Московский государственный университет имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

Осипова Анастасия Андреевна

Адаптивные статистические алгоритмы оценивания параметров конечных смешанных нормальных моделей

ВВЕДЕНИЕ В МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ

Научный руководитель:
д.ф-м.н., профессор
Королев Виктор Юрьевич

Научный консультант:
к.ф-м.н., доцент
Горшенин Андрей Константинович

Москва, 2020

Содержание

1 Введение

Метод скользящего разделения смесей (СРС-метод) является развитием идеи ЕМ-алгоритма, применяемого в статистике для поиска оценок максимального правдоподобия в случаях, когда целевая функция правдоподобия имеет сложную структуру. СРС-метод позволяет в динамическом режиме выделять компоненты смеси и применяется в таких областях, как физика турбулентной плазмы [1], обработка данных финансовых рынков [2], анализ потоков тепла между океаном и атмосферой [3].

Данная работа состоит из трех логических блоков, каждому из которых посвящена отдельная глава. В первом блоке рассматривается адаптивная модификация СРС-метода, изучается ее точность и условия применимости для различных взаимных вариантов расположения компонент сигнала и шума. Также представлен алгоритм автоматического определения момента разладки (момента, когда меняется структура данных и в записи помимо шума оборудования появляется полезный сигнал).

Во втором блоке предлагается алгоритм выделения компонент связности после обработки данных СРС-методом для восстановления "истории" развития каждой из компонент и демонстрируется его применение к данным по потокам тепла в различных климатических зонах.

В третьем блоке рассматривается возможность использования метода для улучшения точности прогнозирования временных рядов путем расширения признакового пространства информацией об эволюции компонент. Для применения этого подхода требуется лишь предположение о подчинении данных рассматриваемого временного ряда структуре конечной смеси нормальных законов, что является достаточно широко используемой моделью описания данных, например, в области эволюции финансовых потоков.

2 Постановка задачи

Требуется провести широкий анализ практического применения адаптивного метода выделения сигнала на фоне зашумленных данных. Можно выделить три основных части, в каждой из которых решается отдельная подзадача, связанная с данным методом:

- Анализ эффективности адаптивного метода скользящего разделения смесей (адаптивного СРС-метода)
- Выделение компонент связности из оценок, полученных алгоритмом
- Расширение пространства признаков при построении нейронных сетей

Каждая из подзадач является достаточно самостоятельной и полноценной задачей, поэтому под каждую из них отведён отдельный раздел, где содержится математическая постановка задачи, описание решения и полученные результаты, а для второй и третьей подзадач – связь с предыдущими этапами.

3 Анализ эффективности адаптивного метода

На первом этапе работы была продемонстрирована эффективность предложенной в работе [4] процедуры определения параметров полезного сигнала при условии наличия шума для различных соотношений между их параметрами на примере 24 модельных выборок, охватывающих большинство возможных реальных сценариев. Также в данном разделе обсуждаются вопросы прикладного подхода к обнаружению момента появления полезного сигнала в наблюдениях.

Результаты работы по данному этапу были опубликованы в статье [5].

3.1 Предпосылки

Наблюдения (сигналы) в реальных системах зачастую регистрируются с округлениями и дополнительной шумовой составляющей, которая возникает из-за случайных флуктуаций в работе экспериментального оборудования или внешних факторов. Очевидно, что такие модификации получаемой выборки не связаны непосредствено с проводимым экспериментом, однако влияют на его результаты. Данная проблема характерна для широкого спектра исследовательских задач, в том числе в медицинских приложениях [6, 7], при анализе сигналов с негауссовским шумом [8–10], предобработке изображений [11].

Особенности работы с округленными данными изучались в статье [12]. Для учета влияния случайного шума в статье [4] была предложена модель для исходных наблюдений на основе случайной величины (с.в.) Z, которая может быть представлена в виде суммы независимых с.в. X (полезный сигнал) и Y (аддитивный шум) с различными смешанными конечными нормальными распределениями. Предполагается, что до начала эксперимента может быть получена выборка реализаций только с.в. Y достаточно большого объема. Данное требование не является ограничительным, поскольку обычно основные сложности регистрации связаны непосредственно с экспериментом (ограниченное время наблюдения за процессом, разрешающая способность оборудования и т.п.), в то время как предварительный запуск детектирующих приборов и получение данных с них являются достаточными простыми процедурами. Оценивание параметров проводится в режиме сдвигающегося окна с помощью метода скользящего разделения смесей [13].

3.2 Постановка задачи

Обозначим X – полезный сигнал, Y – аддитивный шум, $Z = X + Y$ – наблюдаемые в эксперименте величины:

$$X \sim \sum_{j=1}^{k} p_j \Phi\left(\frac{x - a_j}{\sigma_j}\right), \tag{1}$$

$$\text{где } \sum_{j=1}^{k} p_j = 1,\ p_j \geqslant 0,\ a_j \in \mathbb{R},\ \sigma_j > 0,\ j = \overline{1,k},$$

$$Y \sim \sum_{j=1}^{\widetilde{k}} \widetilde{p}_j \Phi\left(\frac{x - \widetilde{a}_j}{\widetilde{\sigma}_j}\right), \tag{2}$$

$$\text{где } \sum_{j=1}^{\widetilde{k}} \widetilde{p}_j = 1,\ \widetilde{p}_j \geqslant 0,\ \widetilde{a}_j \in \mathbb{R},\ \widetilde{\sigma}_j > 0,\ j = \overline{1,\widetilde{k}},$$

$$Z \sim \sum_{l=1}^{k\cdot\widetilde{k}} \widehat{p}_l \Phi\left(\frac{x - \widehat{a}_l}{\widehat{\sigma}_l}\right), \tag{3}$$

$$\text{где } \sum_{j=1}^{k\cdot\widehat{k}} \widehat{p}_j = 1,\ \widehat{p}_j \geqslant 0,\ \widehat{a}_j \in \mathbb{R},\ \widehat{\sigma}_j > 0,$$

$$\widehat{p}_{(r-1)\widetilde{k}+j} = p_r \widetilde{p}_j,\ \widehat{a}_{(r-1)\widetilde{k}+j} = a_r + \widetilde{a}_j,\ \widehat{\sigma}^2_{(r-1)\widetilde{k}+j} = \sigma_r^2 + \widetilde{\sigma}_j^2, \tag{4}$$
$$r = \overline{1,k},\ j = \overline{1,\widetilde{k}},$$

и $\Phi(x)$ – функция распределения стандартного нормального закона. В данном случае, все величины в выражении (1), включая и число компонент k, считаем неизвестными, а в выражении (2) – предварительно оцененными, например, с помощью какой-либо модификации EM-алгоритма.

Тогда, как показано в статье [4], оценки параметров неизвестного распределения с.в. X (1) по оценкам (4) с.в. Z (3) определяются следующими соотношениями:

$$a_r = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \left(\widehat{a}_{(r-1)\widetilde{k}+j} - \widetilde{a}_j\right), \tag{5}$$

$$\sigma_r^2 = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \left(\widehat{\sigma}^2_{(r-1)\widetilde{k}+j} - \widetilde{\sigma}_j^2\right), \tag{6}$$

$$p_r = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \widehat{p}_{(r-1)\widetilde{k}+j} \cdot \widetilde{p}_j^{-1}. \tag{7}$$

4 Выделение компонент связности

На втором этапе данный метод был применен к экспериментальным данным, описывающим поведение турбулентных потоков тепла в между океаном и атмосферой, а также решена подзадача о выделении компонент связности из полученных СРС-методом оценок.

В статье [3] подробно описано теоретическое обоснование применения данного метода для статистического оценивания коэффициентов стохастического дифференциального уравнения Ланжевена.

4.1 Постановка задачи

В физике стохастическим дифференциальным уравнением (СДУ) Ланжевена принято называть следующее соотношение:

$$dX(t) = a(t)dt + b(t)dW, \tag{8}$$

определяющее случайный процесс $X(t)$, где $W(t)$ – винеровский процесс, а а коэффициенты $a(t)$ и $b(t)$ – случайны. СДУ вида (8) широко используются, например, в задаче ассимиляции данных при анализе разномасштабной изменчивости геофизических переменных [14]. В финансовой математике известны специальные версии уравнения (8). В частности, весьма популярна модель геометрического броуновского движения вида

$$dX(t) = aX(t)dt + bX(t)dW, \tag{9}$$

где $a \in \mathbb{R}$, $b > 0$. Известно много обобщений модели (9) с конкретными видами зависимости a и b от $X(t)$ и других случайных процессов, например модели Леланда [15], Барлса–Сонера [16], Хестона [17], Кокса–Ингерсолла–Росса [18], Халла–Уайта [19] и другие так называемые модели стохастической волатильности (см. также [20–22]).

При отсутствии априорной информации о «физической» структуре процесса $X(t)$ для успешного изучения и прогнозирования его эволюции первостепенную важность приобретает задача статистического оценивания функциональных коэффициентов $a(t)$ и $b(t)$. В силу их случайности данная задача допускает как минимум две принципиально разные формулировки. Во-первых, можно пытаться найти (случайные же) оценки самих функций $a(t)$ и $b(t)$, то есть найти их точечные аппроксимации, и, во-вторых, пытаться статистически оценить распределения случайных величин $a(t)$ и $b(t)$. Во втором случае, зная какие-либо свойства этих коэффициентов, например структуру их функциональной зависимости от исходного процесса $X(t)$ (как в упомянутых выше моделях Леланда, Барлса–Сонера, Хестона, Кокса–Ингерсолла–Росса или Беляева), можно найти оценки числовых параметров этих моделей.

Рассматривается второй вариант постановки задачи.

5 Расширение пространства признаков

На последнем, третьем этапе работы освещается вопрос использования рассматриваемого метода для выделения дополнительной информации и расширения за счет этого пространства признаков нейронных сетей для улучшения качества получаемого прогноза временных рядов.

5.1 Постановка задачи

Пусть даны измерения некоей величины X, составляющие временной ряд: $X_1, ..., X_N$. Требуется предложить и реализовать метод расширения пространства признаков на основе изучаемого адаптивного метода. Целью является улучшение точности прогнозирования данного временного ряда с помощью нейронных сетей на среднюю длину окна (около 30 измерений).

Рассматривается следующий набор архитектур нейронных сетей: LSTM-сети, Feedforward и Deep Feedforward сети, CNN-сети. Для каждой архитектуры планируется провести обучение для трех вариантов представления пространства признаков:

- Необогащенное пространство признаков – при получении предсказания на вход сети подается лишь окно с предшествующими искомому промежутку значениями величины X (так называемые лаги временного ряда),
- Обогащенное моментами пространство признаков – к входному вектору нейросети добавляются заранее посчитанные моменты выборки,
- Обогащенное моментами и компонентами пространство признаков – к входному вектору добавляются заранее посчитанные моменты выборки, а также параметры компонент, оцененные при помощи адаптивного метода и выделения связности, описанных в предыдущих разделах.

5.2 Ход работы

Данный этап пока не завершен и находится в процессе доработки.

Список литературы

[1] *Королев, ВЮ.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. Учебное пособие / ВЮ Королев. — 2011.

[2] *Skvortsova, NN.* Estimation of dynamic and diffusive components in edge turbulent particle fluxes in the L-2M stellarator and the FT-2 tokamak / NN Skvortsova, GM Batanov, DV Malakhov, AE Petrov, VV Saenko, KA Sarksyan, NK Kharchev, Yu V Kholnov, V Yu Korolev, Yu V Zhukov et al. // 21st IAEA Fusion Energy Conference. — 2006.

[3] *Горшенин, Андрей Константинович.* Статистическое оценивание распределений случайных коэффициентов стохастического дифференциального уравнения Ланжевена / Андрей Константинович Горшенин, Виктор Юрьевич Королев, Анастасия Андреевна Щербинина // *Информатика и её применения.* — 2020. — Vol. 14, no. 3. — Pp. 3–12.

[4] *Gorshenin, AK.* Adaptive Detection of Normal Mixture Signals with Pre-Estimated Gaussian Mixture Noise / AK Gorshenin // *Pattern Recognition and Image Analysis.* — 2019. — Vol. 29, no. 3. — Pp. 377–383.

[5] *Gorshenin, AK.* Efficiency of the Method for Detecting Normal Mixture Signals with Pre-Estimated Gaussian Mixture Noise / AK Gorshenin, AA Shcherbinina // *Pattern Recognition and Image Analysis.* — 2020. — Vol. 30, no. 3. — Pp. 470–479.

[6] *Márquez-Figueroa, Sandra.* Optimal extraction of EMG signal envelope and artifacts removal assuming colored measurement noise / Sandra Márquez-Figueroa, Yuriy S Shmaliy, Oscar Ibarra-Manzano // *Biomedical Signal Processing and Control.* — 2020. — Vol. 57. — P. 101679.

[7] *Almgren, Hannes.* The effect of global signal regression on DCM estimates of noise and effective connectivity from resting state fMRI / Hannes Almgren, Frederik Van de Steen, Adeel Razi, Karl Friston, Daniele Marinazzo // *NeuroImage.* — 2020. — Vol. 208. — P. 116435.

[8] *Asadi, Hamid.* Signal enumeration in Gaussian and non-Gaussian noise using entropy estimation of eigenvalues / Hamid Asadi, Babak Seyfe // *Digital Signal Processing.* — 2018. — Vol. 78. — Pp. 163–174.

[9] *Ilter, Mehmet Cagri.* The Joint Impact of Fading Severity, Irregular Constellation, and Non-Gaussian Noise on Signal Space Diversity-Based Relaying Networks / Mehmet Cagri Ilter, Hamza Umit Sokun,

Halim Yanikomeroglu, Risto Wichman, Jyri Hämäläinen // *IEEE Access.* — 2019. — Vol. 7. — Pp. 116162–116171.

[10] *Guo, Junchao.* An enhanced modulation signal bispectrum analysis for bearing fault detection based on non-Gaussian noise suppression / Junchao Guo, Hao Zhang, Dong Zhen, Zhanqun Shi, Fengshou Gu, Andrew D Ball // *Measurement.* — 2020. — Vol. 151. — P. 107240.

[11] *Li, Yongsong.* Noise estimation for image sensor based on local entropy and median absolute deviation / Yongsong Li, Zhengzhou Li, Kai Wei, Weiqi Xiong, Jiangpeng Yu, Bo Qi // *Sensors.* — 2019. — Vol. 19, no. 2. — P. 339.

[12] *Gorshenin, Andrey Konstantinovich.* Data noising by finite normal and gamma mixtures with application to the problem of rounded observations / Andrey Konstantinovich Gorshenin // *Informatika i Ee Primeneniya [Informatics and its Applications].* — 2018. — Vol. 12, no. 3. — Pp. 28–34.

[13] *Korolev, V Yu.* Probabilistic and statistical methods of decomposition of volatility of chaotic processes / V Yu Korolev // *Izd. Mosk. Gos. Univ., Moscow.* — 2011.

[14] *Belyaev, Konstantin.* An optimal data assimilation method and its application to the numerical simulation of the ocean dynamics / Konstantin Belyaev, Andrey Kuleshov, Natalia Tuchkova, Clemente AS Tanajura // *Mathematical and Computer Modelling of Dynamical Systems.* — 2018. — Vol. 24, no. 1. — Pp. 12–25.

[15] *Leland, Hayne E.* Option pricing and replication with transactions costs / Hayne E Leland // *The journal of finance.* — 1985. — Vol. 40, no. 5. — Pp. 1283–1301.

[16] *Barles, Guy.* Option pricing with transaction costs and a nonlinear Black-Scholes equation / Guy Barles, Halil Mete Soner // *Finance and Stochastics.* — 1998. — Vol. 2, no. 4. — Pp. 369–397.

[17] *Heston, Steven L.* A closed-form solution for options with stochastic volatility with applications to bond and currency options / Steven L Heston // *The review of financial studies.* — 1993. — Vol. 6, no. 2. — Pp. 327–343.

[18] *Cox, John C.* A theory of the term structure of interest rates / John C Cox, Jonathan E Ingersoll Jr, Stephen A Ross // Theory of valuation. — World Scientific, 2005. — Pp. 129–164.

[19] *Hull, John.* The pricing of options on assets with stochastic volatilities / John Hull, Alan White // *The journal of finance.* — 1987. — Vol. 42, no. 2. — Pp. 281–300.

[20] *Derman, Emanuel*. Riding on a smile / Emanuel Derman, Iraj Kani // *Risk*. — 1994. — Vol. 7, no. 2. — Pp. 32–39.

[21] *Dupire, Bruno*. Pricing with a smile / Bruno Dupire et al. // *Risk*. — 1994. — Vol. 7, no. 1. — Pp. 18–20.

[22] *Ширяев, Альберт Николаевич*. Основы стохастической финансовой математики. Том 1. Факты. Модели. 2-е изд., испр. / Альберт Николаевич Ширяев. — Фазис, 2004.