本日の内容

6.3. Model Validation (モデルの検証)

6.3.1 Predictive Model Validation (予測モデルの検証)
6.3.2 Spatio-Temporal Validation Statistics (時空間モデルの検証のための統計値)
6.3.3 Spatio-Temporal Cross-Validation Measures(時空間モデルの交差検証のmeasure)
6.3.4 Scoring Rules (スコアリングのルールを用いた検証)
6.3.5 Field Comparison　(フィールドの比較による検証)

6.4. Model Selection (モデルの選択)

6.4.1 Model Averaging
6.4.2 Model Comparison via Bayes Factors
6.4.3 Model Comparison via Validation
6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up (6章まとめ)

まえおき　Model validationとは

- 観察データをもとに構築したモデルが，どれだけ現実世界のプロセスを反映できているかを確認すること．
- つまり，ここで紹介する様々な診断指標や感度テストを用いて仮定の正しさを調べることで，実世界でのモデルの有効性を調べる．

6.3.1 Predictive Model Validation（予測モデルの検証）p268-

（最もシンプルなケース）

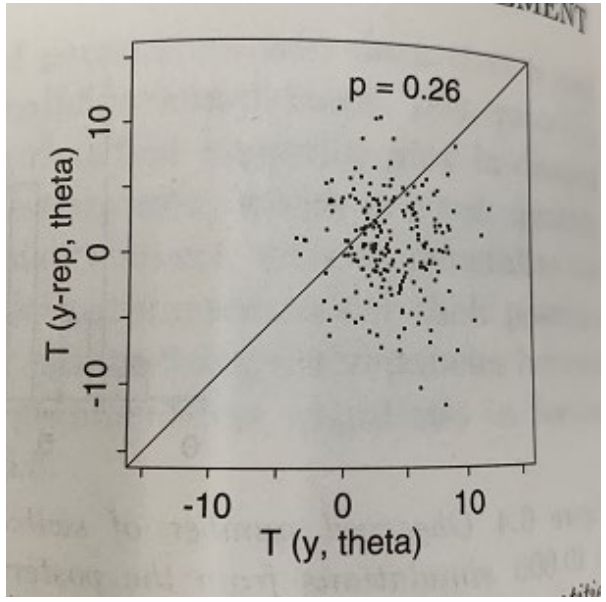
・予測モデルからの事後予測分布（ppd: posterior predictive distribution）や経験的予測分布（epd: empirical predictive distribution）から抽出したサンプルが，どれだけ観察データに近いかを評価する．この評価をするための指標を**predictive diagnostics**と言い，特に事後予測分布からのサンプルを評価する指標を**posterior predictive diagnostics**と呼び，経験的予測分布のそれを**empirical predictive diagnostics**と呼ぶ．

（posterior predictive diagnostics　事後予測診断指標の例）

・**Discrepancy measure $T(\cdot)$**（逸脱度の指標）

↳何の逸脱度をみるかはモデラーが指定．（例：overall fit，scoring rule, etc.）

$T(Z_p; Y, \theta)$（→事後分布サンプルからのDiscrepancy measure）と$T(Z; Y, \theta)|Z$（→観察データからのDiscrepancy measure）を比較する



- **Posterior predictive p-value p_B**を計算

$$p_B = \Pr\left(T(Z_p; Y, \theta) \geq T(Z; Y, \theta)|Z\right)$$

観察データZのもとで，予測分布からのサンプルが観察データを超える確率

→　45度線にあれば，観察データに近い値が得られてる．

→　p値が 0 や1だと，モデルは上手く行っていない　．

㊟　p値は予備的な診断材料であり，正式な統計テストではない！

- **Discrepancy measureを用いた例**（シドニーのレーダのデータとIDEモデル500回サンプリング）

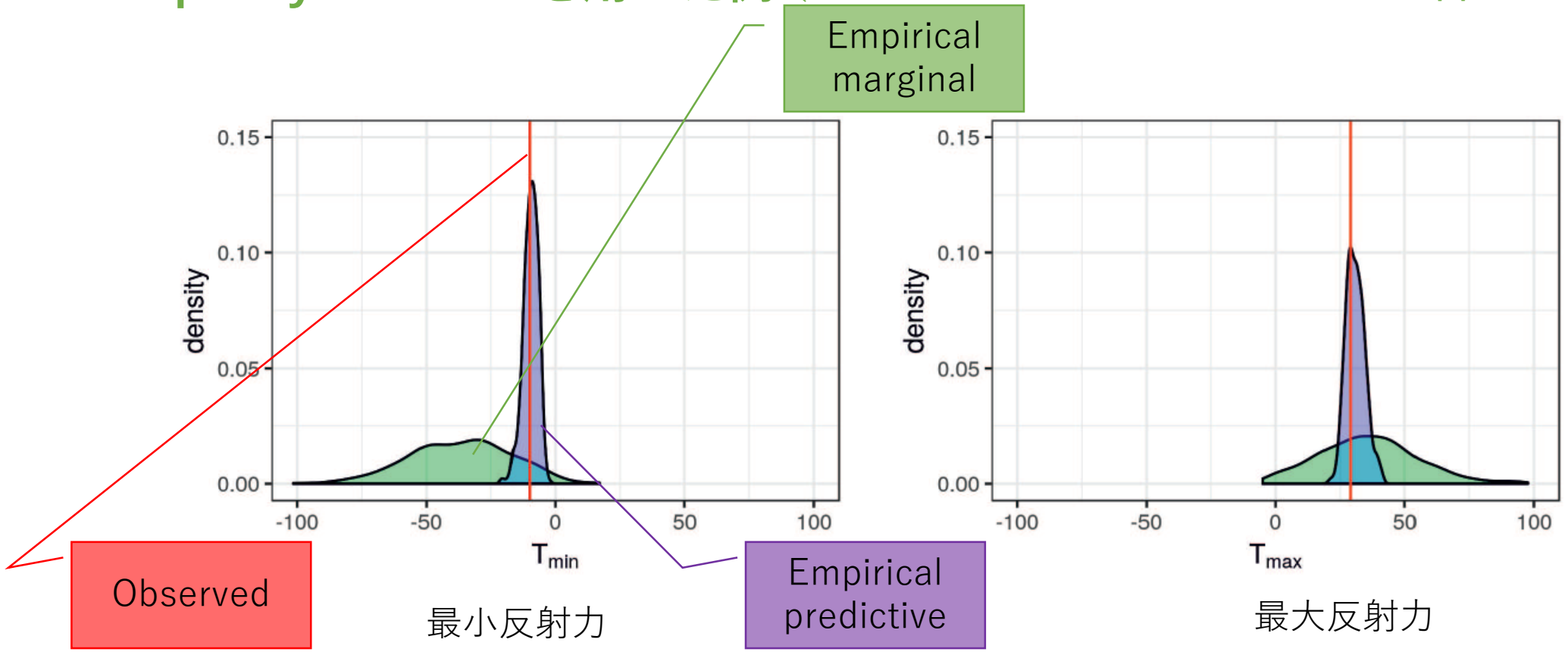


Figure 6.6: Empirical marginal distribution (green) and empirical predictive distribution (blue) densities for the minimum ($T_{\min}$, left) and maximum ($T_{\max}$, right) radar reflectivities across all grid boxes with centroid at $s_1 = 26.25$ (i.e., a vertical transect) for the times shown in Figure 6.1. In both panels, the red line denotes the observed statistic.

P値：0.09（緑），0.442（紫）　　P値：0.364（緑），0.33（紫）

P値はどのケースでも0.05より大きい→reasonable fit

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）

6.3.2 Spatio-Temporal Validation Statistics（時空間モデルの検証のための統計値）

6.3.3 Spatio-Temporal Cross-Validation Measures(時空間モデルの交差検証のmeasure)

6.3.4 Scoring Rules（スコアリングのルールを用いた検証）

6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）

6.4.1 Model Averaging

6.4.2 Model Comparison via Bayes Factors

6.4.3 Model Comparison via Validation

6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

6.3.2 Spatio-Temporal Validation Statistics　（時空間モデルの検証のための統計値）　p269-

1. MSPE (Mean Squared Prediction Error: 予測の平均二乗誤差)
2. RMSPE (Root Mean Squared Prediction Error: 平均二乗平方誤差)
3. MAPE (Mean Absolute Prediction Error: 平均絶対値予測誤差)
4. ACC (Anomaly Correlation Coefficient：アノマリー相関係数)
5. その他の簡易な指標

要は色々ある！

1/5. MSPE (Mean Squared Prediction Error: 予測の平均二乗誤差）最も一般的

$$MSPE = \frac{1}{Tm}\sum_{j=1}^{T}\sum_{i=1}^{m}\{Z_v(s_i;t_j) - \hat{Z}_v(s_i;t_j)\}^2$$

Validation サンプルと予測値との差の2乗の平均

（人気の理由）

・二乗和損失誤差の期待値に関する経験的な指標であり，最小化するとS-T kriging predictor（つまり，最良線形予測量）となるため.

・さらに，MSPEは予測量のバイアスに関する部分と分散に関する部分とに分解でき，モデル構築の大部分はそのトレードオフを検討するものであるため有益.

2/5. Root Mean Squared Prediction Error（RMSPE: 二乗平均平方根誤差）これも一般的

- MSPEの平方根をとったもの
- RMSPEのほうがMSPEよりも好まれる場合があるが，これは単位が観察されたものと同じになり分かりやすいため.

3/5. Mean Absolute Prediction Error（MAPE: 平均絶対値予測誤差）

$$MAPE = \frac{1}{Tm}\sum_{j=1}^{T}\sum_{i=1}^{m}\left|Z_v(s_i;t_j) - \hat{Z}_v(s_i;t_j)\right|.$$

- 利点：外れ値の影響を防ぎたいときに有効
- 欠点：バイアスと分散の部分に分けられない.

4/5. Anomaly correlation coefficient（ACC: アノマリー相関係数）時空間プロセスの連続値の検証によく用いられる

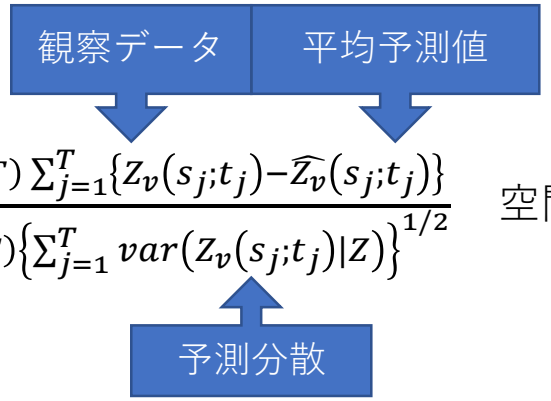
$$ACC = \frac{\sum_{j=1}^{T}\sum_{i=1}^{m}\left(Z_v(s_i;t_j) - Z_a(s_i)\right)\left(\widehat{Z_v}(s_i;t_j) - Z_a(s_i)\right)}{\sqrt{\sum_{j=1}^{T}\sum_{i=1}^{m}\left(Z_v(s_i;t_j) - Z_a(s_i)\right)\sum_{j=1}^{T}\sum_{i=1}^{m}\left(\widehat{Z_v}(s_i;t_j) - Z_a(s_i)\right)}}$$

（検証用の観察データ）と（長期間の観察データの平均値）とのanomaly

（予測値）と（長期間の観察データの平均値）とのanomaly

- 普通のピアソンの相関係数と同じ式.
- 予測のanomalyの分散のパターンが観察のそれと同じなら，ACCは最大値１．完全に反対なら最小値―1となる.
- 欠点：この相関係数に限ったことではないが，予測値の観察データに対するバイアスは考慮していない.
- 利点：フィールド間のフェーズの違い（シフト）を発見できる.

5/5. その他の基準　（サンプルされたデータのみ，つまり同じデータを2回用いて検証する場合に適用できる統計値）



$$V_1(s_i) = \frac{(1/T)\sum_{j=1}^{T}\{Z_v(s_j;t_j) - \widehat{Z_v}(s_j;t_j)\}}{(1/T)\left\{\sum_{j=1}^{T} var(Z_v(s_j;t_j)|Z)\right\}^{1/2}}$$

空間における予測量のバイアスに関しての指標（バイアスなければ 0 に近づく）

$$V_2(s_i) = \left[\frac{(1/T)\sum_{j=1}^{T}\{Z_v(s_j;t_j) - \widehat{Z_v}(s_j;t_j)\}^2}{(1/T)\left\{\sum_{j=1}^{T} var(Z_v(s_j;t_j)|Z)\right\}^{1/2}}\right]^{1/2}$$

MSPEの 正確さの指標．モデルからの予測誤差が妥当であれば，１に近づく．

$$V_3(s_i) = \left[(1/T)\sum_{j=1}^{T}\{Z_v(s_j;t_j) - \widehat{Z_v}(s_j;t_j)\}^2\right]^{1/2}$$

予測のgoodnessの指標．小さければ予測はよい．

・上記の統計値を空間の関数としてプロットするとよい

→空間のある特定の領域の予測が他と比べてよいかなどがわかる

注意：同様の指標が時間に対しても存在する．

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）

6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）

6.3.3 Spatio-Temporal Cross-Validation Measures(時空間モデルの交差検証を行う際のmeasure)

6.3.4 Scoring Rules（スコアリングのルールを用いた検証）

6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）

6.4.1 Model Averaging

6.4.2 Model Comparison via Bayes Factors

6.4.3 Model Comparison via Validation

6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

- **まえおき**

- 今まで紹介したvalidationのsummary measureは主に，with-in sample validation (つまり，同じデータを2回使っている）のときに用いられており，同じデータが2回使われているという点で楽観的なmeasureである（一回目が訓練データとして，2回目がvalidationのときに使用)

- もし可能なら，取り分けておいたvalidation sampleを用いるのが良い．がそのような状況はなかなかないかも．

- そんなときにはcross-validation methodsを用いるのだが，ここでは，時空間データを対象としたいくかのcross-validation statisticsを紹介．

6.3.3 Spatio-Temporal Cross-Validation Measures（時空間交差検証のmeasure）　p272-

1. LOOCV(Leave-one-out-cross-validation)の際の残差を用いたシンプルな指標
2. PCV (Predictive Cross-Validation)
3. SCV (Standardized-Cross-Validation)

1/3. LOOCV(Leave-one-out-cross-validation)の際の残差を用いたシンプルな指標

$$\left\{\frac{Z(s_i;t_j)-E\left(Z(s_i;t_j)|Z^{(-i,-t_j)}\right)}{\left\{var\left(Z(s_i;t_j)|Z^{(-i,-t_j)}\right)\right\}^{1/2}}\right\}$$

$Z(s_i;t_j)$のデータを取り除いたデータ

・外れ値や潜在的な時空間の依存性などをみつけるのによい.

2/3. PCV (Predictive Cross-Validation)

$$PCV \equiv \left(\frac{1}{mT}\right)\sum_{j=1}^{T}\sum_{i=1}^{m}\left\{Z(s_i;t_j)-E\left(Z(s_i;t_j)|Z^{(-i-t_j)}\right)\right\}^2$$

先ほどの$V_3(s_i)=\left[(1/T)\sum_{j=1}^{T}\left\{Z_v(s_j;t_j)-\widehat{Z_v}(s_j;t_j)\right\}^2\right]^{1/2}$ 予測のgoodnessの指標と似てる.

・予測モデルのふるまいがよいなら，PCV= 0 に近づく.

3/3. SCV (Standardized Cross-Validation)

$$SCV \equiv \left(\frac{1}{mT}\right)\sum_{j=1}^{T}\sum_{i=1}^{m}\frac{\{Z(s_i;t_j) - E(Z(s_i;t_j)|Z^{(-i-t_j)}\}^2}{var(Z(s_i;t_j)|Z^{(-i,-t_j)})}$$

先ほどの$V_2(s_i) = \left[\frac{(1/T)\sum_{j=1}^{T}\{Z_v(s_j;t_j)-\widehat{Z_v}(s_j;t_j)\}^2}{(1/T)\{\sum_{j=1}^{T} var(Z_v(s_j;t_j)|Z)\}^{1/2}}\right]^{1/2}$ MSPEの正確さの指標と似てる.

・予測モデルのふるまいがよいなら，SCV=1に近づく.

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）

6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）

6.3.3 Spatio-Temporal Cross-Validation Measures(時空間交差検証のmeasure)

6.3.4 Scoring Rules（スコアリングのルールを用いた検証）

6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）

6.4.1 Model Averaging

6.4.2 Model Comparison via Bayes Factors

6.4.3 Model Comparison via Validation

6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

- **まえおき**

- 4章5章で紹介した統計モデルの利点：確率的予測を与える．→一つの解（予測）ではなく予測分布.
- 予測における様々な不確実性を考慮できる点で優れている．しかし，ちょっと困った点もあり，理想的には，予測分布の妥当性を調べたいが，観察データは1セットしかない.

- 予測分布の妥当性を一つの観察セットと比較して評価する際に**スコア**，**スコア関数**を考える.
- $p(z)$を予測分布，Zをvalidation valueとすると，スコア関数は $S(p(z),Z)$と表記.

- 確率的予測の検証に関しては長い歴史があり，初期は，「proper」なscoring functionが好まれていた.

6.3.4 Scoring Rules（スコアリングのルール）　p273-

A．時間や空間に対して平均化あるいはまとめることができるような，一変量の場合

1. BRS (Brier Score: ブライアスコア)
2. Ranked Probability Score (RPS)
3. Continuous Ranked Probability Score (CRPS)
4. Dawid-Sebastiani Score (DSS)とLogarithmic Score (LS)
5. Skill Score (SS)

B．多変量の予測の場合（一般的ではないが，過程モデルの変数間に時空間的依存性があるときに大事）

1. Energy Score (ES)
2. DSSの多変量バージョン
3. Variogram Score of Order p

A．時間や空間に対して平均化あるいはまとめることができるような，一変量の場合

1/5. BRS (Brier Score) ---カテゴリカル変数による確率的予測分布を比較するときに用いる．（特に0,1の二項データ）

$$BRS(p, Z) = (Z-p)^2$$

ここで，Zは二値データの観察とし，モデルより1が出る確率が $p = \Pr(Z=1|\text{data})$

・ゴルフと同様に，この値が小さいほうがよい．
・簡便のため，時空間の記号を省くが，実際には，いくつもの予測に対して平均的なBRSを計算
・BRSは予測に関する三つのコンポーネント（reliability信用性, resolution解像度, and uncertainty不確実性)に分解できる
・BRS=reliability-resolution+uncertainty
・その他にも使えるスコアは色々ある：Heidke skill score, Peirce skill score, Gilbert skill scoreなど

RのVerificationというパッケージにある

2/5. RPS (Ranked Probability Score)---BRSを複数のカテゴリーへと拡張したもの

$$RPS(p,Z) = \frac{1}{J-1}\sum_{i=1}^{J}\left(\sum_{j=1}^{i} Z_j - \sum_{j=1}^{i} p_j\right)^2$$

Jはカテゴリーの数，p_jはj番目のカテゴリーの予測確率であり，もしそのカテゴリーが起きたのなら，$Z_j = 1$，そうでないなら，$Z_j = 0$．

- J=2のときがBRSに同等.
- 予測がパーフェクトだとRPS=0, 予測が最悪だとRPS=1

RのVerificationというパッケージにある

3/5. CRPS (Continuous Ranked Probability Score)---RPSを連続変数へと拡張したもの

Indicator variable　$Z \leq x$なら１，そうでないら０

$$CRPS(F,Z) = \int \left(1\{Z \leq x\} - F(x)\right)^2 dx$$

- 上式をより計算可能なように書き換える．ここで，累積分布関数 F が有限の1次のモーメントを持つとしたとき，次のように書き換えられる：

$$CRPS(F,Z) = E_F|z - Z| - \frac{1}{2}E_F|z - z'|$$

ここでzとz‘は独立なランダム変数であり，分布関数Fに従う

- 上式のCRPSは標準的な累積予測分布関数においては効率よく計算できるが，BHM (Bayesian Hierarchical Model)などから得られるような複雑な分布においては計算しにくく，その場合は経験的な累積予測分布関数を用いればよい：

$\widehat{F_m}(x) = \frac{1}{m}\sum_{i=1}^{m} 1\{Z_i \leq x\}$　empirical 累積予測分布関数

$$CRPS\left(\widehat{F_m}, Z\right) = \frac{1}{m}\sum_{i=1}^{m}|Z_i - Z| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|Z_i - Z_j|$$

ここで $\{Z_i\}$ はindependent and identically distributed (iid)である必要.

この仮定は，ある場所での複数回の観察が大きく時間をまたいでいるならよいが，時空間モデルのデータセットの場合そうなっていないものも多いので注意が必要である.

RのVerificationというパッケージにある

4/5. Dawid-Sebastiani Score (DSS)とLogarithmic Score (LS)

予測分布の評価を1次と2次の中心モーメントである期待値（μ_F）と分散（σ_F^2）のみを用いて行う

$$DSS(F,Z) = \frac{(Z-\mu_F)^2}{\sigma_F^2} + 2log\sigma_F$$

特に，ガウス予測密度関数$f(z)$におけるDSSはLogarithmic Score (LS)と呼ばれ，次式で表される：

$$LS(F,Z) = -\log f(Z)$$

→機械学習において最も使われている
→LSが低いほうがよい

4/5. Skill Score (SS)

予測モデルの**スキル**（S）：（定義）様々な予測ケースにおけるスコアリングルールの平均

$$S = \frac{1}{N}\sum_{i=1}^{N} S(F_i, Z_i)$$

スキルスコア(SS)を用いて，予測モデルと別のreference prediction methodモデルからの予測を比較

$$SS_M = \frac{S_M - S_{ref}}{S_{opt} - S_{ref}}$$

ここでS_M はモデルMのスキル，S_{ref} はreference methodのスキル，S_{opt}は仮説上の最適なpredictorのスキル

- モデルMからの予測が最適なときに最大値１
- モデルMがreference methodと同じスキルを持つときに0
- モデルMがreference methodよりも低いskillをもつときは0以下

※SS_Mは一般的にproper ruleではない．

RのVerificationというパッケージにある

B．多変量の予測の場合（一般的ではないが，過程モデルの変数間に時空間的依存性があるときに大事）

1/3. Energy Score (ES)

$$ES(F,Z) = E_F\|z - Z\| - \frac{1}{2}E_F\|z - z'\|$$

$\|\cdot\|$はユークリッドノルムであり（ベクトル空間に対して距離を与えるもの）
zと z‘は多変量累積分布関数Fからの独立のランダム変数

・ESがよいパフォーマンスを示すには，モデルの中の変数の依存性の構造を正確に特定する必要がある．

2/3. DSSの多変量バージョン

予測分布の評価を1次と2次の中心モーメントである期待値（μ_F）と分散（σ_F^2）のみを用いて行う

$$DSS_{mv}(F,Z) = log|C_F| + (Z - \mu_F)'C_F^{-1}(Z - \mu_F)$$

ここで$\mu_F = E(Z|data)$は多変量予測累積分布関数Fの平均のベクトル,$C_F = var(Z|data)$ その共分散行列である

3/3. Variogram Score of Order *p*

$$VS_p(F,Z) = \sum_{i=1}^{mT}\sum_{j=1}^{mT} w_{ij}\left(\left|Z_i - Z_j\right|^p - E_F\left|z_i - z_j\right|^p\right)^2$$

ここで，w_{ij} は負でない重みづけであり，z_iz_jは多変量累積分布関数Fからのランダムなベクトル **z** のi番目とj番目の要素であり，そして簡便的にデータのベクトルをZ＝（Z_1,,,,Z_{mT}）’と記す．wにより，特定の差のペア（とても離れているやつとか）の重みづけを小さくしたりすることができる

-> see Lab 6.1

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）

6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）

6.3.3 Spatio-Temporal Cross-Validation Measures(時空間交差検証のmeasure)

6.3.4 Scoring Rules（スコアリングのルールを用いた検証）

6.3.5 Field Comparison　（場の比較による検証）

6.4. Model Selection（モデルの選択）

6.4.1 Model Averaging

6.4.2 Model Comparison via Bayes Factors

6.4.3 Model Comparison via Validation

6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

6.3.5 Field Comparison（場の比較）p278-

まえおき

場の比較の例）
短期的な降水量のnowcastsを行うモデルからの予測と，天気レーダのデータからの二つの場を比較することでモデルを検証したい

- 今まで紹介してきたMSPE, MAPE, RMSPE, ACCなどの指標はよく場の比較の際にも用いられるが，それよりもさらに,**プロセスとデータの間の時空間的な特徴を比較するための特別なsummary**をここでは紹介

- 場の比較において最もチャレンジングなことの一つは，feature location, orientation, scaleの違いをどう考慮するかである.
（特徴のある場）（方向性）　（大きさ）

A．Field-Matching Methods

1. Threat Score (TS)

B．Field Significance　フィールドが有意に異なるか

1. Enhanced False Discovery Rate (EFDR)法

A. Field Matching Methods

1/1. Threat Score(TS) --- Critical Success Indexとも言う.

（あるイベントの予測に成功した回数/そのイベントが予測あるいは観察された回数）の割合

$$\mathrm{TS} = \frac{A_{11}}{A_{11} + A_{10} + A_{01}}$$

- A_{11}は予測したイベントが起こると期待された場所で事実そこで起きたエリア
- A_{10}はイベントが起こると予測されたが起きなかったエリア
- A_{01}はイベントは起きたが，そこで起きるとは予測されていなかったエリア

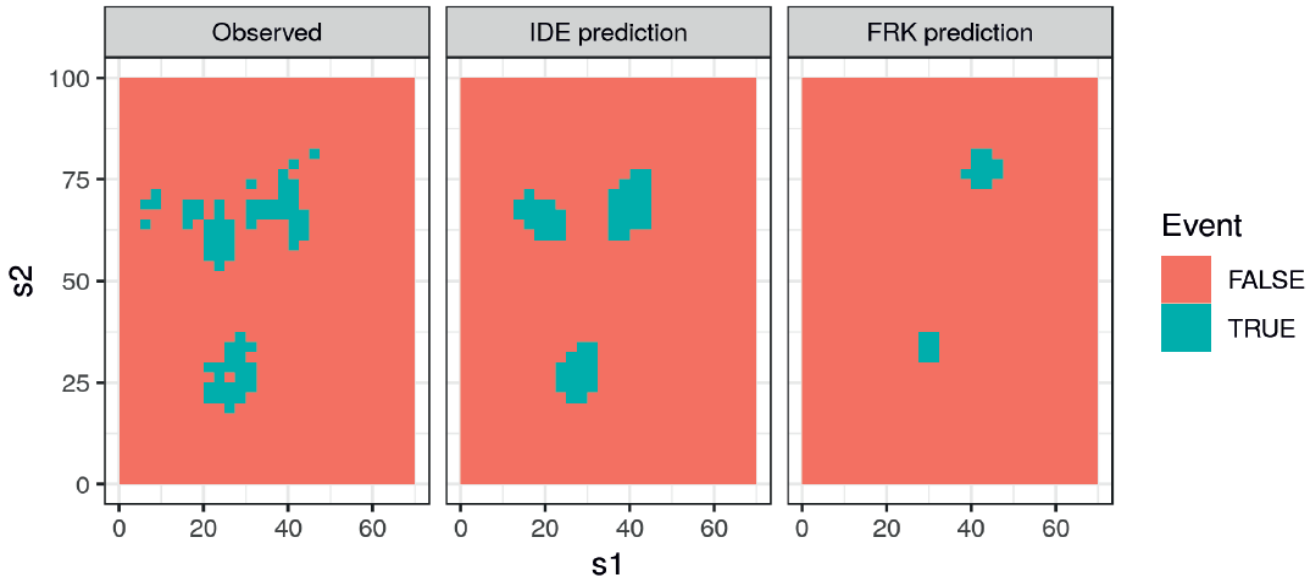
Lab6.1で図示 (Sydney radar data set)

(想定)
- 9:35-9:45の10分間のデータは含まない.
- これらの時間における反射性をIDEモデル，FRKモデルの時空間基底関数を用いて予測する.

(TSの計算)
- イベントが起きたor起きなかったを明らかにする必要がある.
- Threshold parameterを設定する.
- Thresholdよりも大きい観察・予測→起きた， Thresholdよりも小さい観察・予測→起きなかった

（実際にはいくつかのthreshold parameterを設定して結果の頑健性を確認する）

TSを用いたフィールド比較の例



- Thresholdは25dBZ
- IDE predictionのほうがイベントをとらえている

Figure 6.8: Plots showing the presence or absence of events at 09:35, obtained by thresholding the observations (left) or the IDE/FRK predictions (center and right) at 25 dBZ.

Table 6.1: Threat scores (TS) calculated using (6.26) for both the IDE predictions and the FRK predictions at 09:35 for different thresholds.

Threshold (dBZ)	TS for IDE	TS for FRK
15.00	0.73	0.32
20.00	0.58	0.21
25.00	0.37	0.11

- Thresholdを15から25の間で変化
- どのthresholdでもIDEモデルのほうがよい

RのSpatial IVxというパッケージにはTSを含む様々なField matching法があるのでみてみるとよい！

B. Field Significance

（問い例）2001-2010年における北アメリカのグリッドにおける最高気温は，1971-1980年と有意に異なるか？　など

時空間的依存性を考慮する必要がある．

1/1. Enhanced False Discovery Rate (EFDR)法

- 空間フィールドをwavelet scaleに基づいて分解．
- 空間的依存性をDecorrelated wavelet coefficientsに基づいてテスト

（テスト例）1970年代と1990年代における太平洋のSST(海面水温）の異常性の違い

左図）水温の差．赤いほうが1990年代高い　　右図）有意に異なるエリア．EFDR法を用いる

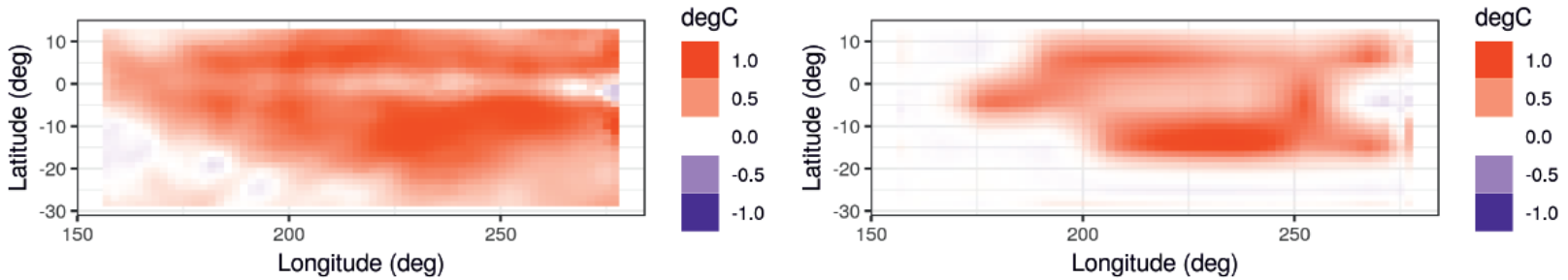


Figure 6.9: Left: Difference between the average SST anomalies in the 1990s and the average SST anomalies in the 1970s. Right: The field significance map of SST anomaly differences that were found to be significantly different from zero at the 5% level. The plot is based on the EFDR procedure and was obtained using the package **EFDR**.

RのEFDRというパッケージを用いた

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）
6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）
6.3.3 Spatio-Temporal Cross-Validation Measures(時空間交差検証のmeasure)
6.3.4 Scoring Rules（スコアリングのルールを用いた検証）
6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）どのモデルが最もよいかを決定する方法の紹介

6.4.1 Model Averaging
6.4.2 Model Comparison via Bayes Factors
6.4.3 Model Comparison via Validation
6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

6.4.1 Model Averaging p282- モデル選択というよりモデルの統合？

- 一つのモデルのみに焦点をあてるのではなく，いくつかのモデルを平均化することで，よりよい予測を得ようとする.

$$[g|\boldsymbol{Z}] = \sum_{l=1}^{L} [g|\boldsymbol{Z}, M_l] P(M_l|\boldsymbol{Z})$$

- Zはモデルを訓練した観察データ
- $[g|\boldsymbol{Z}, M_l]$はデータZ,モデルM_lを与えられたときのgの事後分布
- $P(M_l|\boldsymbol{Z})$は，データZが与えられたときのモデルM_lの事後確率で次式で与える

$$P(M_l|\boldsymbol{Z}) = \frac{[Z|M_l]P(M_l)}{\sum_{j=1}^{L}[Z|M_j]P(M_j)}$$

- モデルの事前分布$P(M_l)$は与える.
- 個々のモデルの$integerated\ likelihood$は$[Z|M_l] = \int\int [Z|Y,\theta,M_l]\,[Y|\theta,M_l][\theta|M_l]dYd\theta$ であり，$[Z|Y,\theta,M_l]$はモデルM_lにおけるdata model(尤度)であり，$[Y|\theta,M_l]$はモデルM_lのもとでのプロセスYの分布，$[\theta|M_l]$は事前分布である

㊟ しかしながら，階層ベイズモデルのときは，この式は扱いにくく複雑なモデルでは使えない.

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）
6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）
6.3.3 Spatio-Temporal Cross-Validation Measures(時空間交差検証のmeasure)
6.3.4 Scoring Rules（スコアリングのルールを用いた検証）
6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）どのモデルが最もよいかを決定する方法の紹介

6.4.1 Model Averaging
6.4.2 Model Comparison via Bayes Factors
6.4.3 Model Comparison via Validation
6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

6.4.2 Model Comparison via Bayes Factors p283- モデル選択というよりモデルの比較

- **Posterior odds （二つの事後確率の比）**

$$\frac{p(M_l|Z)}{p(M_k|Z)} = \frac{[Z|M_l]P(M_l)}{[Z|M_k]P(M_k)} \equiv B_{l,k}(Z)\frac{P(M_l)}{P(M_k)}$$

- ここで，統合尤度$[Z|M_l]$の比$B_{l,k}(Z)$をBayes factor(ベイズ係数)
- $B_{l,k}(Z)$が大きいほど，モデルM_lへのサポートは大きくなる.

- ちなみにベイズ係数の負の対数尤度をとったら，次のように以前のLS(Logarithmic Score)で表される：

$$-logB_{l,k} = LS(F_l; \boldsymbol{Z}) - LS(F_k; \boldsymbol{Z})$$

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）
6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）
6.3.3 Spatio-Temporal Cross-Validation Measures(時空間交差検証のmeasure)
6.3.4 Scoring Rules（スコアリングのルールを用いた検証）
6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）どのモデルが最もよいかを決定する方法の紹介

6.4.1 Model Averaging
6.4.2 Model Comparison via Bayes Factors
6.4.3 Model Comparison via Validation
6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

6.4.3 Model Comparison via Validation

BHM(ベイズ階層モデル）におけるモデル比較

- 対数スコア(LS)について考える
- N個のMCMCサンプルを平均して得られる.

$$LS_{p,l} = -log\left(\frac{1}{N}\sum_{j=1}^{N}[Z_p|Z, Y^{(j)}, \theta^{(j)}, M_l]\right), \qquad l = 1, \ldots., L$$

ここで，Z_pは我々のモデルで予測したい時空間データ，$Y^{(j)}$は j 番目のMCMCサンプルのプロセス，$\theta^{(j)}$はj番目のMCMCサンプルのパラメータの要素，l番目のモデルに基づく.

- 複数のモデル（L個）について$LS_{p,l}$を計算し比較する→小さいほうがよいモデル

Validationで使う取り置きサンプルがない場合はcross-validationを行う

- K-fold cross-validationによる推定値のLSは

$$LS_{cv,l} = -\frac{1}{K}\sum_{k=1}^{K} log\left(\frac{1}{N}\sum_{j=1}^{N}[Z_k|Z^{(-k)}, Y^{(j)}, \theta^{(j)}, M_l]\right), \qquad l = 1, \ldots., L.$$

Z_kはZのk番目のhold-out sampleの中の要素

㊟　しかしながら，時空間モデルの階層ベイズモデルのときは，かなり計算コストが高くなる.
そこで，訓練データを2回用いた場合（1回目はモデルのパラメータ推定，2回目は予測の評価）のバイアスを補正する方法で評価することを考える→Information Criteria

本日の内容

6.3. Model Validation（モデルの検証）

6.3.1 Predictive Model Validation（予測モデルの検証）
6.3.2 Spatio-Temporal Validation Statistics（時空間検証統計値）
6.3.3 Spatio-Temporal Cross-Validation Measures(時空間交差検証のmeasure)
6.3.4 Scoring Rules（スコアリングのルールを用いた検証）
6.3.5 Field Comparison　（フィールドの比較による検証）

6.4. Model Selection（モデルの選択）どのモデルが最もよいかを決定する方法の紹介

6.4.1 Model Averaging
6.4.2 Model Comparison via Bayes Factors
6.4.3 Model Comparison via Validation
6.4.4 Information Criteria

6.5. Chapter 6 Wrap-Up（6章まとめ）

6.4.4 Information Criteria（情報量基準）

1. Akaike Information Criterion (AIC)
2. Bayesian Information Criterion (BIC)
3. Deviance Information Criterion (DIC)
4. Watanabe-Akaike Information Criterion (WAIC)
5. Posterior Predictive Loss (PPL)

まえおき

情報量基準は，バイアスと分散のトレードオフを示し，訓練データと同じデータを用いて評価したときなどに生じる過適合によるバイアスにペナルティーを課す方法である．

Akaike Information Criterion (AIC)

$$AIC(M_l) \equiv -2log\left[Z|\hat{\theta}, M_l\right] + 2p_l$$

- パラメータθは最尤法により推定される．
- $-log\left[Z|\hat{\theta}, M_l\right]$は，モデル$M_l$のLS（ログスコア）であり，推定されたパラメータ$\hat{\theta}$が最尤推定値，$p_l$はモデル$M_l$で推定されたパラメータの数．つまり，LS＋パラメータ数によるペナルティー．
- AICは低いほうがよい．
- モデルM_lにランダム効果や依存性があると，AICは不適（何故なら有効パラメータ数はp_lでなくなるため）．
- 異なるベイズ階層モデル間のモデル選択にも不適（ベイズ階層モデルでは最尤推定値の推定が困難）．

Bayesian Information Criterion (BIC)

$$BIC(M_l) \equiv -2log\left[Z|\hat{\theta}, M_l\right] + \log(m^*)p_l$$

ここで，m^*はサンプルサイズ（つまり，時空間観察データの数），$\hat{\theta}$は最尤推定値，p_lはモデルM_lで推定されたパラメータの数

- AICと同様に低いほうがよい.
- $m^* > 7$ において，AICよりもより大きいペナルティーを課す
- AICと同様に，階層ベイズモデル間での比較には不適.
- ランダム効果などがあるときの有効パラメータ数に関するペナルティーの補正がない.

Deviance Information Criterion (DIC)　逸脱度情報量基準

$$DIC(M_l) \equiv -2log[Z|E(\theta|Z), M_l] + 2p_l^D$$

ここで$E(\theta|Z)$はモデルM_lにおける θ の事後分布の期待値であり，p_l^Dは有効パラメータ数で次式：

$$p_l^D \equiv \overline{D_l} - \widehat{D_l}$$

$\widehat{D_l}$は推定モデルの逸脱度 $-2log[Z|E(\theta|Z), M_l]$であり，$\overline{D_l}$は事後平均逸脱度であり次式:

$$\overline{D_l} = \int -2(log[Z|\theta, M_l])[\theta|Z, M_l]d\theta$$

- DICは階層ベイズモデルのMCMCの適用において比較的単純に計算できる
- （問題点）有効パラメータ数の推定に関連することと，混合モデルには適切でない

→これらのlimitationsを克服しようとする試みがいくつかある.

Watanabe-Akaike Information Criterion (WAIC)

$$WAIC(M_l) = -2\sum_{i=1}^{m*} log\left(\int [Z_i|\hat{\theta}, M_l]\,[\theta|Z, M_l]d\theta\right) + 2p_l^w$$

ここで，有効パラメータ数p_l^wは次式：

$$p_l^w = \sum_{i=1}^{m^*} var_{\theta|Z}\,(log[Z_i|\theta, M_l]$$

階層ベイズモデルにおけるWAICの（DICと比べた）利点：
① Θの事後分布を直接平均化している
② 有効パラメータ数に関してより現実的なペナルティを与えている
③ 階層ベイズモデルとベイズ混合モデルの両方に適している

㊟仮定に依存性のある時空間モデルには不適

Posterior Predictive Loss (PPL)

$$PPL(M_l) = \sum_{i=1}^{m*}(z_i - E(z_i|Z, M_l))^2 + \sum_{i=1}^{m*} var(z_i\,|Z, M_l)$$

ここで$E(z_i|Z, M_l))$は観察z_iの予測平均で$var(z_i|Z, M_l)$は予測分散

- 依存プロセスのあるBHMに対してより適している

AICやBICはRで計算できる．パッケージINLAにあるInlaという関数はdevianceとWAICを計算する．Spatio Temporal, FRK, IDEなどのパッケージはモデルの当てはめによる尤度を計算し，いくつかの情報量基準を計算することができる．

6.5 Wrap Up

- “It is worth pointing out again that many of these methods are often **not appropriate in fully Bayesian contexts, or when one has dependent random effects**. In that sense, there is still a lot of work to be done in developing model-selection approaches for complex spatio-temporal models”

- Murphy (1993) --- List of nine “attributes” to consider when trying to describe the quality of a forecast

1. Bias
2. Accuracy
3. Skill
4. Reliability
5. Resolution
6. Sharpness
7. Discrimination
8. Uncertainty

ここでは主に１と２を扱ったが，他の視点も大事で，時空間統計学のまだ未発展な部分に挑戦していくとよい