近似確率伝搬法による一般化線形モデルの予測誤差表現について

統計数理研究所　数理・推論研究系

坂田　綾香

Special thanks to Yukito Iba

今日の内容

● 予測誤差(の推定量)を評価する

動機
- 予測誤差を何のために評価するのか

目標
- 予測誤差の推定量の解釈と比較
- 効率的な計算法の開発

手法
- ファクターグラフ表現を用いた近似推論法
  - ファクターグラフ表現…確率分布をグラフで表す
  - 近似推論法…確率伝搬法

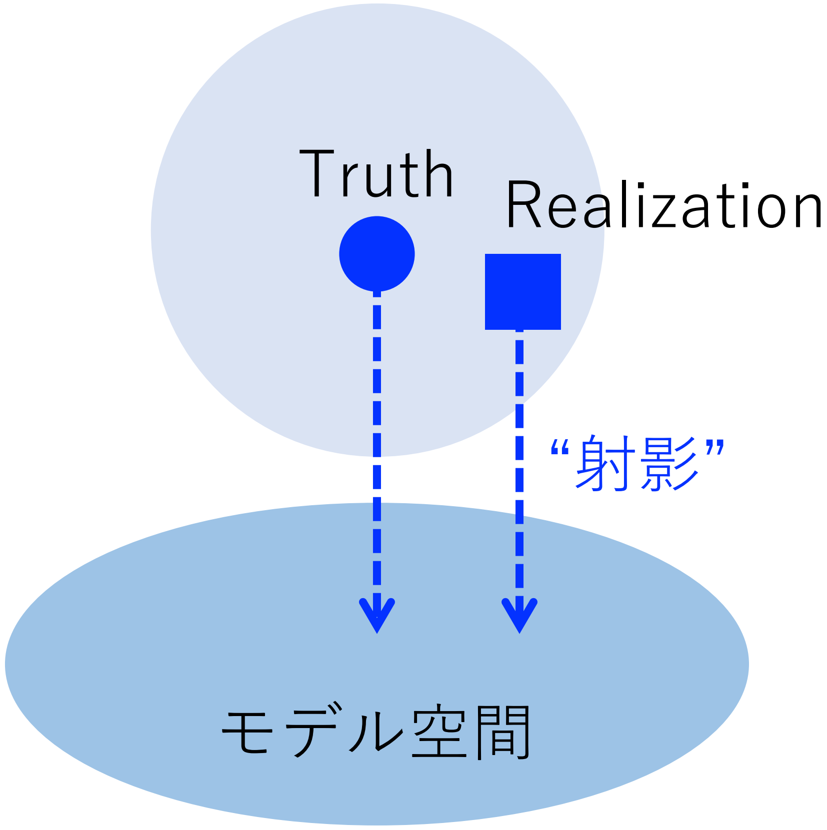
結果
- 確率伝搬法を使うことで出来ること，
  もたらされる解釈について説明

予測誤差とは

予測に基づく統計的モデリング

統計学におけるモデリング

- 与えられたデータを説明するモデルを作る
  - モデル：確率分布
  - 一般に，データが生成される過程は未知
- 可能な確率分布のうち，どれが適切なのか？
  - 用意したモデル候補から
    適切なものを選ぶ
  - これを**モデル選択**という
- データは一つの実現値である
  - 実現値をもとに，確率分布の性質を知ることが目標



”The Elements of Statistical Learning”
より

モデル選択

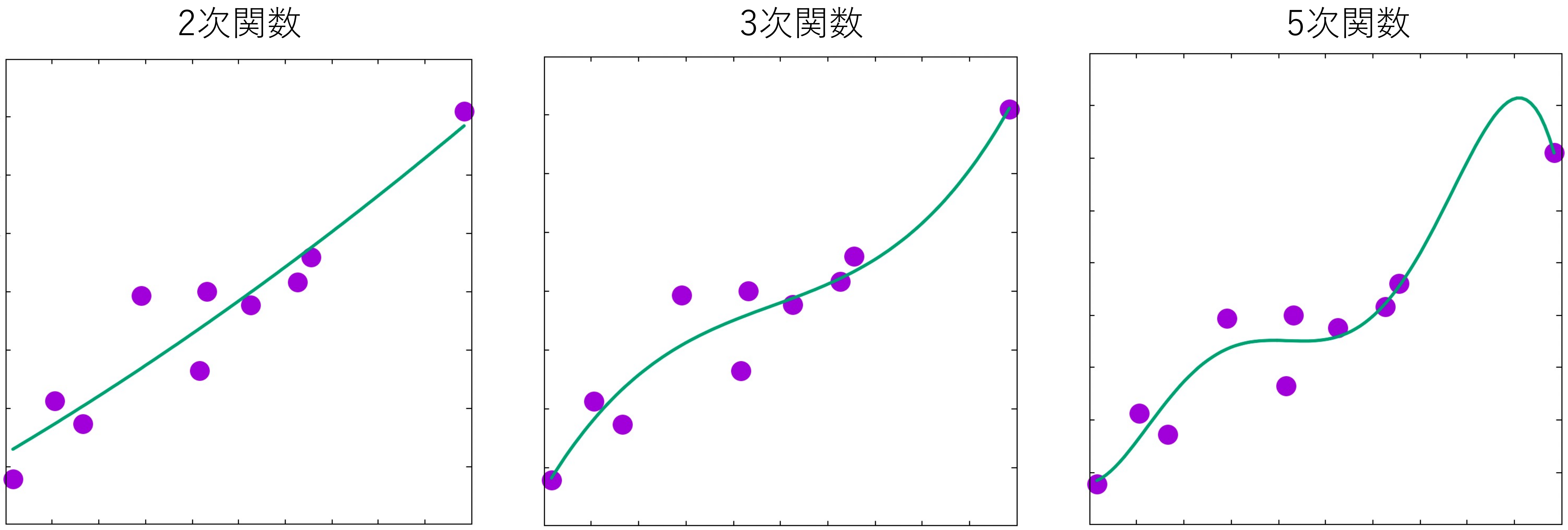
(例) データ $\mathcal{D}=\{z_1,\dots,z_M\}$を得たとする

- パラメータθをもつ確率分布$p(Z|\theta)$を導入する
- データ$\mathcal{D}$のもとで，パラメータθを推定する → $\hat{\theta}(\mathcal{D})$とする
  - (例) 最尤推定

$$\hat{\theta}_{\mathrm{ML}}(\mathcal{D})=\max_{\theta} L(\theta;\mathcal{D}), \qquad L(\theta;\mathcal{D})=\sum_{\mu=1}^{M}\ln p\left(Z=z_\mu|\boldsymbol{\theta}\right)$$

- モデル候補$\mathcal{M}=\left\{p_1\left(Z|\hat{\boldsymbol{\theta}}_1(\mathcal{D})\right),p_2\left(Z|\hat{\boldsymbol{\theta}}_2(\mathcal{D})\right),\dots,p_K\left(Z|\hat{\boldsymbol{\theta}}_K(\mathcal{D})\right)\right\}$の中から<u>適切な</u>モデルを選ぶ

モデル選択の例



どのモデルでデータを記述することが適切か？

客観的な指標が必要：**予測性能**

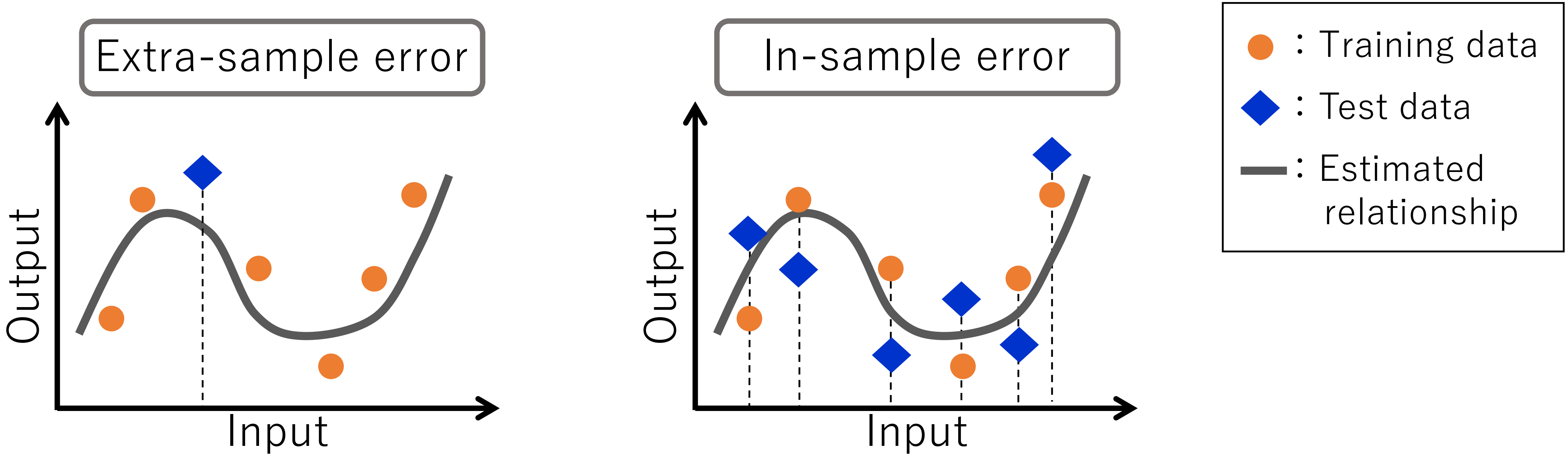
「予測」に基づくモデリング

- 与えられたデータに対するあてはまりの良さは必要
- 一方で，与えられたデータに過適合してしまうことも問題
  - 将来の予測ができなくなる可能性がある

「予測」に基づくモデリングの立場

- "真のモデル"から生成される 新しいデータについて高い**予測能力**を持つモデルを良いモデルとする
  - "真のモデル"を見つけることが目的ではない．
    - 結果的に"真のモデル"が得られるのであれば，それはそれで良いこと．
  - 予測能力を**予測誤差**で定量化

予測誤差とは



訓練データと
異なる入出力関係を
テストデータとする

訓練データと
同じ入力に対する異なる出力を
テストデータとする

推定量：交差検証誤差 ⟷ 推定量：情報量規準, Cp規準

漸近等価 [Stone (1977)]

予測誤差の定義と汎化ギャップ

以下では回帰問題を考える．

- 仮定する確率分布を$f(\cdot\,|\boldsymbol{\theta})$とする．$\boldsymbol{\theta}(\boldsymbol{F},\boldsymbol{x})=\frac{1}{\sqrt{N}}\boldsymbol{F}\boldsymbol{x}$はパラメータ．
- 回帰係数：$\boldsymbol{x}\in\mathbb{R}^N$
- 説明変数：$\boldsymbol{F}\in\mathbb{R}^{M\times N}$
- 出力：$\boldsymbol{y}$（M次元ベクトル）

データセット$\mathcal{D}=\{\boldsymbol{y},\boldsymbol{F}\}$

● データ$\mathcal{D}$のもとでの推定値を$\hat{\boldsymbol{x}}(\mathcal{D})$と表現する．

- Training error： $\mathrm{err}_{\mathrm{train}}(\mathcal{D})=-\frac{1}{M}\ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\left(\boldsymbol{F},\hat{\boldsymbol{x}}(\mathcal{D})\right)\right)$
- Ex-sample prediction error： $\mathrm{err}_{\mathrm{pre}}^{(\mathrm{ex})}(\mathcal{D})=-\mathbb{E}_{Z,\mathbf{f}_{\mathrm{new}}}\left[\ln f\left(Z|\boldsymbol{\theta}\left(\mathbf{f}_{\mathrm{new}},\hat{\boldsymbol{x}}(\mathcal{D})\right)\right)\right]$
- In-sample prediction error： $\mathrm{err}_{\mathrm{pre}}^{(\mathrm{in})}(\mathcal{D})=-\frac{1}{M}\mathbb{E}_{\boldsymbol{Z}}\left[\ln f\left(\boldsymbol{Z}|\boldsymbol{\theta}\left(\boldsymbol{F},\hat{\boldsymbol{x}}(\mathcal{D})\right)\right)\right]$

予測誤差と
訓練誤差の差を
汎化ギャップと呼ぶ

予測誤差の推定量

- 予測誤差を定義通りに評価するには，真の確率分布が必要
- 真の確率分布は未知なので厳密評価は不可能

→ 予測誤差の**推定量**を構成する

  - 確率変数の実現値を入れると，それに応じた値を返す関数

推定量の例

  - 交差検証誤差
  - 情報量規準
  - Cp規準

交差検証誤差(cross validation error)

- ここではleave-one-out CV (LOOCV) errorを考える

- μ番目のサンプルを抜いたデータを$\mathcal{D}_{\backslash\mu}$とする.
- $\mathcal{D}_{\backslash\mu}$のもとでの推定値を$\hat{x}(\mathcal{D}_{\backslash\mu})$とする.

■ LOOCV errorの定義は次の通り

$$\mathrm{err}_{\mathrm{LOOCV}}(\mathcal{D}) = -\frac{1}{M}\sum_{\mu=1}^{M} \ln f\left(y_\mu \middle| \frac{1}{\sqrt{N}}\mathbf{f}_\mu \hat{\boldsymbol{x}}(\mathcal{D}_{\backslash\mu})\right)$$

■ 利点：さまざまなモデルに適用可能

■ 欠点：計算量が多い

- **Linear estimator**の場合は解析的に表現できる

Linear estimator

- **Linear estimator**

  モデルによる出力の表現$\hat{\boldsymbol{y}}$が $\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$として与えられる ($\boldsymbol{A} \in \mathbb{R}^{M \times M}$)

(例) 線形回帰問題における最尤推定 ($M > N$)

$$\hat{\boldsymbol{x}} = \operatorname*{argmin}_{\boldsymbol{x}} \left\| \boldsymbol{y} - \frac{1}{\sqrt{N}} \boldsymbol{F}\boldsymbol{x} \right\|_2^2$$

$$\hat{\boldsymbol{x}} = \left(\frac{1}{N} \boldsymbol{F}^\top \boldsymbol{F} \right)^{-1} \frac{1}{\sqrt{N}} \boldsymbol{F}^\top \boldsymbol{y}, \qquad \hat{\boldsymbol{y}} = \frac{1}{\sqrt{N}} \boldsymbol{F}\hat{\boldsymbol{x}} = \frac{1}{N} \boldsymbol{F} \left(\frac{1}{N} \boldsymbol{F}^\top \boldsymbol{F} \right)^{-1} \boldsymbol{F}^\top \boldsymbol{y}$$

- 特に，行列 $\boldsymbol{H} = \frac{1}{N} \boldsymbol{F} \left(\frac{1}{N} \boldsymbol{F}^\top \boldsymbol{F} \right)^{-1} \boldsymbol{F}^\top$を**Hat matrix**と呼ぶ [Hoaglin & Welsch (1978)].

Linear estimatorの場合のLOOCV error

- Linear estimatorの場合，leave-one-out sample $\mathcal{D}_{\backslash\mu}$のもとでの出力表現は次のように与えられる

$$\hat{y}_\mu\left(\mathcal{D}_{\backslash\mu}\right) = \frac{\hat{y}_\mu(\mathcal{D}) - A_{\mu\mu} y_\mu}{1 - A_{\mu\mu}}$$

[Cook (1977), Seber & Lee (2003)]

- 特にガウス分布の場合，LOOCV errorは次の通り

$$\mathrm{err}_{\mathrm{LOOCV}}(\mathcal{D}) = \frac{1}{2M} \sum_{\mu=1}^{M} \left(\frac{y_\mu - \hat{y}_\mu(\mathcal{D})}{1 - A_{\mu\mu}(\mathcal{D})} \right)^2$$

- ■ フルデータ$\mathcal{D}$のもと１度だけ推定すれば良い
- ■ Linear estimatorのみ適用可能

Importance sampling cross validation error

- 正則化つき最尤法

$$\hat{\boldsymbol{x}}_\lambda(\mathcal{D}) = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, \varphi_\lambda(\mathcal{D},\boldsymbol{x}), \qquad \varphi_\lambda(\mathcal{D},\boldsymbol{x}) = \ln f\big(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{F},\boldsymbol{x})\big) - \underline{h(\boldsymbol{x};\boldsymbol{\lambda})}$$

正則化

- ベイズ推定(最大事後確率推定)で表す
  - 事後分布　$P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D};\beta) \propto f^\beta\big(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{F},\boldsymbol{x})\big)\exp\big(\beta h(\boldsymbol{x};\boldsymbol{\lambda})\big)$
  - 推定値の表現　$\hat{x}_{\lambda,i}(\mathcal{D}) = \lim_{\beta\to\infty}\int dx\, x_i P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D};\beta)$
- 対応するLOOCV error

$$\mathrm{err}_{\mathrm{LOOCV}}(\mathcal{D};\beta) = -\frac{1}{M}\sum_{\mu=1}^{M}\ln\int dx\, f^\beta\big(y_\mu|\theta_\mu(\mathbf{f}_\mu,\boldsymbol{x})\big)P_{\mathrm{post}}\big(\boldsymbol{x}|\mathcal{D}_{\backslash\mu};\beta\big)$$

Importance sampling cross validation error

$$
\begin{aligned}
\mathrm{err}_{\mathrm{LOOCV}}(\mathcal{D};\beta) &= -\frac{1}{M}\sum_{\mu=1}^{M}\ln\int dx\, f^{\beta}\big(y_\mu|\theta_\mu(\mathbf{f}_\mu,\boldsymbol{x})\big)P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D}_{\backslash\mu};\beta)\\
&= -\frac{1}{M}\sum_{\mu=1}^{M}\ln\int dx\, f\big(y_\mu|\boldsymbol{\theta}_\mu(\mathbf{f}_\mu,\boldsymbol{x})\big)\ \frac{\prod_{\nu\neq\mu} f^{\beta}(y_\nu|\theta_\nu(\mathbf{f}_\nu,\boldsymbol{x}))\exp\big(\beta h(\boldsymbol{x};\boldsymbol{\lambda})\big)}{\int dx\prod_{\nu\neq\mu} f^{\beta}(y_\nu|\theta_\nu(\mathbf{f}_\nu,\boldsymbol{x}))\exp\big(\beta h(\boldsymbol{x};\boldsymbol{\lambda})\big)}\\
&= \frac{1}{M}\sum_{\mu=1}^{M}\ln\frac{\int dx\prod_{\nu\neq\mu} f^{\beta}(y_\nu|\theta_\nu(\mathbf{f}_\nu,\boldsymbol{x}))\exp\big(\beta h(\boldsymbol{x};\boldsymbol{\lambda})\big)}{\int dx\prod_{\nu=1}^{M} f^{\beta}(y_\nu|\theta_\nu(\mathbf{f}_\nu,\boldsymbol{x}))\exp\big(\beta h(\boldsymbol{x};\boldsymbol{\lambda})\big)}\\
&= \frac{1}{M}\sum_{\mu=1}^{M}\ln\int dx P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D};\beta)\,\frac{1}{f^{\beta}\big(y_\mu|\theta_\mu(\mathbf{f}_\mu,\boldsymbol{x})\big)}
\end{aligned}
$$

この表現を**Importance Sampling Cross Validation error** と呼ぶ.

[Vehtari & Lampinen, *Neural Computat*ion (2002)]

Widely Applicable Information Criterion

- 次の関数を定義：$\mathcal{T}(\eta) = \dfrac{1}{M}\displaystyle\sum_{\mu=1}^{M} \ln \int d\boldsymbol{x} P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D};\beta) f^{\eta\beta}(y_\mu|\theta_\mu(\mathbf{f}_\mu,\boldsymbol{x}))$

  → $-\mathcal{T}(1)$：Training error, $\mathcal{T}(-1)$：ISCV

- $|\eta| \ll 1$での表現を適用した汎化ギャップの推定量 (**Functional Variance**)

$$\hat{\Delta}(\mathcal{D}) = \frac{\beta}{M}\sum_{\mu=1}^{M} \mathbb{V}_{\mathrm{post}}\left[\ln f\left(y_\mu|\theta_\mu(f_\mu,x)\right)\right]$$

$\mathbb{V}_{\mathrm{post}}[\cdot]$は事後分布における分散

  - 予測誤差の推定量「訓練誤差+FV」を **Widely applicable information criterion**とよぶ.

[Watanabe, *JMLR* (2010) & *JJSDS* (2021)]

■ 利点：計算量が少ない

■ 欠点：収束半径を考慮する必要がある

Linear estimatorのin-sample error

$$\mathrm{err}^{(\mathrm{in})}_{\mathrm{pre}}(\mathcal{D}) = \frac{1}{2\sigma^2 M}\mathbb{E}_{\boldsymbol{z}}[\|\boldsymbol{z} - \hat{\boldsymbol{y}}(\mathcal{D})\|_2^2]$$

$$\mathrm{err}_{\mathrm{train}}(\mathcal{D}) = \frac{1}{2\sigma^2 M}\|\boldsymbol{y} - \hat{\boldsymbol{y}}(\mathcal{D})\|_2^2$$

$$\Rightarrow \quad \Delta^{(\mathrm{in})}(\mathcal{D}) = \frac{1}{\sigma^2 M}\sum_{\mu=1}^{M}\mathrm{Cov}_{\boldsymbol{y}}\left(y_\mu, \hat{y}_\mu(\mathcal{D})\right)$$

[Efron (2004)]

- Linear estimatorの場合： $\mathrm{Cov}_{\boldsymbol{y}}\left(y_\mu, \hat{y}_\mu(\mathcal{D})\right) = \mathrm{Cov}_{\boldsymbol{y}}\left(y_\mu, \sum_\nu A_{\mu\nu} y_\nu\right)$
- さらに$y \sim \mathcal{N}(\mu, \sigma^2 I_M)$のとき$\mathrm{Cov}_{\boldsymbol{y}}\left(y_\mu, \sum_\nu A_{\mu\nu} y_\nu\right) = A_{\mu\mu}$なので

$$\Delta^{(\mathrm{in})}(\mathcal{D}) = \frac{1}{\sigma^2 M}\mathrm{Tr}(A)$$

…**Degrees of freedom**と呼ばれる

線形回帰問題では，Hat matrixの性質より$\mathrm{Tr}(H) = N$ =モデルパラメータ数

→ **Akaike Information criterion**と一致する

Beyond Linear estimator

$$\mathrm{err}^{(\mathrm{in})}_{\mathrm{pre}}(\mathcal{D}) = \frac{1}{2\sigma^2 M}\mathbb{E}_{\boldsymbol{z}}[\|\boldsymbol{z} - \hat{\boldsymbol{y}}(\mathcal{D})\|_2^2]$$

$$\mathrm{err}_{\mathrm{train}}(\mathcal{D}) = \frac{1}{2\sigma^2 M}\|\boldsymbol{y} - \hat{\boldsymbol{y}}(\mathcal{D})\|_2^2$$

$$\Rightarrow \quad \Delta^{(\mathrm{in})}(\mathcal{D}) = \frac{1}{\sigma^2 M}\sum_{\mu=1}^{M}\mathrm{Cov}_{\boldsymbol{y}}\left(y_\mu, \hat{y}_\mu(\mathcal{D})\right)$$

[Efron (2004)]

- $\boldsymbol{y} = \boldsymbol{\mu} + \sigma^2\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}_M)$のとき

Stein's Lemma： $\frac{1}{\sigma^2}\mathrm{Cov}_{\boldsymbol{y}}\left(y_\mu, \hat{y}_\mu(\mathcal{D})\right) = \mathbb{E}_{\boldsymbol{y}}\left[\frac{\partial}{\partial \varepsilon_\mu}\hat{y}_\mu(\mathcal{D})\right]$

$$\hat{\Delta}^{(\mathrm{in})}(\mathcal{D}) = \frac{1}{M}\sum_{\mu=1}^{M}\frac{\partial}{\partial \varepsilon_\mu}\hat{y}_\mu(\mathcal{D})$$

⇒ **Generalized degrees of freedom**と呼ばれる

- 利点：一般の推定量に適用可能
- 欠点：数値的微分が必要．ガウスノイズの場合のみ不偏

いろいろな予測誤差の推定量がある

必要とされること

(1) 効率的な計算方法

- 近似評価法が必要

(2) 推定量の間の関係性の理解

- 特に非漸近領域において

(3) 実用上の指針

- 近似評価法の精度，適用限界

- ファクターグラフ表現
- 近似確率伝搬法

を使って議論する

問題設定

一般化線形モデル(GLM)

- 回帰係数：$\boldsymbol{x} \in \mathbb{R}^N$
- 説明変数：$\boldsymbol{F} \in \mathbb{R}^{M \times N}$
- 出力：$\boldsymbol{y}$ (M次元ベクトル)

$\Big\}$ データセット$\mathcal{D} = \{\boldsymbol{y}, \boldsymbol{F}\}$

- 対数尤度 $\ln f(\boldsymbol{y}|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{y} - a(\boldsymbol{\theta}) + b(\boldsymbol{y})$
  - $\theta_\mu = \dfrac{1}{\sqrt{N}} \sum_{i=1}^{N} F_{\mu i} x_i, \qquad a(\boldsymbol{\theta}) = \sum_{\mu=1}^{M} a(\theta_\mu), \qquad b(\boldsymbol{y}) = \sum_{\mu=1}^{M} b(y_\mu)$
  - $\mathbb{E}[y_\mu] = \dfrac{\partial a(\theta_\mu)}{\partial \theta_\mu}, \qquad \mathbb{E}\left[\left(y_\mu - \mathbb{E}[y_\mu]\right)^2\right] = \dfrac{\partial a(\theta_\mu)}{\partial \theta_\mu^2}$

□ 線形回帰：$a(\theta) = \frac{1}{2}\theta^2, b(y) = -\frac{1}{2}y^2$

□ ロジスティック回帰： $a(\theta) = \ln\left(1 + e^\theta\right), b(y) = 0$

制約付き最尤推定

$$\hat{\boldsymbol{x}}_\lambda(\mathcal{D}) = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, \varphi_\lambda(\mathcal{D}, \boldsymbol{x}), \qquad \varphi_\lambda(\mathcal{D}, \boldsymbol{x}) = \ln f\big(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{F}, \boldsymbol{x})\big) - \underbrace{h(\boldsymbol{x}; \boldsymbol{\lambda})}_{\text{正則化}}$$

- 正則化パラメータ$\boldsymbol{\lambda}$を選ぶことがモデル選択に対応する
  - モデル候補：$\mathcal{M} = \left\{f\big(\boldsymbol{Y}|\boldsymbol{\theta}\big(\boldsymbol{F}, \hat{\boldsymbol{x}}_1(\mathcal{D})\big)\big), f\big(\boldsymbol{Y}|\boldsymbol{\theta}\big(\boldsymbol{F}, \hat{\boldsymbol{x}}_2(\mathcal{D})\big)\big), \ldots\right\}$
- 定式化：最大事後確率を用いた推定として表現
  - 事後分布　$P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D}; \beta) \propto f^\beta\big(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{F}, \boldsymbol{x})\big) \exp\big(\beta h(\boldsymbol{x}; \boldsymbol{\lambda})\big)$
  - 推定値の表現　$\hat{x}_{\lambda,i}(\mathcal{D}) = \lim_{\beta\to\infty} \int dx\, x_i P_{\mathrm{post}}(\boldsymbol{x}|\mathcal{D}; \beta)$

■ 尤度⟷ハミルトニアン，事前分布⟷外場

■ 最大事後確率推定⟷基底状態探索

近似確率伝搬法

Generalized Approximate Message Passingとは

Generalized Approximate Message Passingとは

マルコフネットワーク(特に**ファクターグラフ**)上で定義されるアルゴリズム

●目的：変数の周辺化（分配関数や周辺化分布の評価）

■ Message Passing

- グラフ構造が規定する条件付き独立性のもと，効率的に周辺化
  - “メッセージ”の更新により計算

■ Approximate Message Passing

- メッセージの一部をガウス近似したmessage passing

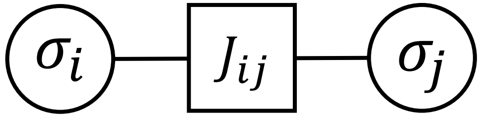
■ Generalized Approximate Message Passing

- 一般の尤度と事前分布に適用可能なAMP

ファクターグラフ表現の例

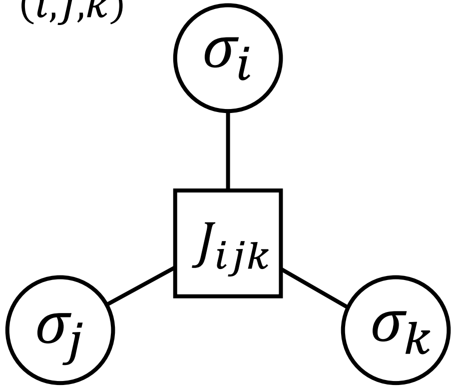
- 2-body spin-glass model

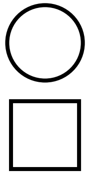
$$P(\boldsymbol{\sigma}) \propto \exp\left(\beta \sum_{(i,j)} J_{ij}\sigma_i\sigma_j\right)$$
$$\equiv \prod_{(i,j)} \psi_{ij}(\sigma_i, \sigma_j; J_{ij})$$



- 3-body spin-glass model with external field

$$P(\boldsymbol{\sigma}) \propto \exp\left(\beta \sum_{(i,j,k)} J_{ijk}\sigma_i\sigma_j\sigma_k\right)$$
$$\equiv \prod_{(i,j,k)} \psi_{ijk}(\sigma_i, \sigma_j, \sigma_k; J_{ijk})$$





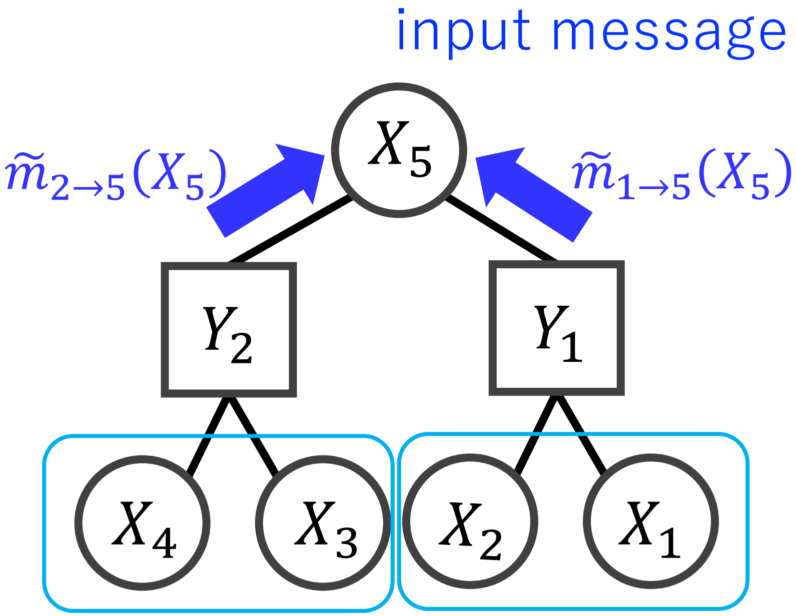
: 変数ノード

: ファクターノード

- Message Passingの手続き
  - ファクターψをかける
  - 変数の和を実行する

ファクターグラフ上でのSum-Product

- 例: $P(X_5|\boldsymbol{Y}) = \sum_{X_4} \cdots \sum_{X_1} P(\boldsymbol{X}|\boldsymbol{Y})$を計算する



$$P(\boldsymbol{X}|\boldsymbol{Y}) \propto \psi_1(X_1, X_2, X_5; Y_1)\psi_2(X_3, X_4, X_5; Y_2)$$

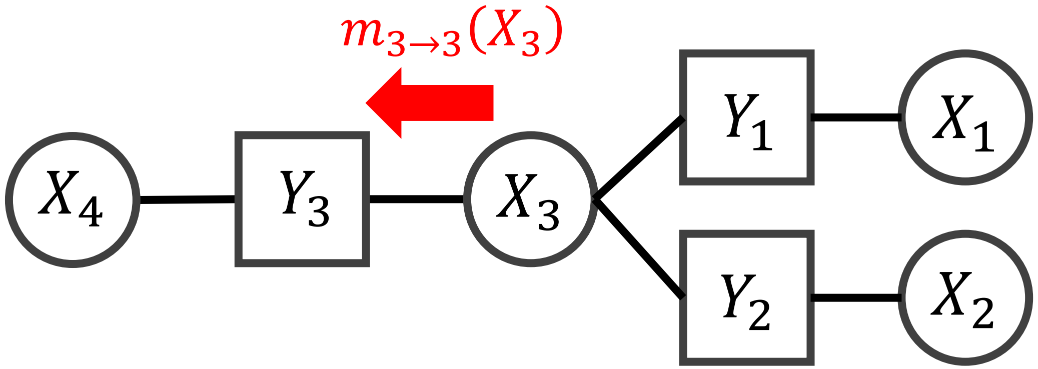
$$P(X_5|\boldsymbol{Y}) = \underbrace{\boxed{\sum_{X_3,X_4} \psi_2(X_3, X_4, X_5; Y_2)}}_{\tilde{m}_{2\to5}(X_5)} \underbrace{\boxed{\sum_{X_1,X_2} \psi_1(X_1, X_2, X_5; Y_1)}}_{\tilde{m}_{1\to5}(X_5)}$$

  - 周辺化分布はinput messageの積により与えられる.

詳しくは→

ファクターグラフ上でのSum-Product

- 例: $P(X_4|\boldsymbol{Y}) = \sum_{X_3}\sum_{X_2}\sum_{X_1} P(\boldsymbol{X}|\boldsymbol{Y})$を計算する



$$P(X_4|\boldsymbol{Y}) = \underbrace{\sum_{X_3} \psi_3(X_3,X_4;Y_3) \underbrace{\underbrace{\sum_{X_2}\psi_2(X_2,X_3;Y_2)}_{\tilde{m}_{2\to3}(X_3)} \underbrace{\sum_{X_1}\psi_1(X_1,X_3;Y_1)}_{\tilde{m}_{1\to3}(X_3)}}_{m_{3\to3}(X_3)}}_{\tilde{m}_{3\to4}(X_4)}$$

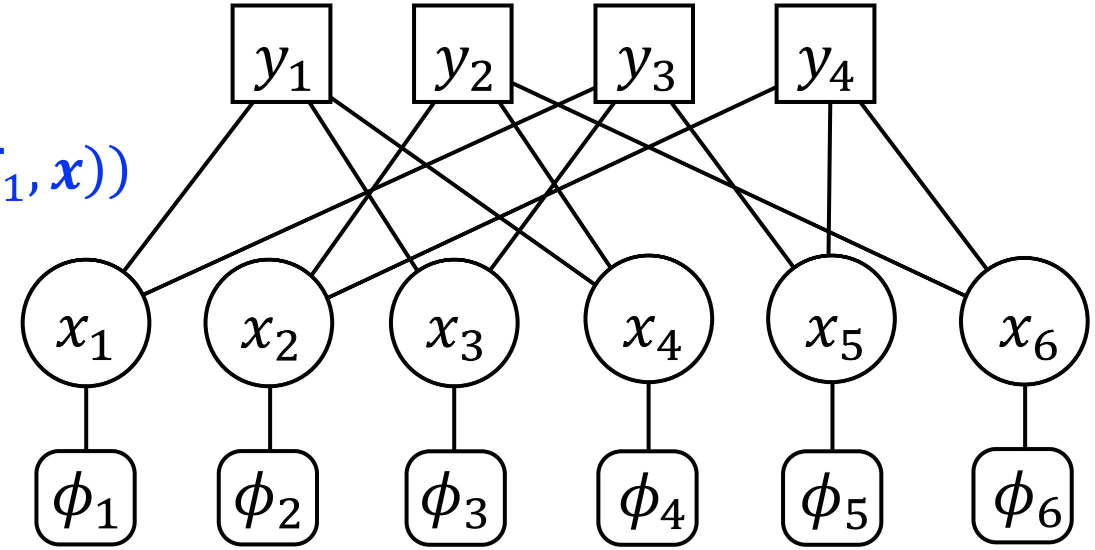
  - $m_{i\to\mu}(X_i)$: i番目の変数ノードからμ番目のファクターノードへのoutput message

Message PassingのGLMへの適用

- 事後分布　$P_{\text{post}}(\boldsymbol{x}|\mathcal{D};\beta) \propto f^{\beta}\big(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{F},\boldsymbol{x})\big)\exp\big(\beta h(\boldsymbol{x};\boldsymbol{\lambda})\big)$
- 推定値の表現　$\hat{x}_{\lambda,i}(\mathcal{D}) = \lim_{\beta\to\infty}\int dx\, x_i P_{\text{post}}(\boldsymbol{x}|\mathcal{D};\beta)$

ファクター＝
尤度$f^{\beta}(y_1|\theta_1(\mathbf{f}_1,\boldsymbol{x}))$

- グラフの構造は
説明変数から決まる



Montanari, “*Graphical Models Concepts in Compressed Sensing*”

arXiv:1011.4328

Output messageの意味

- ファクターグラフがツリーであるとする.
- 変数ノードiとファクターノードa間のエッジを切り離したとする.
- このとき変数ノードiが含まれるツリーを$\mathbb{T}(i \setminus a)$とする.

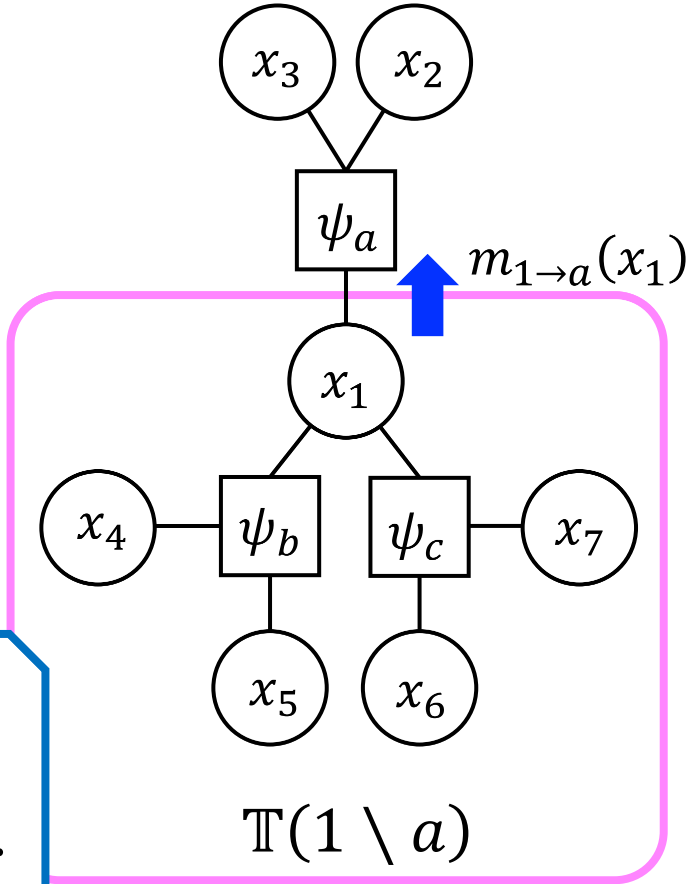
Output message $m_{i\to a}(x_i)$は

$\mathbb{T}(i \setminus a)$における$x_i$の厳密な周辺化分布である.

Output message $m_{i\to a}(x_i)$は

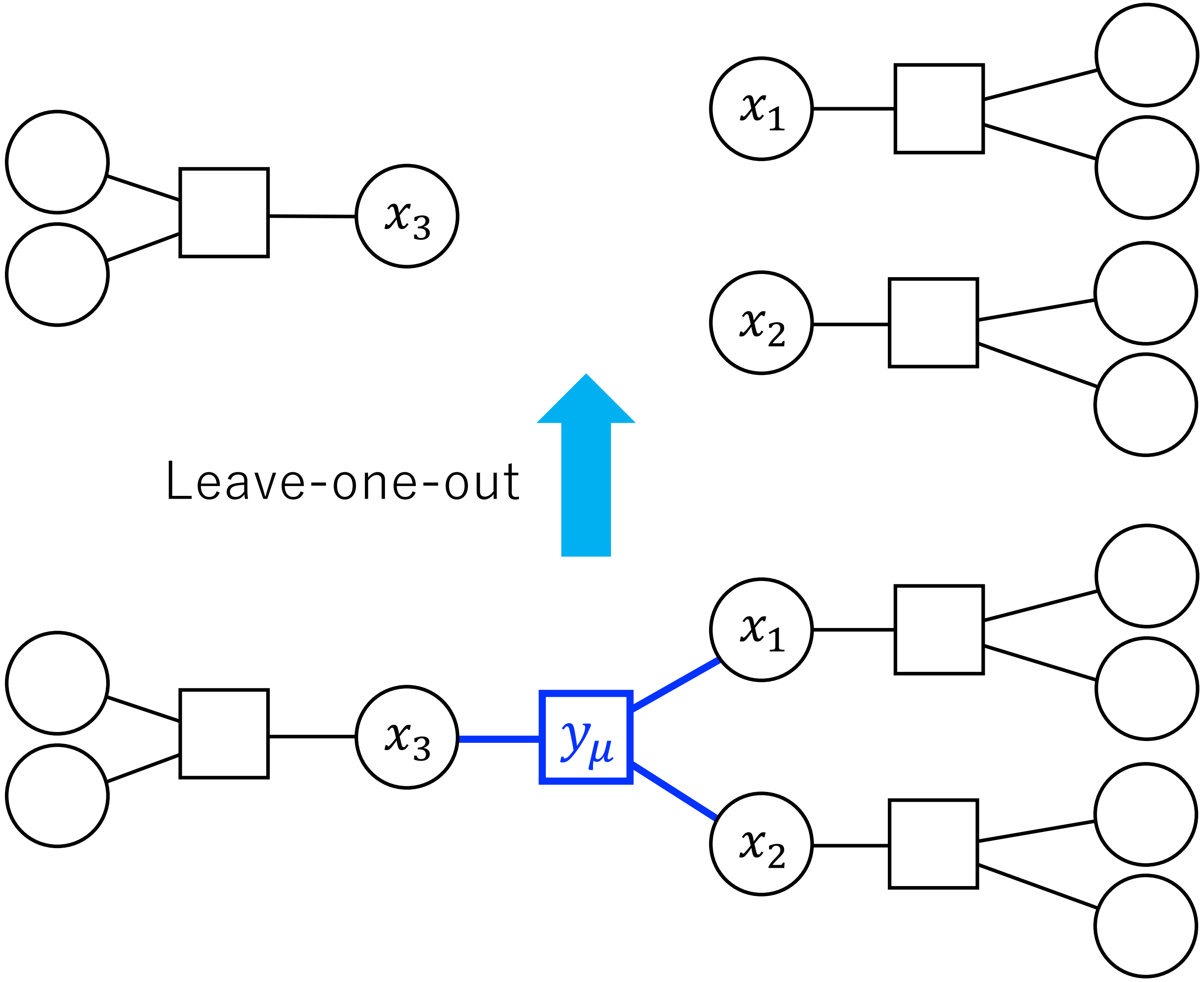
LOO sample $\mathcal{D}_{\setminus a}$のもとでの$x_i$の厳密な周辺化分布である.

- $\hat{x}_{i\to a} \equiv \lim_{\beta\to\infty} \int dx_i\, x_i m_{i\to a}(x_i)$は$\mathcal{D}_{\setminus a}$の下での推定値



[Mezard & Montanari, (2009)]

ファクターグラフによるLOOサンプル下での推定の表現



メッセージを利用すれば，
LOO sampleのもとで
M回の推定を行わずに，
LOOCV errorの近似値を
得ることが出来る

ツリーではない場合

- ツリーでないグラフ（ループを含むグラフ）に
  Message Passingを適用することも可能

その場合，グラフがツリーであると仮定したうえで,
真の分布とのKullback-Leibler divergenceを最小化するような
“試行”周辺化分布を求めている と解釈できる

  - Bethe-Peierls近似に相当

- 厳密性の保証はないが，近似計算法として
  妥当な精度が得られる場合がある.

  - ターボ符号，低密度パリティ検査符号など [Gallager (1962), MacKay (1999)]

Message PassingからApproximate Message Passingへ

- Input message

$$\tilde{m}_{\mu\to i}(x_i) \propto \int d\boldsymbol{x}_{\mathrm{scope}(\mu)\backslash i}\, f^{\beta}\left(y_\mu | \theta_\mu(\mathbf{f}_\mu, \boldsymbol{x})\right) \prod_{j\in\mathrm{scope}(\mu)\backslash i} m_{j\to\mu}(x_i)$$

⬇ **ガウス近似**

$$\tilde{m}_{\mu\to i}(x_i) \propto \exp\left(-\frac{\beta}{2\tilde{s}_{\mu\to i}}\left(x_i - \tilde{x}_{\mu\to i}\right)^2\right)$$

※ $\tilde{x}_{\mu\to i}$は平均，　$\beta^{-1}\tilde{s}_{\mu\to i}$は分散

- Output message

$$m_{i\to\mu}(x_i) \propto \phi_i(x_i;\boldsymbol{\lambda}) \prod_{\eta\in\mathcal{F}(i)\backslash\mu} \tilde{m}_{\eta\to i}(x_i) \propto \phi_i(x_i;\boldsymbol{\lambda}) \exp\left(-\sum_{\eta\in\mathcal{F}(i)} \frac{\beta}{2\tilde{s}_{\eta\to i}}\left(x_i - \tilde{x}_{\eta\to i}\right)^2\right)$$

⬇ 平均と分散を評価

$$\hat{x}_{i\to\mu} = \int dx_i\, m_{i\to\mu}(x_i), \qquad s_{i\to\mu} = \beta \int dx_i\, m_{i\to\mu}(x_i)\left(x_i - \hat{x}_{i\to\mu}\right)^2$$

$\tilde{x}_{\mu\to i}$と$\tilde{s}_{\mu\to i}$，$\hat{x}_{i\to\mu}$と$s_{i\to\mu}$の更新アルゴリズムを**Approximate Message Passing**と呼ぶ.

AMPの使用が妥当な場合

次の性質が満たされるとき，ガウス近似は妥当である．

i. 変数の数Nが十分大きい

ii. 説明変数の各成分が$O(N^{-1/2})$

- 説明変数が標準化(standardization)されている

iii. 説明変数間の相関が無視できるほど小さい

- 多重共線性がないように説明変数を設計

iv. 説明変数がスパースではない

i～ivの仮定のもと，中心極限定理からAMPの妥当性は保証される．

これから示す結果

- GAMPの適用可能な範囲で，一般化線形モデルにおける次の結果を紹介

■ 一般化自由度は熱揺らぎから評価可能

■ Linear estimation ruleにおけるLOOCV error表現の拡張を実現する

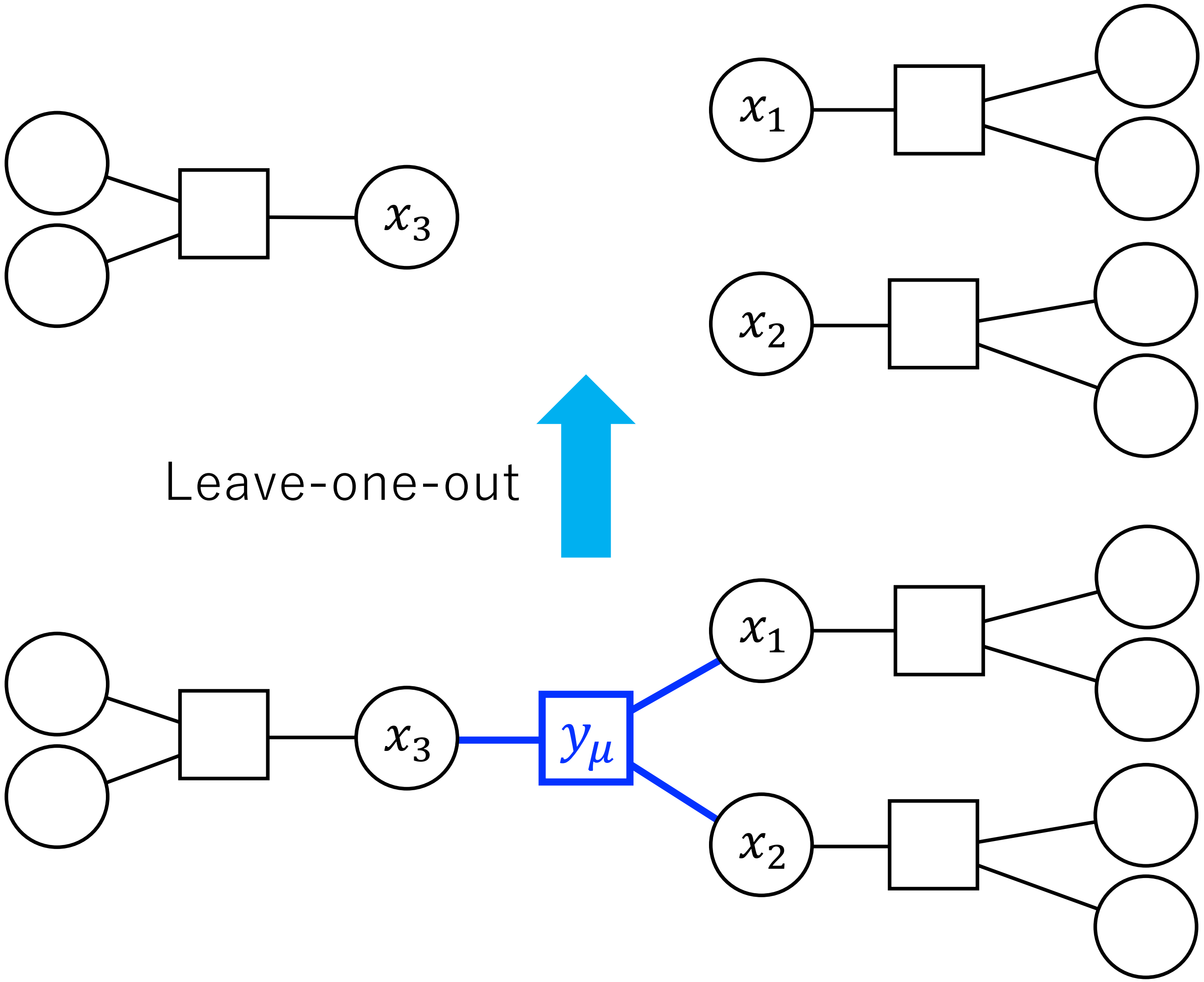
■ WAIC導出の根拠となる，ISCVの重みに関する級数展開は
線形ガウスモデルにおいて スピングラス転移点で発散する

- 一般化線形モデルで収束半径を導出することも可能

● GAMPの根拠となる仮定が破れているときについては最後に少し説明

GAMPによる予測誤差の表現

ファクターグラフによるLOOサンプル下での推定の表現



y_μがないグラフでの
x_1の期待値：$\hat{x}_{1\to\mu}$

(メッセージ$m_{1\to\mu}$
の期待値)

LOOCV error

- 一般化線形モデルにおける
  LOO sampleのもとでの推定値$\hat{\theta}_\mu^{\backslash\mu} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} F_{\mu i}\hat{x}_{i\to\mu}$と
  フルデータのもとでの推定値$\hat{\theta}_\mu = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} F_{\mu i}\hat{x}_i$の関係

$$\frac{\hat{\theta}_\mu - \hat{\theta}_\mu^{\backslash\mu}}{V_\mu} = y_\mu - a'(\hat{\theta}_\mu),\ V_\mu = \frac{1}{N}\sum_{i=1}^{N} F_{\mu i}^2 s_{i\to\mu} \xrightarrow{\text{代入}} \mathrm{err}_{\mathrm{LOOCV}}(\mathcal{D}) = -\frac{1}{M}\sum_{\mu=1}^{M}\ln f\left(y_\mu|\hat{\theta}_\mu^{\backslash\mu}\right)$$

- 特にガウス分布の場合

$$y_\mu - \hat{y}_\mu^{\backslash\mu} = (y_\mu - \hat{y}_\mu)(1 + V_\mu)$$

Linear estimator $\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$での表現

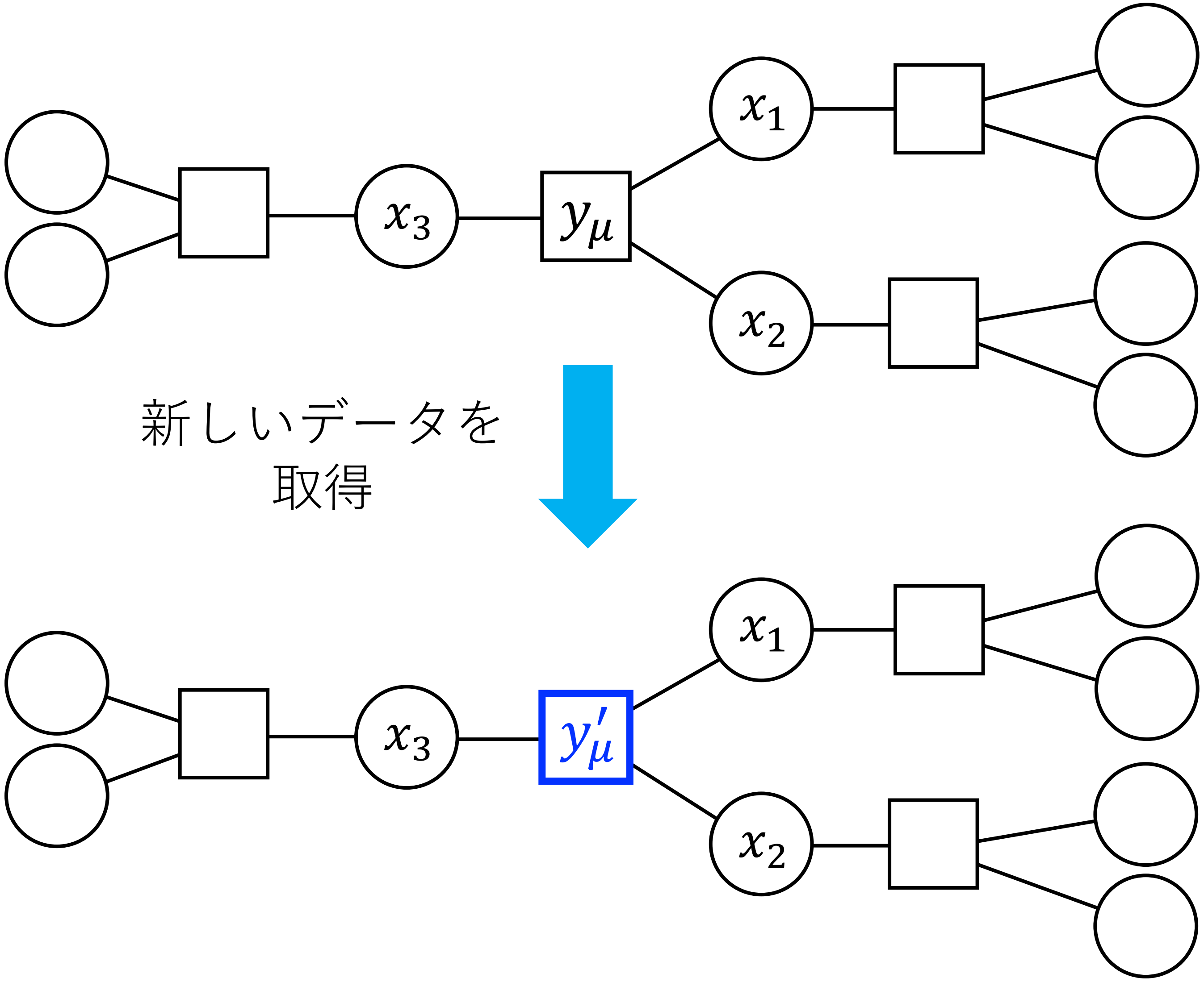
$$y_\mu - \hat{y}_\mu(\mathcal{D}_{\backslash\mu}) = \frac{y_\mu - \hat{y}_\mu(\mathcal{D})}{1 - A_{\mu\mu}}$$

Linear estimation ruleにおける表現の一般化

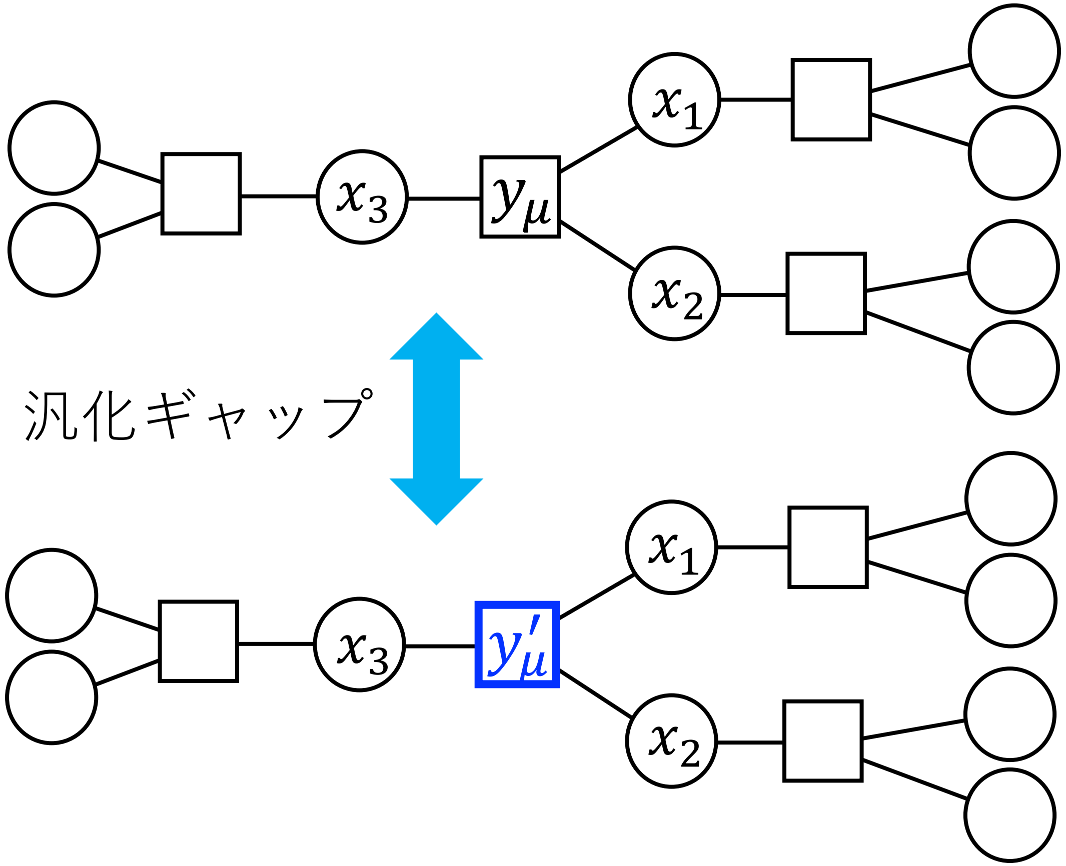
$$A_{\mu\mu} \leftrightarrow \frac{V_\mu}{1 + V_\mu} = \lim_{\beta\to\infty}\beta\left(\langle\theta_\mu^2\rangle_{\theta_\mu} - \langle\theta_\mu\rangle_{\theta_\mu}^2\right)$$

※ $\langle\cdot\rangle_{\theta_\mu}$ ：Scope[μ]の局所平均

ファクターグラフによるin-sample errorの表現

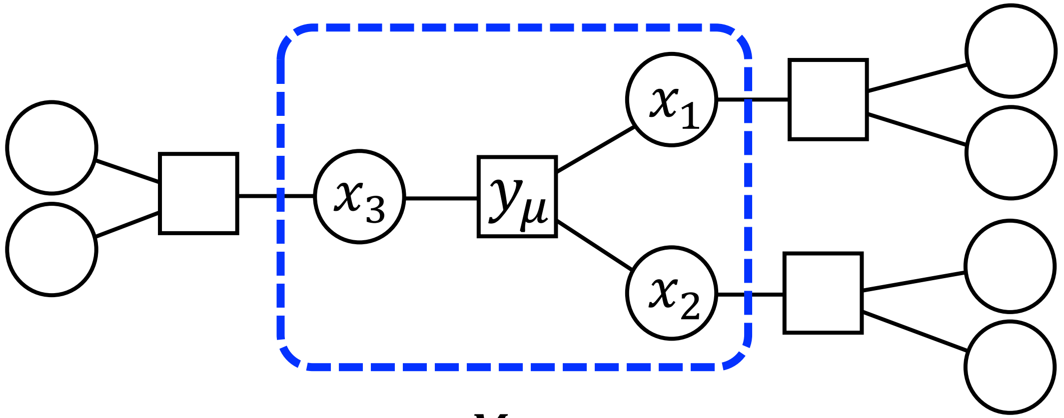


GAMPによる一般化自由度の推定量



$y_\mu \to y'_\mu$の変化に伴う

汎化ギャップ∝ Scope[μ]の “**熱揺らぎ**”



$$\hat{\Delta}^{(\mathrm{in})} = \lim_{\beta\to\infty} \frac{\sigma^2\beta}{M} \sum_{\mu=1}^{M} \left(\langle \theta_\mu^2 \rangle_{\theta_\mu} - \langle \theta_\mu \rangle_{\theta_\mu}^2 \right)$$

※ $\langle\cdot\rangle_{\theta_\mu}$ ：Scope[μ]の局所平均

- Linear estimatorにおける$\hat{\Delta}^{(\mathrm{in})} = \mathrm{Tr}(A)$の拡張

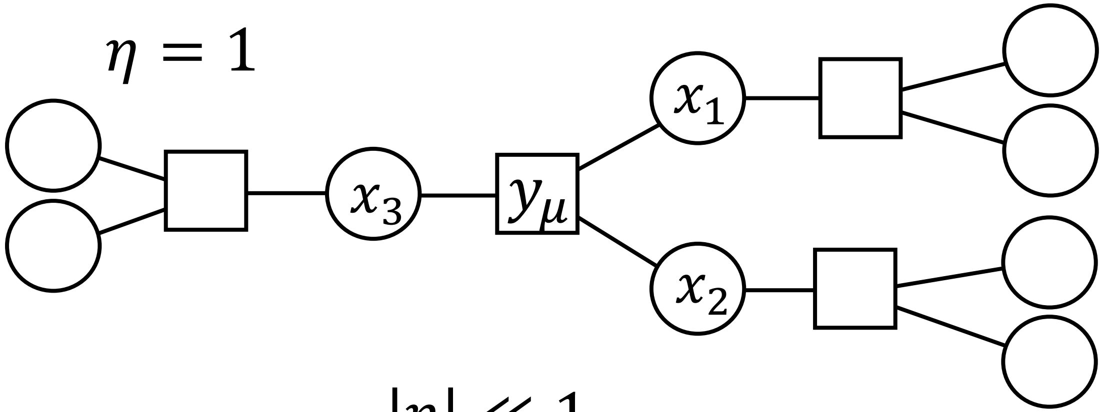
「予測における揺らぎと応答」 [Iba (private communication), Watanabe *JJSDS* (2021)]

ファクターグラフによるISCVの表現

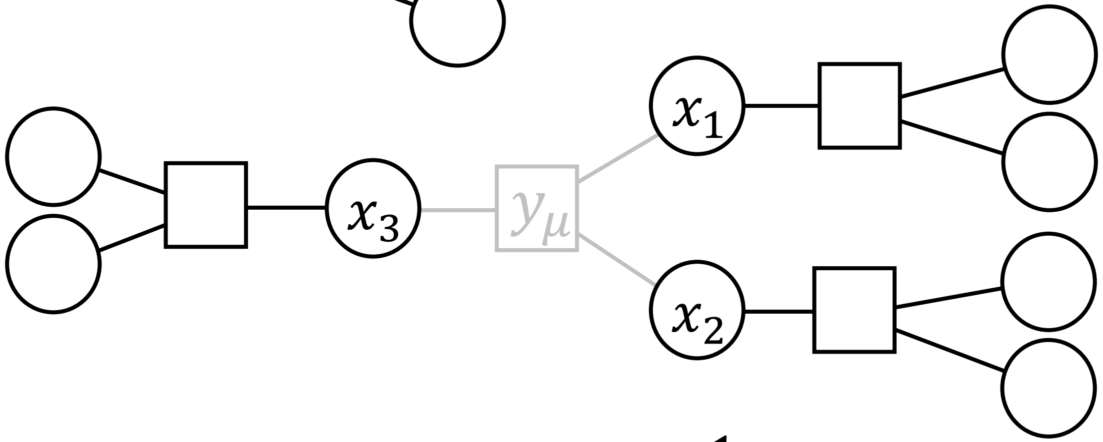
WAICの導出：

$|\eta| \ll 1$での汎化ギャップを
$|\eta| = 1$に適用

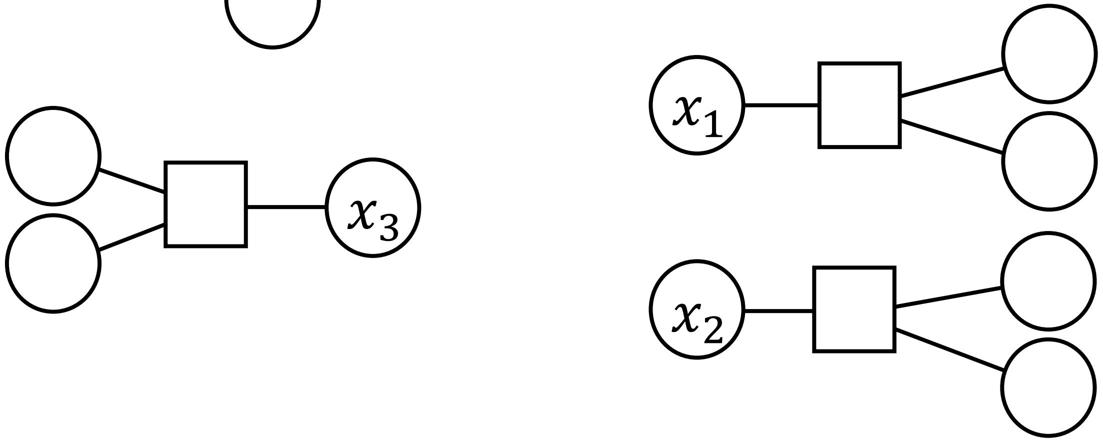
$\eta = 1$



$|\eta| \ll 1$

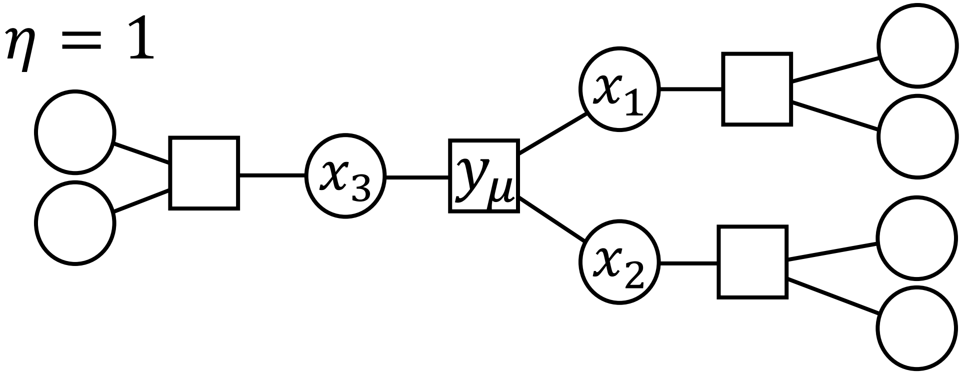


$\eta = -1$

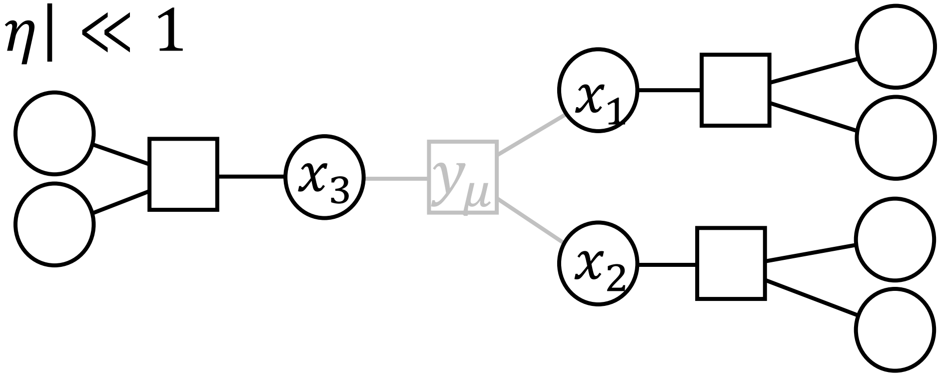


GAMPによるWAICの表現と適用限界

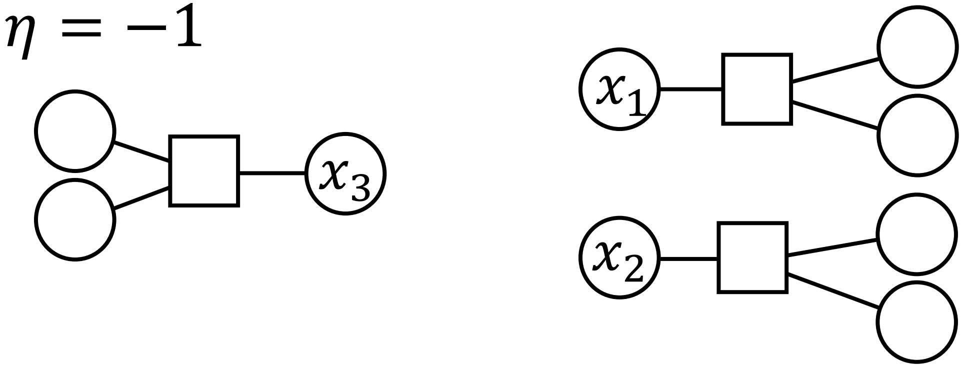
$\eta = 1$



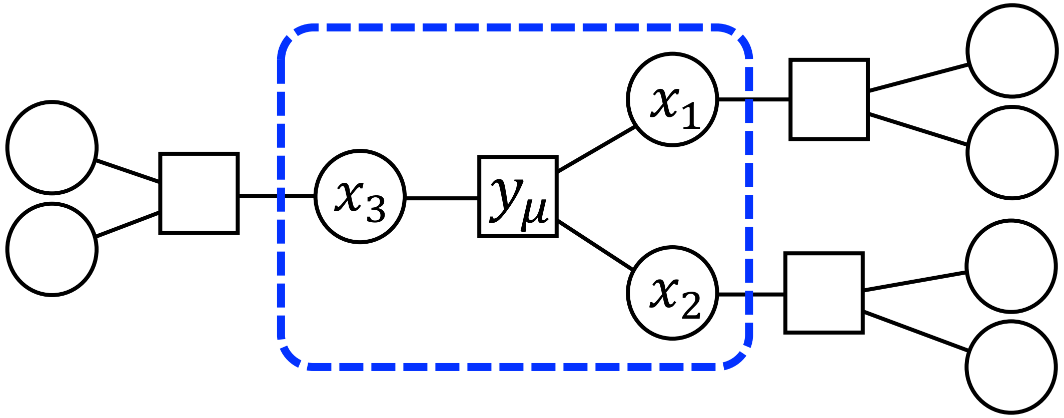
$|\eta| \ll 1$



$\eta = -1$



汎化ギャップは，$|\eta| \ll 1$のとき
Scope[μ]の “**熱揺らぎ**”と**残差**から評価可能

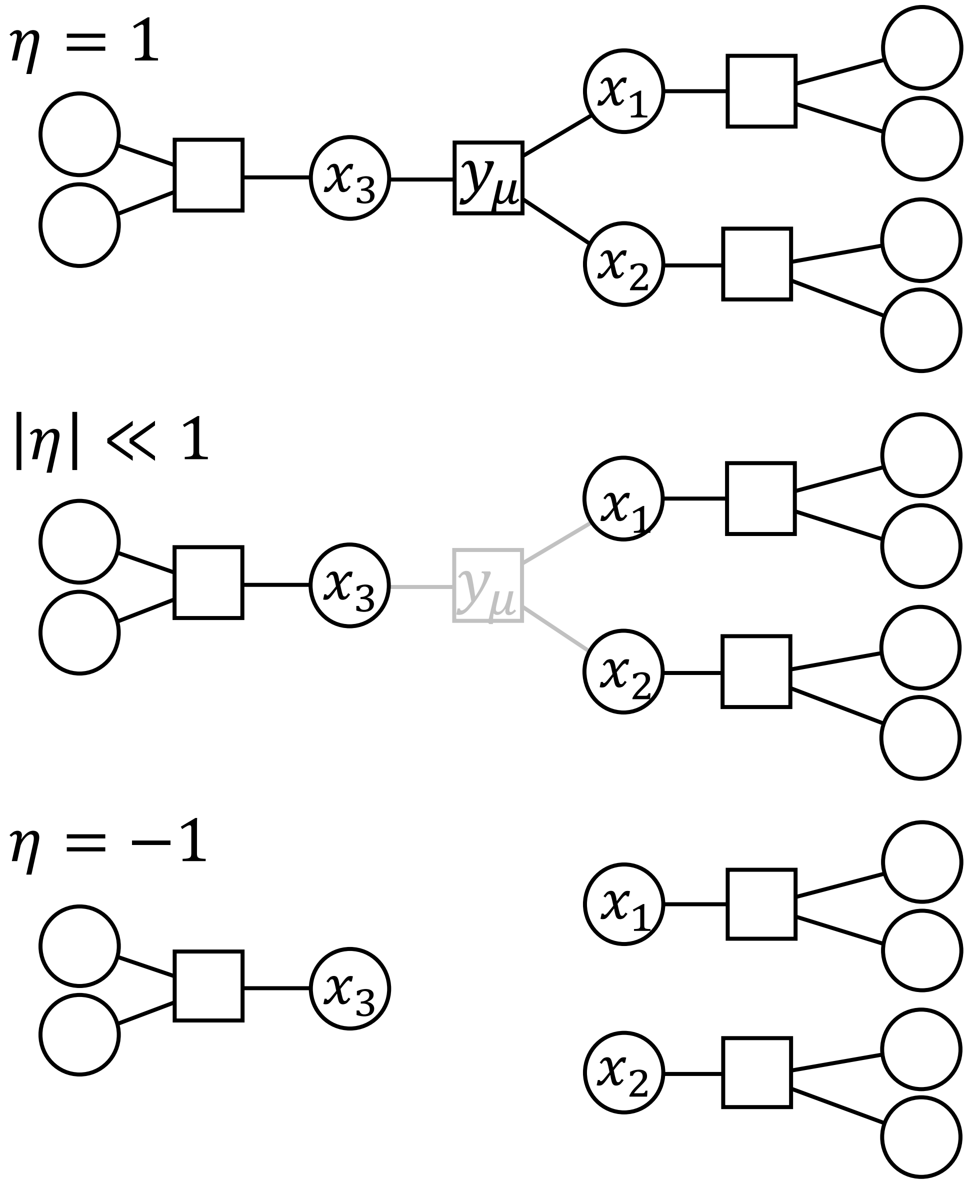


$$\hat{\Delta} = \lim_{\beta\to\infty} \frac{\beta}{M} \sum_{\mu=1}^{M} (y_\mu - \hat{y}_\mu)^2 \left(\langle \theta_\mu^2 \rangle_{\theta_\mu} - \langle \theta_\mu \rangle_{\theta_\mu}^2 \right)$$

- σ^2を推定量(残差^2)で表した一般化自由度

※ σ^2を推定量で置き換える方法の提案 [Efron (2004)]

GAMPによるWAICの表現と適用限界



尤度がガウス分布の場合,
ISCVの級数展開は次の条件が成立するとき収束する

$$|\eta| < \frac{1}{\frac{V}{1+V}}, \qquad V = \frac{1}{M}\sum_{\mu=1}^{M} V_\mu$$

$|\eta| = 1$で発散しないためには $\frac{V}{1+V} < 1$

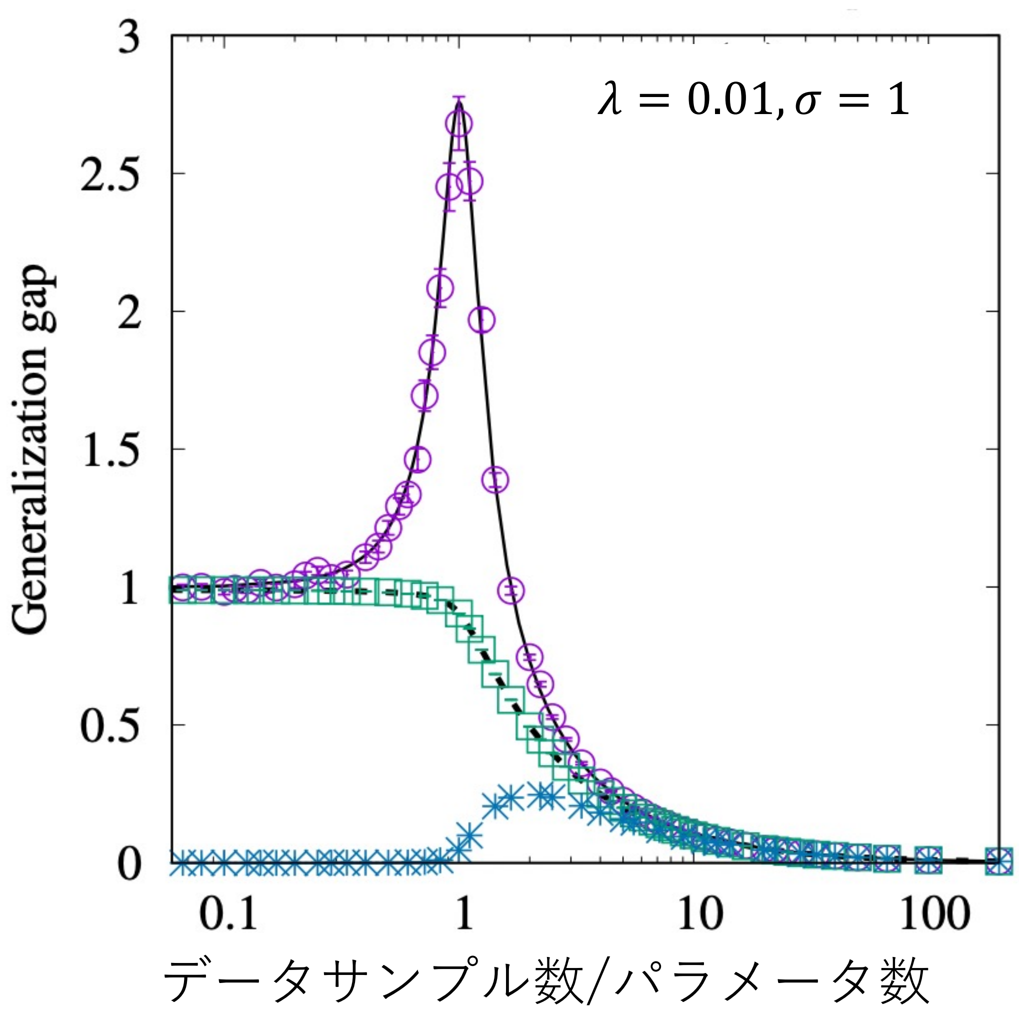
これは**スピングラス転移点**に対応する.

意味：

一般化自由度<データ数の場合に適用可能

※ 正則化による推定値の縮小がある場合は必ずしもモデルパラメータ>データ数で発散するわけではない

実際の値とGAMPの比較：Ridge回帰の場合



—— LOOCV errorの汎化ギャップ

------ In-sample errorの汎化ギャップ

○ GAMPによるLOOCV errorの汎化ギャップ

□ GAMPによるIn-sample errorの汎化ギャップ

* GAMPによるFV

- GAMPは実際の値を精度良く評価できている.
- 漸近極限ではすべての推定量は一致するが,
  データサンプル数が小さくなるにつれて
  ずれる様子が見える.
- Logistic回帰などでも同様.

GAMPにより得られたこと

(1) 汎化ギャップの効率的な計算

- 汎化ギャップを評価するための追加の計算が必要ない
  - GAMPによる表現は，Linear estimatorに対する表現の拡張と理解できる

(2) 「揺らぎと応答」としての汎化ギャップの表現

- $y_\mu \to y'_\mu$による汎化ギャップ $\propto \mathrm{Scope}[\mu]$に対する事後分布の分散
  - 未知データへの応答は，事後分布の性質から把握できる

(3) ISCVの吸数展開の収束半径

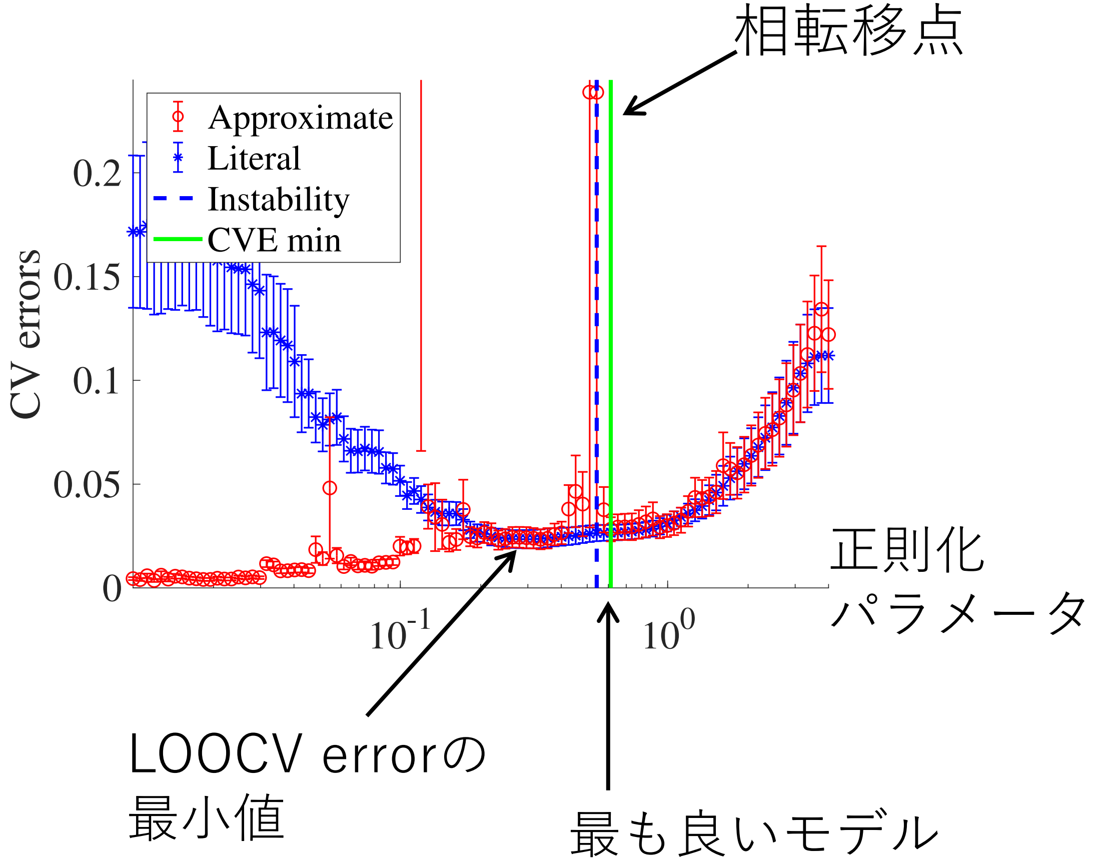
- Gaussian likelihoodの場合はスピングラス転移が起こらなければWAICは妥当
  - そのほかのlikelihoodの場合にも収束半径は導出できる

Sakata, J Phys A (2023), (arXiv:2206.12832)

実データへの適用について

- 実データにおいては，GAMPの仮定は必ずしも満たされない

- 説明変数に相関がある場合の補正
  - 分散共分散行列を正確に評価する方法
    - Obuchi & Kabashima, JSTAT (2016)
  - GAMPを利用した“相転移点”の発見とモデル候補からの除外
    - Obuchi & Sakata, JSTAT (2019)



おわりに

■ 統計学・機械学習におけるモデリングと
物理学におけるモデリングの考え方の違い

- Ground truthがない問題に折り合いをつけるための客観的方法

■ その方法の中にも“物理”がある

- モデルを作る＝ハミルトニアンの設計
- 正則化つき最尤推定＝外場のある問題での基底状態探索
- 推定アルゴリズム＝ダイナミクス
- 汎化ギャップ＝相互作用，温度に関する摂動への応答

■ 推論方法の開発，計算の高度化 → 統計学・機械学習の深化へ

- 特に統計力学は非漸近領域の記述が得意