Глава 2

Моделирование величины уловов методом усреднения байесовских моделей

Введение

Излагаемая методика является комбинацией известных идей и алгоритмов. Ее фундамент составляют:

— **байесовский подход к оцениванию,** в рамках которого неизвестная модель считается случайной;

— **парадигма Монте-Карло,** в силу которой «тотальное» усреднение байесовской модели подменяется выборочным;

— **алгоритм Метрополиса – Гастингса,** используемый как генератор последовательности моделей, образующих марковскую цепь, вдоль которой и происходит усреднение вычисляемых характеристик.

В целом данный подход известен в англоязычной литературе под названием Markov Chain Monte Carlo Model Composition (MC^3). Применительно к общей задаче моделирования он разрабатывался приблизительно с конца 1980-х и был опубликован в 1990-х годах (4)[1]. С целью иллюстрации работоспособности данная методика применялась для решения различных прикладных задач, в том числе предпринимались попытки адаптировать ее и к оцениванию величины уловов (5; 7).

Далее в разделе 2.1 изложены принципы использования байесовских моделей для прогнозирования сложных систем, характеризующихся высокой степенью неопределенности; в разделах 2.2 и 2.3 рассмотрены детали алгоритмической реализации этого метода и приведены результаты вычислений, иллюстрирующие работу алгоритма на реальных данных.

[1]Здесь приведена лишь одна ссылка. На самом деле по данной тематике существует довольно обширная библиография.

2.1. Общее описание методики

Рассмотрим задачу построения модели для системы с векторным входом и скалярным откликом. Назначение искомой модели — оценка отклика системы в ситуации, когда входное воздействие известно.

В стандартной параметрической постановке задачи вероятностно-статистического моделирования присутствуют три обязательные компоненты:

— **априорные сведения** о моделируемой системе, представленные в формализованном виде некоторым параметрическим семейством моделей; предполагается, что среди моделей этого семейства находится оптимальная модель, доставляющая наилучшие прогнозные оценки;

— **результаты наблюдения** за поведением моделируемой системы;

— **алгоритм идентификации,** обеспечивающий вычисление оценок параметров оптимальной модели.

Качество получаемой модели определяется тем, насколько адекватна каждая из указанных компонент: априорная информация должна быть объективно точна (насколько это возможно), данные наблюдения — репрезентативны, алгоритм оценивания — состоятелен. Кроме того, точность получаемых оценок зависит от полноты информации, содержащейся в априорных сведениях и данных наблюдения. Вообще говоря, более полная исходная информация делает и формулировку задачи, и ее решение более точными и определенными. Специфика рассматриваемой проблематики, однако, такова, что невозможно рассчитывать на полное удаление неопределенности из постановки задачи. Под неопределенностью в данном контексте понимается как неопределенность, связанная с неизвестными параметрами, значения которых необходимы для однозначной идентификации модели (т.е. *параметрическая неопределенность*), так и неопределенность выбора самого семейства моделей (*структурная неопределенность*). Метод «усреднения байесовских моделей» в известной мере учитывает оба вида неопределенности.

Изложение в рамках данного раздела организовано следующим образом. В секции 2.1.1 вводятся первоначальные обозначения. Далее (пункты 2.1.2, 2.1.3) излагаются основные принципы метода усреднения байесовских моделей. В секции 2.1.4 указан известный алгоритм, позволяющий такое усреднение реализовать практически.

2.1.1. Множество моделей

Обозначим Y — отклик системы (результирующую переменную, значение которой требуется «предсказать»). Объясняющие признаки разделим на числовые (непрерывные) и категориальные. Их количество соответственно обозначим m' и m''. Пусть $X^{(i)}$, $i = 1, \ldots, m'$, — числовые, а $\tilde{X}^{(i)}$ — категориальные признаки ($i = 1, 2, \ldots, m''$). Областью допустимых значений числовых признаков является числовая ось или ее подмножество. Каждой категориальной переменной $\tilde{X}^{(i)}$ соответствует некоторое количество l_i индексированных значений (уровней): $a_1, \ldots, a_{l_i}$. Дополнительно введем

вспомогательные бинарные переменные $X^{(ij)}$, $j = 1, \ldots, l_i$:

$$X^{(ij)} = \begin{cases} 1, & X^{(ij)} = a_j^{(i)}, \\ 0, & \text{иначе.} \end{cases}$$

Вспомогательные переменные вводятся в модель вместо исходных категориальных признаков. Их число определяется числом уровней l_i каждой категориальной переменной и их общим количеством m''. Всего для построения модели получаем m переменных всех типов:

$$m = m' + \sum_{i=1}^{m''} l_i.$$

Наиболее простую (но, при удачном выборе регрессионных переменных, достаточно эффективную) модель получим из предположения линейной связанности входных и выходных переменных. Конкретный вид линейной модели определяется набором включаемых в нее признаков. Для полного комплекта переменных отклик определяется уравнением

$$Y = \alpha + \sum_i \alpha_i X^{(i)} + \sum_i \sum_j \alpha_{ij} X^{(ij)} + \sigma\varepsilon, \tag{2.1}$$

где α, α_i и α_{ij} — параметры модели, ε — центрированная случайная величина, имеющая стандартное нормальное распределение, σ — положительный параметр.

Переменные Y, X, ε в (2.1) можно интерпретировать и как случайные величины, и как их значения: после получения эмпирической информации обозначения Y, $X^{(i)}$ и $X^{(ij)}$ можно заменить соответствующими выборочными данными. Вектор ε содержит ненаблюдаемые непосредственно величины. Далее предполагается, что Y, $X^{(i)}$, $X^{(ij)}$, ε — векторы и матрицы, содержащие n строк (n — объем выборки).

В целях упрощения обозначений введем матрицу данных $Z = \|z_k^{(i)}\|$, содержащую все наблюденные значения всех признаков (как числовых, так и категориальных) и вектор β, составленный из параметров α_{ij} и α_i. Таким образом каждой переменной (отвечающей исходному или вспомогательному признаку) соответствует столбец $Z^{(i)}$ матрицы Z и компонента $\beta^{(i)}$ вектора параметров β. Тогда

$$Y = \alpha + \sum_i \beta^{(i)} Z^{(i)} + \sigma\varepsilon = \alpha + Z\beta + \sigma\varepsilon. \tag{2.2}$$

Для удобства интерпретации модели предварительно центрируем входные признаки таким образом, чтобы

$$\sum_{k=1}^{n} z_k^{(i)} = 0, \quad \forall i. \tag{2.3}$$

Тогда α имеет смысл среднего значения результирующей переменной Y, в то время как $\beta^{(i)}$ определяют величину отклонения отклика модели от среднего уровня при изменении значения i-той переменной на единицу.

Далее понадобится также обозначение для подматриц матрицы Z специального вида. А именно, пусть $\mathcal{I}$ — множество индексов, тогда $Z^{(\mathcal{I})}$ будет означать матрицу, составленную из столбцов матрицы Z, номера которых содержатся в $\mathcal{I}$.

Полная модель (содержащая все признаки) неидентифицируема по параметрам: при включении вспомогательных переменных, соответствующих всем уровням категориальных переменных, мы получим вырожденную[2] матрицу Z. Поэтому интерес представляют лишь неполные модели, которым соответствуют невырожденные подматрицы $Z^{(\mathcal{I})}$.

Будем далее рассматривать множество моделей $\mathcal{M} = \{M_j\}$, содержащее только те модели M_j, которые не являются полными ни по одной категориальной переменной. Каждая модель M_j определяется своим набором объясняющих признаков (множеством индексов $\mathcal{I}_j$) и, соответственно, своей матрицей данных $Z^{(\mathcal{I}_j)}$.

Обозначим для краткости $Z_j \equiv Z^{(\mathcal{I}_j)}$ матрицу данных модели M_j, β_j — вектор параметров в этой же модели, тогда

$$Y = \alpha + \sum_{i \in \mathcal{I}_j} \beta^{(i)} Z^{(i)} + \sigma_j \varepsilon = \alpha + Z_j \beta_j + \sigma_j \varepsilon, \tag{2.4}$$

где дисперсия величины ε по-прежнему равна единице; σ_j — положительные скалярные параметры.

Условимся, что «нулевой признак» (константный столбец из единиц) входит в модель всегда, т.е. минимальной (максимально простой) является модель вида:

$$Y = \alpha + \sigma_0 \varepsilon. \tag{2.5}$$

Традиционный метод построения регрессионной модели (2.4) сводится к направленному перебору моделей (с последовательным наращиванием или, напротив, уменьшением их сложности) и вычислению оценок параметров β_j методом наименьших квадратов (МНК). Лучшая модель выбирается среди рассмотренных с помощью того или иного статистического критерия (обычно — с помощью F-критерия). Классические МНК-оценки детально проработаны и относительно просто вычисляются, однако они не дают возможности использовать априорное знание и прошлый опыт также гибко и универсально, как методика байесовского оценивания.

2.1.2. Байесовское моделирование

Характерной особенностью байесовского подхода является то, что искомая модель рассматривается как случайный элемент некоторого априорного множества моделей. Далее будут рассматриваться линейные по параметрам модели вида (2.4). Для них байесовость означает, что входящие в модель объясняющие переменные (выбираемые из некоторого предопределенного набора) и их количество являются случайными величинами.

Согласно классической формуле Байеса[3]

$$p(M_j|Y) = \frac{p(Y|M_j)p(M_j)}{\sum_i p(Y|M_i)p(M_i)}. \tag{2.6}$$

[2] В том смысле, что $\operatorname{rank}(Z) < \min(n, m)$.

[3] Здесь, и далее в тексте, обозначение $p(\cdot)$ следует понимать как вероятность (если аргументом этой функции является событие, связанное со значением случайной величины дискретного типа) или как плотность вероятности (для распределений непрерывного типа).

Условная вероятность $p(M_j|Y)$ количественно оценивает насколько адекватна модель M_j в условиях, характеризуемых выборкой Y. Возможность решения задачи прогнозирования зависит от практической осуществимости вычислительных процедур, доставляющих оценки этих вероятностей. Зная их значения, можно обоснованно выбирать модель из заданного семейства моделей.

Рассмотрим функции, входящие в правую часть формулы (2.6). Вероятности $p(M_j)$ в рамках байесовского моделирования должны быть известны. Они являются частью априорной информации. Функции $p_j(Y) \equiv p(Y|M_j)$ зависят от вида и типа модели M_j. Индекс j в обозначении вероятности или плотности распределения вероятности здесь и далее указывает на то, что эти характеристики являются условными, т.е. вычисляются при фиксированной структуре модели M_j (при фиксированном составе признаков).

Вообще говоря, в рамках байесовского подхода все параметры трактуются как случайные величины. При этом, как указывалось выше, выгодно оставаться в границах параметрической постановки задачи. Поэтому будем исходить из того предположения, что распределения регрессионных коэффициентов могут быть заданы параметрически. Учитывая гибкость параметрических семейств вероятностных распределений, можно утверждать, что для практических задач данное предположение совершенно не ограничивает общности рассмотрения. Пусть θ — вектор всех параметров, определяющих распределения вида $p_j(\alpha, \beta|\theta)$ для всех моделей M_j. Как именно следует задавать эти распределения, определяется, строго говоря, вне процедуры идентификации модели; в методологии байесовского оценивания это тоже, по существу, априорная информация. Кроме того, в этот же вектор добавим все σ_j.

«Вторичный» параметр θ также должен быть задан априорно с помощью вероятностных распределений[4]. Распределения вероятности $p_j(Y|\theta)$ выводятся из вида модели (в данном случае, определяемого уравнением (2.4)) и распределения $p_j(\alpha, \beta|\theta)$. При фиксированном составе признаков зависимость от коэффициентов α и β может быть устранена после соответствующего интегрирования, но появляется зависимость от параметров их вероятностных распределений:

$$p_j(Y|\theta) = \int p_j(Y|\alpha, \beta, \theta)\, p_j(\alpha, \beta|\theta) d\alpha\, d\beta. \tag{2.7}$$

В итоге функцию $p_j(Y) = p(Y|M_j)$ можно теперь представить в виде:

$$\begin{aligned} p_j(Y) &= \int p_j(Y|\theta)\, p_j(\theta) d\theta \\ &= \int p_j(Y|\alpha, \beta, \theta)\, p_j(\alpha, \beta|\theta)\, p_j(\theta)\, d\alpha\, d\beta\, d\theta. \end{aligned} \tag{2.8}$$

Распределения $p_j(\theta) \equiv p(\theta|M_j)$ также, как и априорные вероятности $p(M_j)$, выбираются, исходя из имеющийся информации об исследуемой системе (или о ее ближайших аналогах).

[4]Вырожденные распределения вполне допустимы, т.е. ничто не препятствует жесткой фиксации параметров θ, если имеющаяся информация дает такие возможности

2.1.3. Усреднение байесовских моделей

Если Δ — представляющая интерес характеристика (например, вектор оцениваемых параметров), то апостериорное распределение вероятности для нее можно получить в результате усреднения при фиксированном Y по множеству моделей:

$$p(\Delta|Y) = \sum_{M \in \mathcal{M}} p(\Delta|M, Y) p(M|Y). \tag{2.9}$$

Общее количество моделей равно

$$|\mathcal{M}| = 2^{m'} \prod_{i=1}^{m''} (2^{l_i} - 1)$$

(числу способов, которым можно выбрать подмножество из полного множества объясняющих переменных, соблюдая условие невырожденности матрицы данных). В реальной задаче число объясняющих переменных составляет не менее нескольких десятков. В этих условиях перебор всех моделей требует длительного времени и суммирование (2.9) может оказаться очень трудоемким либо неосуществимым практически. В то же время, далеко не все слагаемые вносят значимый вклад в апостериорное распределение (2.9), поэтому для получения приемлемой точности оценки $p(\Delta|Y)$ полный перебор всех моделей и необязателен.

Стандартным решением данной дилеммы является выборочное усреднение, применяемое в методах Монте-Карло. Если нам удастся получить выборку $M[t]$ объема T из множества моделей $\mathcal{M}$, подчиняющуюся распределению вероятности $p(M|Y)$, то оценка для апостериорного распределения (2.9) вычисляется как сумма

$$p(\Delta|Y) \approx \frac{1}{T} \sum_{t=1}^{T} p(\Delta|M[t], Y). \tag{2.10}$$

Данное равенство обычно тем точнее, чем больше объем выборки. При этом чем сильнее отличие распределения $p(M|Y)$ от равномерного, тем, как правило, меньшим может быть выбрано значение T.

В тех случаях (к числу которых относится и рассматриваемый), когда удается получить весовые коэффициенты w_k пропорциональные вероятностям $p(M_k|Y)$, целесообразно использовать в качестве искомой оценки сумму

$$p(\Delta|Y) \approx \frac{1}{\sum_k w_k} \sum_{k=1}^{N} w_k p(\Delta|M_k, Y). \tag{2.11}$$

Суммирование в (2.11) производится по различным моделям из множества $\mathcal{M}$. Понятно, что при выборочном суммировании целесообразно в первую очередь учитывать модели с наибольшими апостериорными вероятностями.

2.1.4. Алгоритм Метрополиса – Гастингса

Для вычисления среднего значения по некоторому множеству, строго говоря, необходимо исчерпывающее перечисление его элементов. В условиях, когда это затруднительно, прибегают к выборочному усреднению. Центральной проблемой здесь является генерация элементов выборки в соответствии с требуемым распределением вероятности. В рассматриваемой задаче речь идет о множестве моделей $\mathcal{M}$ и распределении $\mathfrak{p}(M|Y)$.

Алгоритм Метрополиса – Гастингса используется для реализации случайного блуждания на множестве $\mathcal{M}$. Усреднение (2.11) производится вдоль последовательности, стартующей из произвольной исходной модели. Все модели в последовательности связаны в марковскую цепь. Множеством состояний цепи является $\mathcal{M}$, а переходные вероятности $\pi(M, M')$ определяются следующим образом.

Назовем две модели «соседними», если по количеству объясняющих переменных они отличаются ровно на одну переменную, причем все переменные, входящие в «короткую» модель, входят также и в «длинную». Определим вероятность перехода для произвольных соседних моделей M и M':

$$\pi(M, M') = \min\left\{1, \frac{\mathfrak{p}(M'|Y)}{\mathfrak{p}(M|Y)}\right\}. \tag{2.12}$$

Остальные переходные вероятности равны нулю.

Отношение $\mathfrak{p}(M'|Y)/\mathfrak{p}(M|Y)$ вычисляется в соответствии с формулой Байеса (2.6):

$$\frac{\mathfrak{p}(M'|Y)}{\mathfrak{p}(M|Y)} = \frac{\mathfrak{p}(Y|M')\mathfrak{p}(M')}{\mathfrak{p}(Y|M)\mathfrak{p}(M)}, \tag{2.13}$$

Последнее выражение в случае равных априорных вероятностей $\mathfrak{p}(M)$ превращается в отношение правдоподобия.

Генерация последовательности моделей сводится, таким образом, к двухшаговому алгоритму: сначала выбирается одна из «соседних» моделей, после чего принимается окончательное решение о «переходе» в новое «состояние». При этом модель-кандидат выбирается среди всех своих соседей равновероятно, а переход происходит с условной вероятностью (2.12). Известно, что стационарное распределение данной марковской последовательности действительно совпадает с условным распределением $\mathfrak{p}(M|Y)$ (см. обзор (3)).

2.2. Алгоритм прогнозирования величины уловов

2.2.1. Особенности задачи прогнозирования уловов

Будем решать задачу моделирования улова отдельно для каждого из объектов промысла и для одного промыслового дня. Термин «промысловый день» применяется здесь в расширительном смысле. Он определяет расчетную ситуацию в полном объеме и включает в себя характеристику всех контролируемых условий и обстоятельств промысла. Предположим, что всего регистрируется $\mathfrak{m}$ таких характеристик

или признаков, которые будем называть объясняющими переменными. Регистрируемые признаки приведены к необходимому формату, в частности, для всех категориальных признаков введены необходимые вспомогательные переменные. Вектор-строка z_k — размерности m содержит описание (т.е. значения всех объясняющих переменных) k-того промыслового дня ($k = 1, \ldots, n$); Z — матрица, построчно составленная из векторов-описаний всех промысловых дней; Y — логарифмы уловов, зафиксированных в течение n промысловых дней (вектор-столбец); $\mathcal{M} = \{M_j\}$ — множество линейных моделей, каждая из которых характеризуется своим набором объясняющих переменных. Целью моделирования является плотность вероятности $p(y|Y)$ случайной величины y — логарифма улова в некоторый промысловый день z.

При моделировании результатов промысла имеет смысл раздельно учитывать нулевые уловы (в частности, дни, в которые промысел данного вида не производился) и уловы не равные нулю. При таком подходе объединенная модель формулируется как смесь вырожденного распределения (плотность распределения которого можно описать с помощью дельта-функции Дирака, центрированной в нуле) и распределения непрерывного типа для ненулевых уловов. Далее в предположении, что исходные данные предварительно просмотрены и судовые дни с нулевыми уловами отфильтрованы, рассматривается модель только для положительных уловов.

Выбор вида априорных распределений (как наблюдаемых величин, так и параметров модели) на данный момент методически не отработан. Применительно к рассматриваемой задаче детально продуманный набор распределений, констант и параметров рассматривается в публикации (5).

Модель строится на основании гипотезы, согласно которой некоторая линейная (по параметрам) функция, определяющая объем вылова, является случайной величиной с логнормальным распределением. Коэффициенты линейной функции (параметры модели) также являются случайными и подчиняются нормальному закону. Вычисления показывают, что в этих предположениях апостериорная плотность вероятности для логарифма величины улова, имеет распределение Стьюдента (в его обобщенной формулировке, в которую входят три параметра: число степеней свободы, параметры положения и точности), что, в принципе, выглядит вполне правдоподобно.

В более точных терминах эти посылки реализуются следующим образом. В качестве «отклика» Y_k рассматривается логарифм уловов. Распределение этой величины предполагается нормальным:

$$p_j(y|\alpha, \beta, \sigma) = f_N^1(y|\hat{y}_j(\alpha, \beta), \sigma^2), \quad k = 1, \ldots, n. \tag{2.14}$$

Здесь $\hat{y}_j(\alpha, \beta)$ — оценка y, полученная в силу модели (2.4), $f_N^d(\cdot|\mu, A)$ — плотность вероятности d-мерного нормального распределения со средним μ и ковариационной матрицей A (в данном случае $d = 1$ и A имеет смысл дисперсии).

Априорные распределения для параметров β также предполагаются нормальными:

$$p_j(\beta|\sigma) = f_N^{d_j}(\beta|0, \sigma^2(g_0 Z_j' Z_j)^{-1}), \tag{2.15}$$

где d_j — размерность модели M_j (число компонент вектора $\beta \equiv \beta_j$), g_0 — неслучайный параметр (проблема выбора значения для g_0 подробно исследована в (6)).

Распределения для среднего логарифма улова α и параметра σ выбираются в соответствии с установившейся традицией как «неинформативные»[5] априорные распределения (согласно рекомендациям Джеффриса (8)):

$$p(\alpha) \propto 1, \quad p(\sigma) \propto 1/\sigma. \tag{2.16}$$

Пусть z — вектор-описание промыслового дня, для которого необходимо получить прогноз улова (точнее, его логарифма), сам прогноз обозначим y, тогда в рассматриваемых предположениях

$$p(y|Y) = \sum_{M_j} f_S(y|n-1, \mu_j, a_j) p(M_j|Y). \tag{2.17}$$

Здесь $f_S(y|\nu, \mu, a)$ — t-распределение с числом степеней свободы ν, параметром сдвига μ и точностью a:

$$f_S(y|\nu, \mu, a) = \frac{\Gamma(\frac{1+\nu}{2})}{\Gamma(\frac{\nu}{2})} \sqrt{\frac{a}{\nu\pi}} \left[1 + \frac{a}{\nu}(x-\mu)^2\right]^{-\frac{1+\nu}{2}}, \tag{2.18}$$

кроме того, использованы следующие обозначения для параметров,

$$\mu_j = \bar{y} + \frac{1}{g_0+1} z B_j Z_j' Y,$$

$$a_j = \frac{n-1}{G_j}\left(\frac{n+1}{n} + \frac{1}{g_0+1} z' B_j z\right),$$

$$\begin{aligned} G_j &= \frac{1}{1+g_0} Y'(I_n - \tilde{Z}_j(\tilde{Z}_j'\tilde{Z}_j)^{-1}\tilde{Z}_j')Y + \frac{g_0}{1+g_0}|Y-\bar{Y}|^2 \\ &= |Y|^2 - n\bar{y}^2 - \frac{1}{g_0+1} Y' Z_j B_j Z_j' Y, \end{aligned}$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} Y_k, \quad B_j = (Z_j Z_j)^{-1},$$

$\bar{Y}$ — n-мерный вектор все компоненты которого равны $\bar{y}$, $\tilde{Z}_j$ — матрица Z_j, дополненная слева столбцом из единиц.

Результатом моделирования является распределение вероятностей оцениваемой величины, в отличие от традиционных классических методов оценивания, предназначенных для вычисления точечных (реже — интервальных) оценок. Если для целей дальнейшего анализа необходима именно точечная оценка, ее можно получить из распределения вероятностей, вычислив среднее, медиану, моду и т.п. Интервальные оценки можно рассчитать, установив соответствующие квантильные точки.

[5]Или иначе – «максимально неопределенные», что означает отсутствие информации о соответствующем параметре. Оба рекомендуемых распределения (2.16) являются несобственными, что не препятствует их использованию в процедурах байесовского статистического вывода.

2.2.2. Формальное описание алгоритма

Ниже излагается ранее опубликованный алгоритм, восстановленный по журнальному описанию (5). Результаты моделирования, приведенные в следующих разделах, получены с его помощью. Приводимое описание содержит более точную и детальную формулировку принципиальных вычислительных особенностей, отсутствующих в оригинальной публикации.

2.2.2.0.1. Входные данные

Y, Z — данные наблюдения в формате описанном в разделе 2.1,

z — вектор признаков промыслового дня, для которого требуется построить прогноз.

2.2.2.0.2. Результаты

$p(y|Y)$ — прогноз величины улова (плотность вероятности).

2.2.2.0.3. Обозначения

n — количество наблюдений (число строк в матрице Z),

m — общее число всех переменных, исключая фиктивный «константный» признак (число столбцов в матрице Z),

$g_0 = 1/\max(n, m^2)$,

k — общее число непрерывных переменных,

l — общее число категориальных переменных,

l_i — количество уровней для i-той категориальной переменной,

k_{cur}, l_{cur}, f_{cur}, m_{cur} — число непрерывных, категориальных, полных категориальных и общее число переменных, включенных в текущую модель,

k_{can}, l_{can}, f_{can}, m_{can} — то же для модели-кандидата,

Z_{cur}, Z_{can} — матрицы данных для текущей модели и модели-кандидата,

2.2.2.0.4. Алгоритм

1. Выбор исходной (стартовой) модели. С помощью датчика случайных чисел сформулировать стартовую модель, включая в неё случайным образом (с вероятностью $1/2$) переменные из полного набора, представленного матрицей Z. При этом необходимо следить за соблюдением условия: ни одна категориальная переменная $\tilde{X}^{(i)}$ не должна быть «полной», т.е. не может быть представлена в модели всеми вспомогательными переменными $X^{(ij)}$, $j = 1, \ldots, l_i$.

   Определить характеристики текущей модели (значения переменных l_{cur}, f_{cur}, m_{cur} и пр.)

2. Начало итераций

   (a) Выбор модели-кандидата

      i. Выбрать параметрическую размерность модели-кандидата. Число параметров увеличивается или уменьшается на единицу:

$$m_{can} = \begin{cases} m_{cur} + 1, & \text{с вероятностью } (m - m_{cur})/m, \\ m_{cur} - 1, & \text{с вероятностью } m_{cur}/m. \end{cases}$$

      ii. Если число переменных увеличено, выбрать «наудачу»новую переменную из числа свободных (контролируя свойство невырожденности матрицы данных). Каждая из свободных переменных может быть выбрана с вероятностью $1/(m - m_{cur} + l - f_{cur})$. Вычислить

$$T = \frac{m - m_{cur} + l - f_{cur}}{m - m_{cur}}.$$

      iii. Если число переменных уменьшено, удалить из модели одну переменную. Каждая из переменных может быть удалена с вероятностью $1/m_{cur}$. Вычислить

$$T = \frac{m - m_{can}}{m - m_{can} + l - f_{can}}.$$

   (b) Принятие/отклонение модели-кандидата

      i. Вычислить

$$Q = \gamma \cdot \left(\frac{G_{cur}}{G_{can}} \right)^{\frac{n-1}{2}},$$

      где

$$\gamma = \begin{cases} \sqrt{(g_0 + 1)/g_0}, & \text{если } m_{can} < m_{cur}, \\ \sqrt{g_0/(g_0 + 1)}, & \text{если } m_{can} > m_{cur}, \\ 1, & \text{иначе}, \end{cases}$$

$$G_{cur} = |Y|^2 - n\bar{Y}^2 - \frac{1}{g_0 + 1} Y' Z_{cur} B_{cur} Z'_{cur} Y,$$

$$B_{cur} = (Z'_{cur} Z_{cur})^{-1},$$

      G_{can} определяется аналогично G_{cur} (с заменой индексирующего маркера cur на can).

      ii. Вычислить

$$L = \begin{cases} l_j, & \text{если } l_{can} < l_{cur}, \\ 1/l_j, & \text{если } l_{can} > l_{cur}, \\ 1, & \text{иначе}, \end{cases}$$

      где j — номер добавленной/удалённной категориальной переменной.

      iii. Модель-кандидат замещает текущую модель с вероятностью $\min(1, T \cdot Q \cdot L)$. В случае, если это произошло, необходимо обновить текущие значения вычисляемых характеристик (формулы (2.17, 2.18)).

3. Вывод/визуализация результатов (формирование таблиц и построение графиков)

2.2.2.0.5. Примечания

1. При переходе к следующей модели необходимо отличать модели, появляющиеся впервые, от моделей посещаемых повторно. Реально обновление расчетных характеристики происходит лишь для «моделей-новичков».

2. Число итераций на втором шаге является параметром алгоритма. Оно определяется по результатам пробных расчетов. Так как усреднение необходимо проводить по стационарному распределению, целесообразно некоторое количество первых итераций не использовать для расчетов.

3. На каждой итерации алгоритма происходит обращение матрицы (вычисление матрицы $\mathsf{B}_{\mathsf{can}}$). Эту операцию имеет смысл реализовать в рекуррентном виде. Учитывая, что на каждой итерации в обращаемую матрицу либо добавляется, либо удаляется ровно один столбец и ровно одна строка, для пересчета матрицы $\mathsf{B}_{\mathsf{can}}$ можно рекомендовать рекуррентную формулу метода окаймления.

4. Данный алгоритм позволяет получить оценку плотности распределения вероятности для величины улова. Поскольку всякое распределение вероятности содержит полную информацию о характеризуемой случайной величине, не составляет большого труда рассчитать прочие характеристики, непосредственно связанные с величиной улова (средний улов, медиану или другие квантильные точки и пр.).

2.3. Результаты моделирования улова по данным промысла в Северной Атлантике

2.3.1. Данные

Для проверки работоспособности алгоритма использовались результаты инспекционных наблюдений за рыбным промыслом в Северной Атлантике (район Большой Ньюфаундлендской Банки). Авторы публикации (5) предоставили свободный доступ к этой информации.

Данные об уловах получены в результате регулярных рейсов инспекционного судна Евросоюза и представлена в текстовом файле в формате таблицы. Строки таблицы соответствуют отдельным промысловым дням, столбцы — значениям переменных (признакам). Всего в этой таблице 6806 строк, каждая строка содержит описание конкретного судового дня, включая значения уловов для пяти основных объектов промысла (Atlantic cod, Greenland halibut, Redfish, Roundnose grenadier, Skate), а также суммарные уловы по всем остальным видам. Итого — 40836 повидовых однодневных уловов. Под уловом понимается суммарный вылов, добываемый за один день (в том числе и получаемый как прилов).

Среди объясняющих переменных — национальная принадлежность судна(представлены данные о судах Испании и Португалии), время лова (год, месяц, день), место лова

(4 зоны в Северной Атлантике у восточного побережья Канады), способ лова (дрифтерная жаберная сеть, якорная жаберная сеть, оттер-трал, парный оттер-трал), тип и характеристики судна и пр.

В расчетах, результаты которых представлены в данном сообщении, в качестве входных переменных взяты

- страна, которой принадлежит судно;
- зона лова (далее условно пронумерованы от 1 до 4);
- месяц, для которого проводится прогнозирование;
- способ лова;
- размер ячейки сети;
- длина судна;
- брутто-регистровый тоннаж;
- мощность двигателя.

Последние четыре признака предварительно подвергались логарифмическому преобразованию.

На основании этой информации был сформирован вектор объясняющих переменных размерности $\mathfrak{m} = 27$.

В качестве моделируемой переменой рассматривался улов по какому-либо одному промысловому виду. Ниже приводятся результаты только для трески и палтуса (Atlantic cod, Greenland halibut).

2.3.2. Прогнозирование результатов промыслового дня

В этом разделе представлены результаты моделирования («прогнозирования») итогов одного промыслового дня: оценки плотности вероятности для величины вылова определенного вида рыбы одним рыболовецким судном за один день.

Результаты, показанные на рис. 2.1, 2.2, выбраны случайно из числа представленных во входном файле. Все прогнозы основаны на результатах счета, в которых прогнозируемые дни исключались из входных данных. На графиках по вертикальным осям отложена плотность распределения вероятности, по горизонтальным — однодневный улов в килограммах. Для удобства сопоставления на графики вынесены также прогнозные оценки (пунктир) и реальные уловы (сплошная линия). В качестве точечного прогноза всюду далее рассматривается медиана.

Характеристики судовых дней, для которых осуществлялось прогнозирование, представлены в табл. 2.1 и 2.2.

Таблица 2.1. Характеристики некоторых промыслов трески

Код	1149	1162	1725	1827	4311	5519
Страна	Порт.	Порт.	Исп.	Исп.	Исп.	Исп.
Зона лова	2	2	4	4	4	2
Месяц лова	март	май	октябрь	июнь	июнь	февраль
Способ лова	ж.с.	ж.с.	п.т.	п.т.	п.т.	п.т.
Сеть (мм)	140	140	130	120	120	130
Длина судна (м)	60,3	60,3	42,0	46,0	43,1	46,0
Тоннаж (т)	896,7	896,7	434,5	614,0	494,5	664,9
Двигатель (кв·ч)	883	883	882	918	845	1877
Улов (кг)	625	1100	1400	10000	180	1800
Прогноз (кг)	420	944	1629	2221	1322	4752

Код — условный номер промыслового дня;

Порт. — Португалия, Исп. — Испания;

ж.с. — жаберная сеть; п.т. — парный («близнецовый») трал;

Таблица 2.2. Характеристики некоторых промыслов палтуса

Код	2279	3846	3978	5545	6485	6728
Страна	Исп.	Исп.	Исп.	Исп.	Исп.	Исп.
Зона лова	4	4	4	1	4	1
Месяц лова	июль	июнь	октябрь	март	март	февраль
Способ лова	п.т.	п.т.	п.т.	п.т.	п.т.	п.т.
Сеть (мм)	130	120	140	130	130	130
Длина судна (м)	33,0	29,0	52,8	46,0	42,0	64,8
Тоннаж (т)	299,1	264,4	491,4	664,9	376,9	996,0
Двигатель (кв·ч)	1037	698	1212	1877	588	1470
Улов (кг)	4237	328	4263	5100	2600	7268
Прогноз (кг)	3656	3280	3325	2619	3763	7480

Код — условный номер промыслового дня;

Исп. — Испания; п.т. — парный («близнецовый») трал;

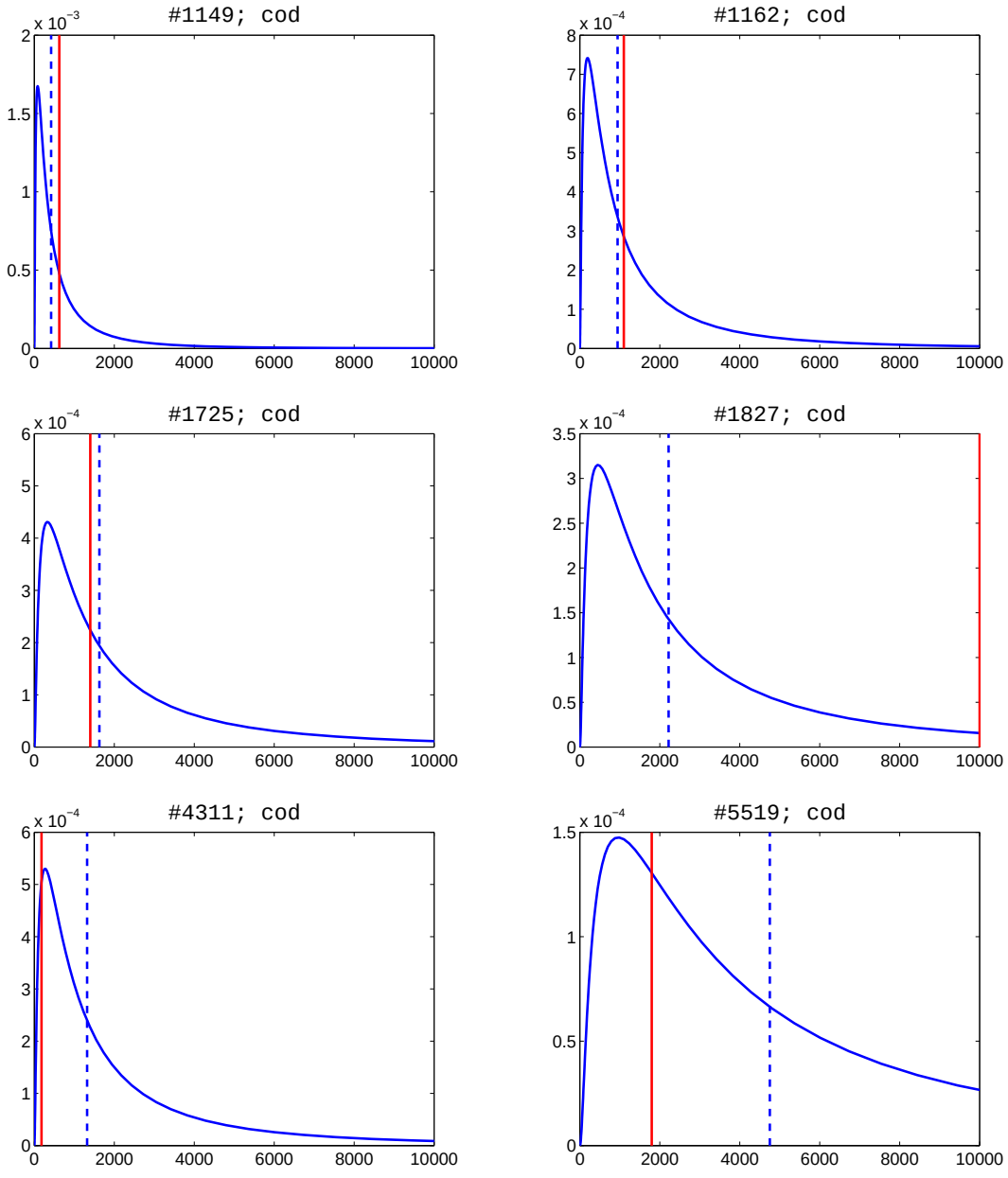

Рис. 2.1. Результаты прогнозирования величины улова трески за один промысловый день

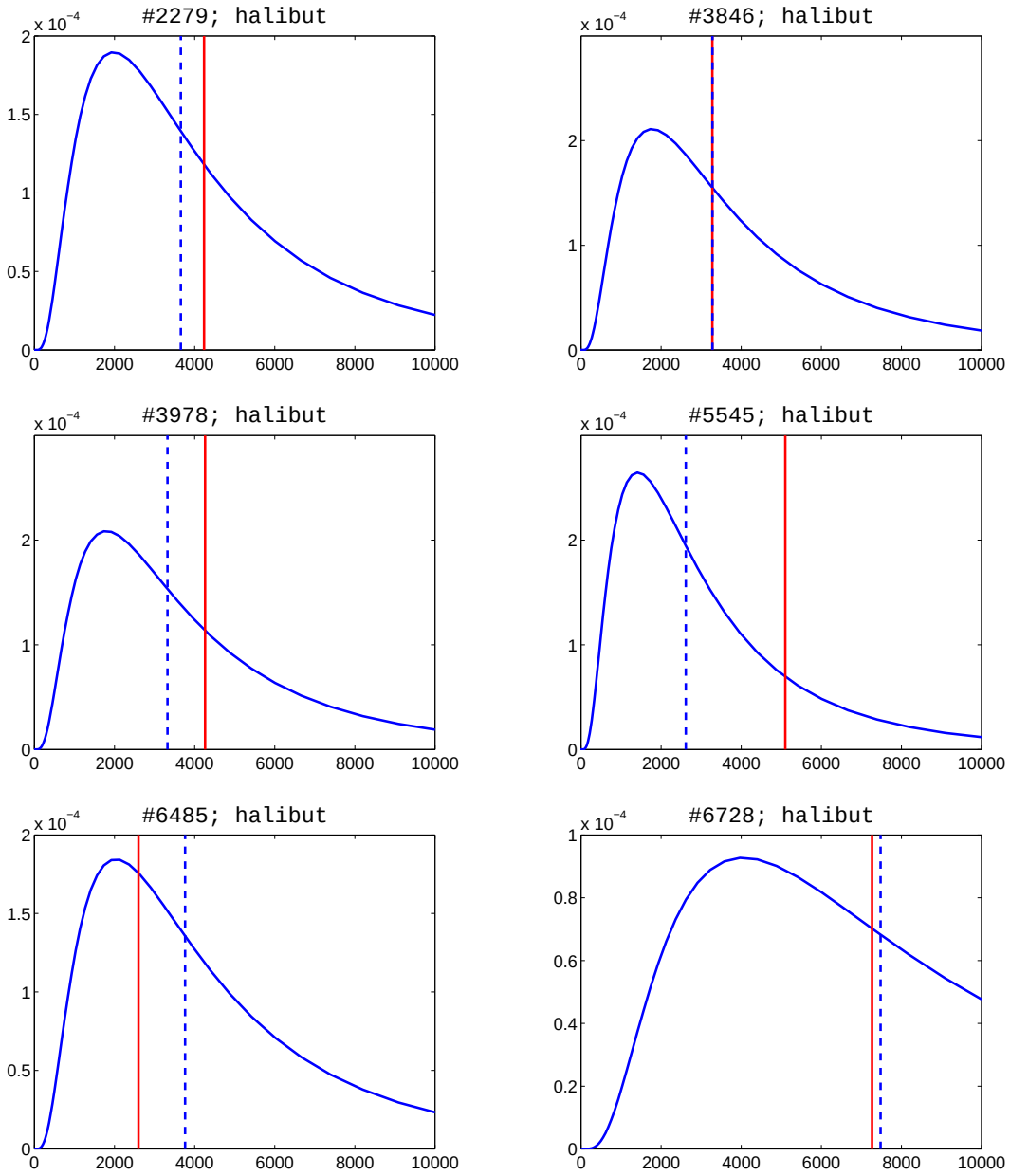

Рис. 2.2. Результаты прогнозирования величины улова палтуса за один промысловый день

2.3.3. Погрешность прогноза

О величине погрешности можно судить, сравнивая точечный прогноз, вычисляемый на основании получаемого распределения вероятности, с известным значением улова. В статистических задачах трудно судить о величине случайной погрешности на основании сравнения результатов в 1-2 расчетных точках. Стандартный подход к тестированию алгоритма оценивания заключается в разбиении входной выборки на «обучающую» и «тестирующую». Первая используется только для настройки модели, вторая — только для прогнозирования и сравнения полученных прогнозов с реальными значениями.

Следует отметить, что, поскольку в данном случае модель, в классическом понимании, не строится, эти два этапа на самом деле явно не разделяются (в том смысле, что невозможно выделить два отдельных последовательно применяемых алгоритма). Однако логика тестирования и принцип разделения выборки на две части соблюдаются.

Исходная выборка случайным образом была поделена на обучающую и тестирующую в соотношении 4:1. Суммарно результаты сравнения прогнозов с эмпирическими данными представлены гистограммами. На графики вынесены частоты (в процентах) «логарифмических погрешностей». А именно, неточность прогноза k-го промыслового дня количественно оценивалась по формуле

$$h(k) = \lg\left(\frac{\hat{Y}(k)}{Y(k)}\right),$$

где $Y(k)$ — реальный улов, $\hat{Y}(k)$ — прогноз. Диапазон возможных значений $h(k)$ разбивался на отрезки с некоторым шагом. Таким образом получалось разбиение на «корзины». После прогона алгоритма прогнозирования на тестирующей выборке вычислялись частоты для всех корзин и строилась стандартная гистограмма (см. рис. 2.3).

В качестве интегрального показателя, характеризующего точность прогнозирования для всей совокупности данных, можно предложить среднюю абсолютную логарифмическую ошибку h, определяемую как среднее значение $|h(k)|$. Значения h для прогнозирования уловов трески и палтуса приведены в подписях к рисункам.

2.3.4. Скорость сходимости метода

О скорости сходимости метода можно судить по скорости, с которой происходит стабилизация оценок распределения вероятности и гистограмм ошибок прогнозирования.

Проведенные вычисления показывают, что для получения оценок близких к предельным (практически от них не отличающихся) вполне достаточно нескольких первых тысяч итераций (в (5) число итераций достигало 1.5 млн., что представляется избыточным).

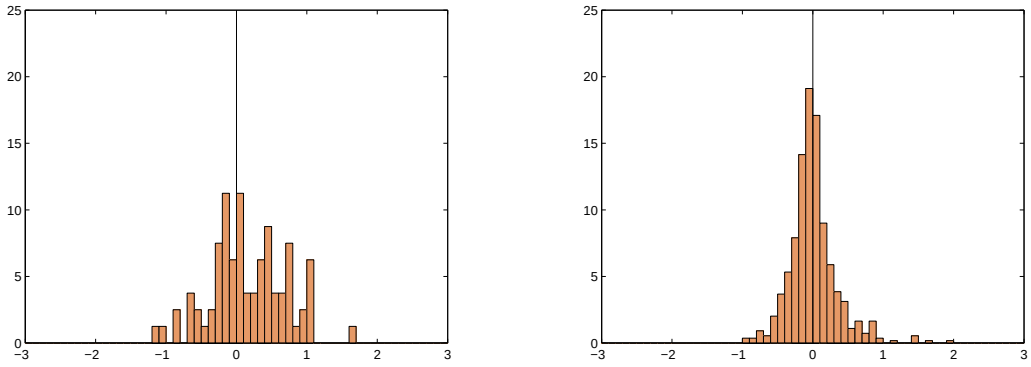

Рис. 2.3. Гистограмма логарифмической оценки точности прогноза с помощью алгоритма MC^3 величины улова: слева — *трески* ($\mathfrak{h} = 0.444$), справа — *палтуса* ($\mathfrak{h} = 0.235$)

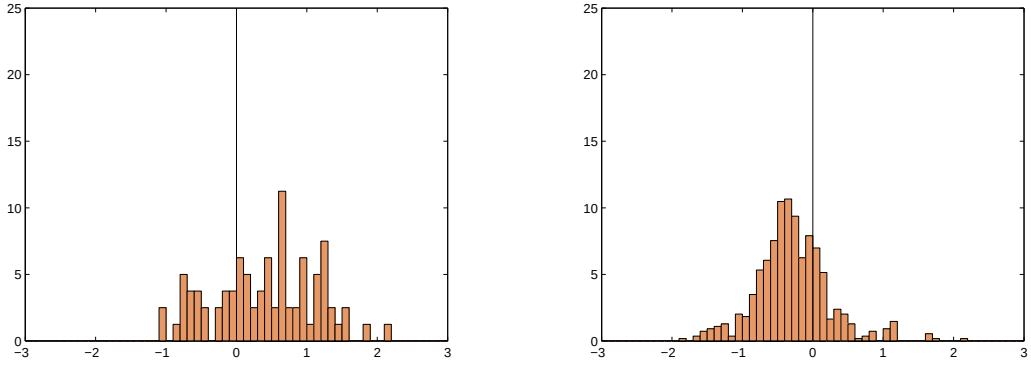

Рис. 2.4. Гистограмма логарифмической оценки точности прогноза с помощью полной регрессионной модели величины улова: слева — *трески* ($\mathfrak{h} = 0.694$), справа — *палтуса* ($\mathfrak{h} = 0.473$)

2.3.5. Сравнение с классической регрессионной моделью

Существенное повышение сложности алгоритма прогнозирования по сравнению с традиционными хорошо изученными и относительно простыми для реализации методами классической линейной регрессии закономерно порождает вопросы, связанные с целесообразностью применения усложненной методики. Отчасти эти сомнения можно снять после сопоставления представленных выше результатов с прогнозом классической линейной регрессионной модели (в случае, если новая методика позволяет получать существенно более точные прогнозы).

Для демонстрации преимуществ исследуемого алгоритма на рис. 2.4 показана гистограмма ошибок прогнозирования (аналогичная изображенным на рис. 2.3). Увеличение средней абсолютной логарифмической ошибки $\hbar$ составляет около 60% для трески и около 100% для палтуса.

Известны и некоторые другие (немногочисленные) примеры, в которых метод усреднения байесовских моделей превосходит по точности классический регрессионный анализ. Разумеется, выводы о превосходстве или степени превосходства одной методики над другой, сделанные на основании нескольких частных случаев, ненадежны. Однако необходимо отметить, что примеров противоположного содержания обнаружить не удалось.

Выводы

На основании опробования алгоритма на данных по рыбному промыслу в Северной Атлантике (Большая Ньюфаундлендская Банка) можно сделать следующие выводы.

1. В целом алгоритм MC^3 показал себя весьма неплохо. Он позволяет построить оценки плотности распределения вероятности в условиях, характеризующихся высокой степенью неопределенности. Эти распределения выглядят естественно и правдоподобно, согласуются с обычными для задач прогнозирования величины улова предположениями, позволяют получать как точечные, так и интервальные прогнозные оценки.

2. Для получения точечной оценки величины улова можно воспользоваться любой характеристикой положения вероятностного распределения. Наиболее распространенными оценками являются среднее, медиана, мода (для унимодальных распределений). Наиболее естественно в данном контексте выглядит медиана.

3. Показатели точности прогноза существенно различны для разных видов промысла (см., например, результаты для трески и палтуса). Для повышения качества прогноза, вероятно, необходима дополнительная информация и более глубокое понимание функциональных и статистических связей между наблюдаемыми характеристиками промысла.

4. Число итераций, необходимых для получения устойчивых оценок может колебаться в широких пределах (от нескольких тысяч до нескольких миллио-

нов и более). Однако для «практически приемлемых» оценок достаточно, по-видимому, нескольких сотен или первых тысяч итераций. Дальнейшие вычисления влияют на среднюю абсолютную логарифмическую погрешность лишь в долях процента. Вероятно, сходимость алгоритма существенно зависит от того, насколько хорошо могут быть представлены данные с помощью небольшого числа моделей (в идеале — какой-либо одной) из множества байесовских моделей, по которым проводится усреднение.

5. Сравнение алгоритма MC^3 и классической регрессионной модели показало преимущество алгоритма MC^3. Итеративное выборочное усреднение даже при относительно небольшом числе итераций (несколько сотен) может повысить среднее качество прогноза (измеряемое в логарифмической шкале) более чем вдвое.

В качестве следующих ближайших шагов можно рекомендовать проведение аналогичной апробации рассматриваемой методики на данных, собранных в интересующих ТИНРО-центр промысловых районах. Для проведения такой апробации необходим доступ к соответствующей информации (представленной в удобном электронном формате). На этапах подготовки априорной информации и интерпретации получаемых результатов крайне желательно заинтересованное участие специалистов по рыбному промыслу.

Литература

[1] Абакумов А.И., Бочаров Л.Н., Каредин Е.П. Модельный анализ многовидовых рыбных промыслов. – Изв. ТИНРО, 2004, т. 138, 220–224.

[2] Зельнер А. Байесовские методы в эконометрии. М.: Статистика, 1980. – 438 с.

[3] Chib S., Greenberg E. Markov chain Monte Carlo simulation methods in econotrics. – Econometric theory, 1996, 12, 409-431.

[4] Madigan D., York J., Bayesian graphical models for discrete data. – International Statistical Review, 1995, 63, 215–232.

[5] Fernandez C., Ley E., Steel M.F.J. Bayesian modelling of catch in a Northwest Atlantic fishery. – Journal of the Royal Statistical Society, Series C (Applied Statistics), 2002, 51, 257–280.

[6] Fernandez C., Ley E., Steel M.F.J. Benchmark Priors for Bayesian Model Averaging. – Journal of Econometrics, 2001, vol. 100, 2, 381–427.

[7] Ferreira E., Tusell F. Un modelo aditivo semiparametrico para estimacion de capturas: el caso de las pesquerias de Terranova (A model semiparametric additive for estimation of captures: the case of pesquerias of Newfoundland). – Economic investigations, 1996, vol. XX, 142-157.

[8] Jeffreys H. Theory of probability. – Oxford: Clarendon, 1966. – 240 p.