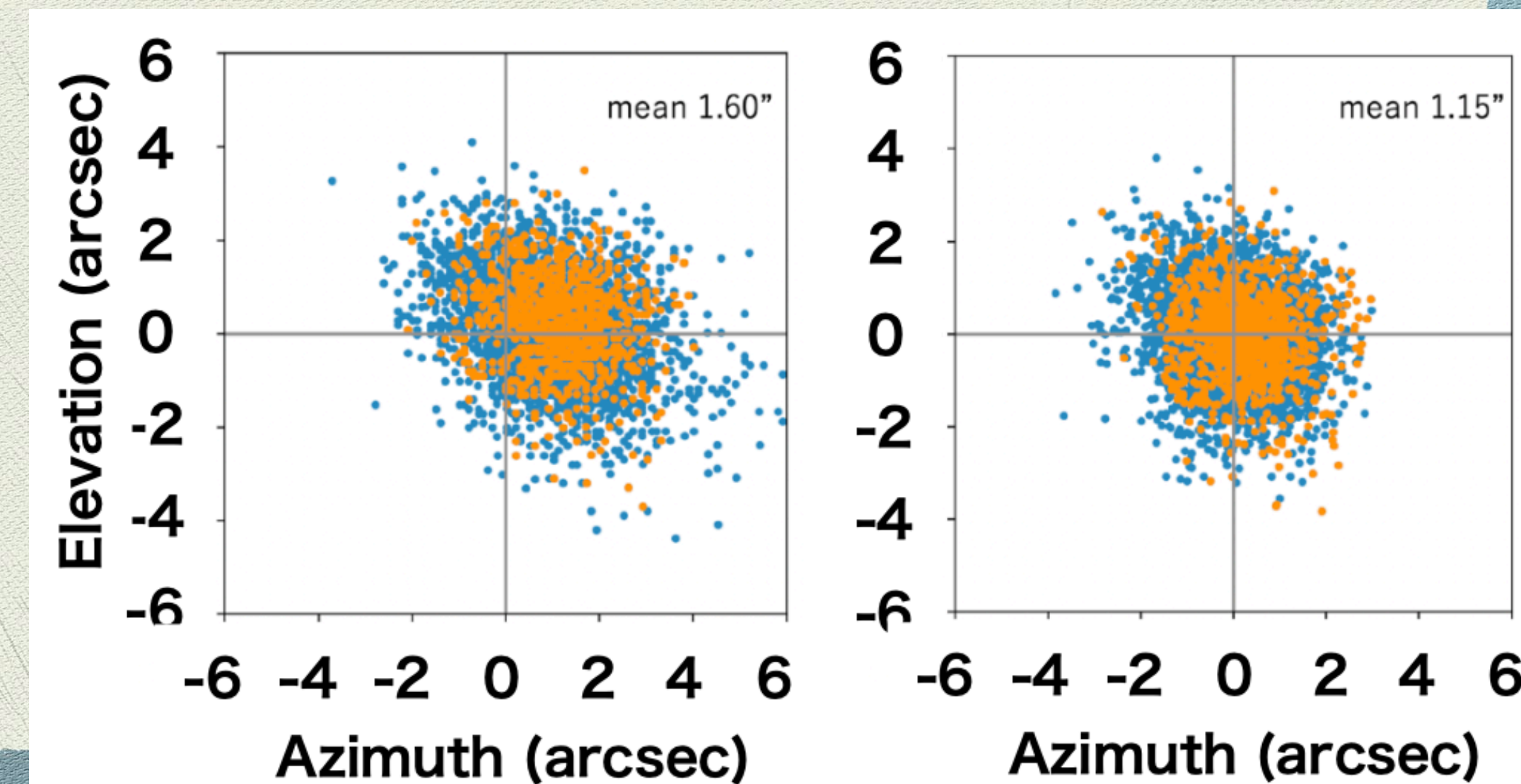
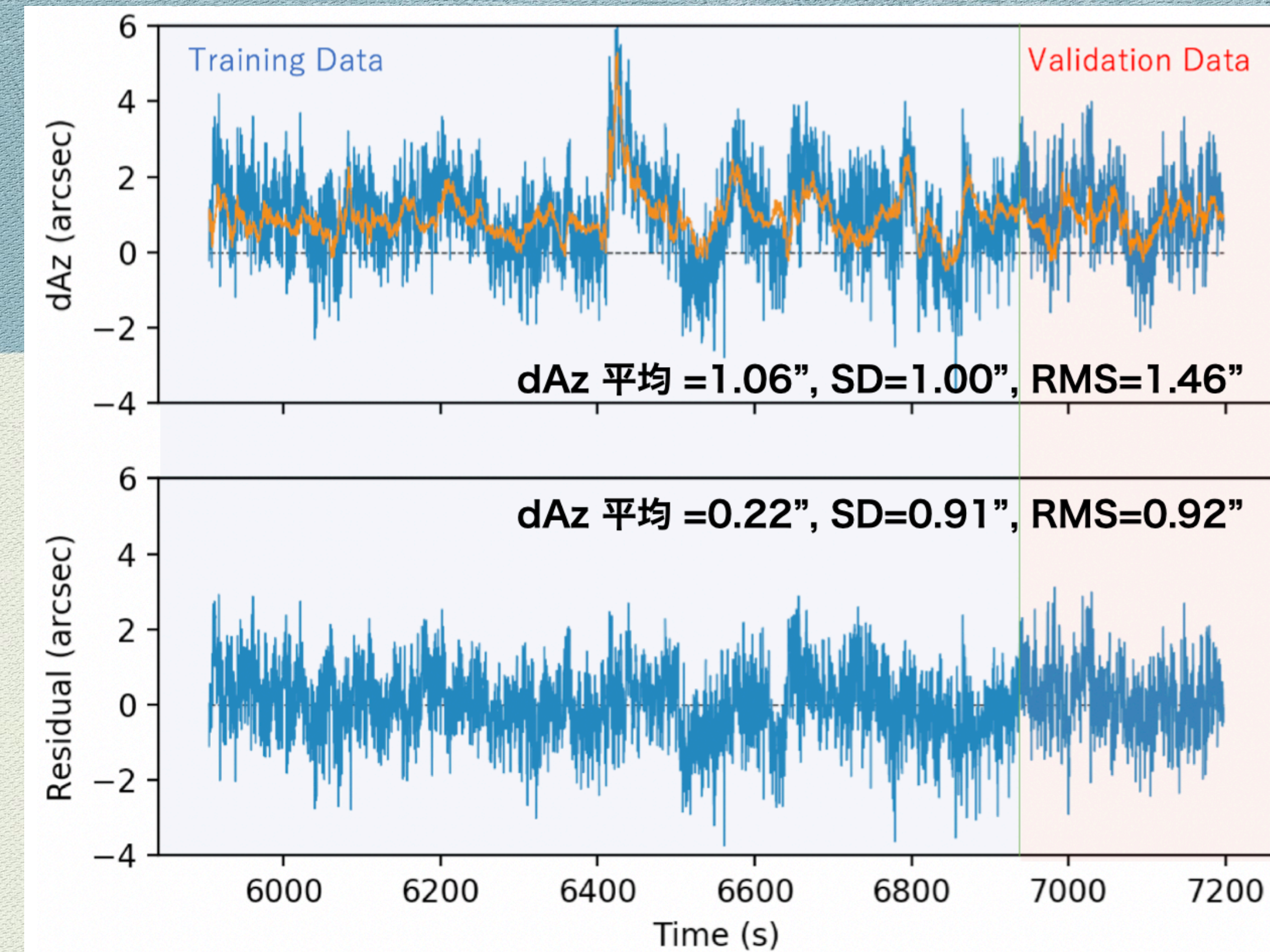
- ◆ 風であおられてASTEの指向方向がズレる
- ◆ 風向・風速のデータからアンテナがズレる量を推定したい
- ◆ 本来は風の流れやアンテナがうける力を流体シミュレーションで予測？ やってられない。

- ◆ 全結合32ユニット (Relu関数) 1 層の単純なNN。パラメー

タ数=2049

$$y_j = f \left(\sum_{i=1}^{62} w_{ij} x_i + b_j \right)$$
$$z = \sum_{j=1}^{32} w_{oj} y_j + b_o$$

- ◆ 入力データ：62次元（風速31+風向31）、4135セット



まとめ

- ◆ MCMC
 - ◆ SEDフィッティング：解の分布を調査
- ◆ ガウス過程
 - ◆ OU過程回帰：不規則な光度曲線の特徴付け
- ◆ 機械判別（スパースロジスティック回帰）
 - ◆ 判別に重要な変数の選択
- ◆ ニューラルネットワーク
 - ◆ 風データから望遠鏡の指向方向を推定