ЛЕКЦИЯ 3

EM-алгоритм: ещё ММП

На этой неделе мы рассмотрим первую вероятностную модель, которую нельзя оценить методом максимального правдоподобия в обычном смысле в силу того, что не хватает нужных данных. Тем не менее, можно на основе принципа ММП построить итерационный алгоритм, который эту задачу поможет хорошо решить.

3.1 Скрытые переменные

В попытке построить вероятностную модель выборки X часто оказывается полезным учесть какой-то второй фактор, которого в данных нет. Конкретнее, предполагается, что есть ещё некоторые величины Y_i которые как-то влияют на распределение X_i и учёт подобного влияния позволяет построить, с одной стороны, технически более логичную модель из простых блоков, а с другой – лучше объяснить происхождение данных с позиции опыта и здравого смысла. Рассмотрим такой пример.

Пример 3.1. *В конце года после экзамена есть хорошая традиция собраться вместе и пообщаться за пиццей и настолками. В такой ситуации всегда надо заказывать достаточно много пицц. Конечно, коробки разных пиццерий стараются делать разными, но иногда это не очень выходит или нужно смотреть на какие-то совсем мелкие детали, вертеть в руках коробку, чтобы отличить одну пиццерию от другой.*

Представим себе ситуацию. Было перепробовано n похожих пицц из двух пиццерий, про каждую разные группы людей сошлись во мнении: от не очень (0) до отлично(5). Можем ли мы понять, не заглядывая в чеки, какова была доля пицц из пиццерии ГогоПицца и БотманПицца? Но ещё, можно ли понять как по качеству каждая пиццерия готовит разные пиццы (распределение по оценкам от не очень, до отлично)?

n пицц $\longrightarrow X_1, \dots, X_n$ — не очень / ок / отлично

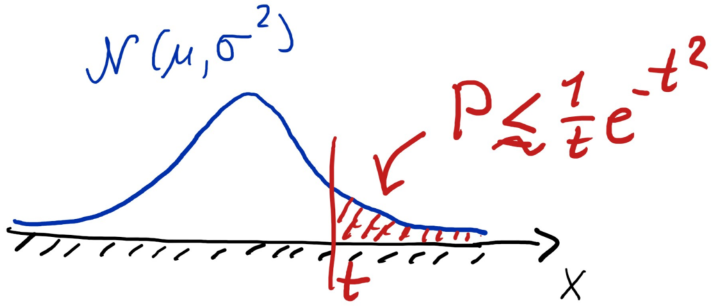
В данном примере формально мы наблюдаем только независимую выборку из оценок пицц $X_1, .., X_n$. Что мы можем по ней сказать? Мы можем в целом понять, какая доля пицц какой оценке соответствует, то есть, построить гистограмму. При этом учитывая простоту модели, оценка ММП (если предположить, что пиццы приходят из одного распределения и независимо) получится удовлетворительной. На этом список наших возможностей заканчивается.

На вопрос из примера так мы не сможем ответить; для этого нужно пересмотреть вероятностную модель. По сути у нас есть две переменных, про которые хочется что-то узнать: название пиццерии Y и качество пиццы X. При этом распределение по качеству в разных пиццериях может быть разным. Давайте предложим следующую идею:

$$\xrightarrow{p_\theta(y)} Y \xrightarrow{p_\theta(x|y)} X$$

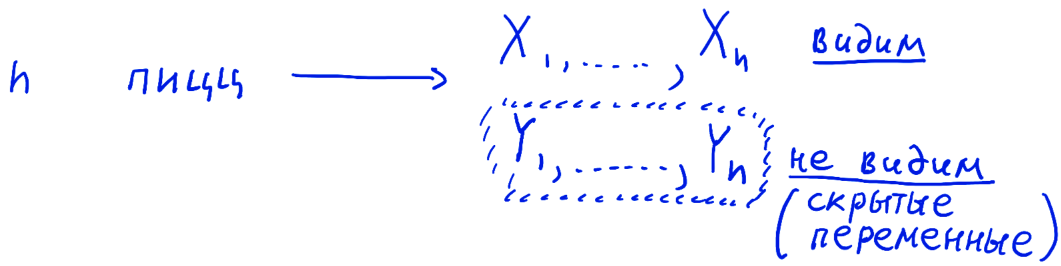
Все переменные X_i, а также все Y_i независимы в совокупности, но X_i и Y_i зависимы: мы предполагаем что сначала семплируется пиццерия Y, а затем на основе результата семплируется пицца. Здесь мы использовали сокращённую нотацию, обозначив за θ все параметры вероятностной модели, а p_θ обозначает конкретное распределение (вероятность или плотность), которое понятно из контекста. У нас всего две пиццерии, поэтому вначале семплируется id пиццерии Y (1 с вероятностью p и 2 с вероятностью $1-p$), затем семплируется качество из распределения, которое у каждой пиццерии своё.

Здесь нам пригодится модельное предположение; предположим, что X при условии $Y = k$ имеет гауссовское распределение $N(\mu_k, \sigma_k^2)$. Это не совсем корректно, так как теоретически в такой модели мы можем получить пиццу с оценкой -100500 или +100500 (впрочем, в реальности тоже бывает...). Но подумайте, из каких факторов складывается качество пиццы? Повар мог не выспаться, доставщик мог долго ехать, на кассе могли долго обсчитывать заказ, – всё это большое количество почти независимых событий, которые ограниченно меняют итоговое качество. Для таких случаев как раз работает ЦПТ: может, наша модель (как любая модель в реальном мире) не точная, но очень большие или очень маленькие значения будут возникать с экспоненциально маленькой вероятностью.



3.2 ЕМ-алгоритм

Как оценивать такую модель? У нас есть только выборка $X_1, .., X_n$, при этом модель включает в себя *скрытые* $Y_1, .., Y_n$, которые в данных не наблюдаются.



ЕМ-алгоритм [4] решает эту проблему, предлагая интересный подход на основе метода максимального правдоподобия, но с естественным способом адресовать неопределённость вокруг набора Y. Помните, мы недавно рассматривали метод максимального правдоподобия как попытку минимизировать аппроксимированную отрицательную кросс-энтропию?

$$-CE(p_{\theta_0}||p_\theta) = \mathbb{E}_{\theta_0}\left[\ln p_\theta(X)\right] \approx \frac{1}{n}\mathbb{E}_{\theta_0}\left[\ln p_\theta(X)|X\right] = \frac{1}{n}l(\theta, X).$$

Мы будем опускать константу $1/n$ как несущественную для оптимизации. В основе знака приближённого равенства лежала идея того, что мы можем в условие матожидания поставить то, что известно из данных, а по остальному взять матожидание. В нашем случае со скрытыми переменными мы получим с тем же подходом

$$\mathbb{E}_{\theta_0}\left[\ln p_\theta(X, Y)\right]? \approx \mathbb{E}_{\theta_0}\left[\ln p_\theta(X, Y)|X\right],$$

где при известных θ_0 мы могли бы явно посчитать это выражение и прооптимизировать по θ.

В этом смысле ЕМ-алгоритм предлагает следующее:

1. Задать произвольный стартовый θ_0;

2. (Е-шаг) Посчитать матожидание справа;

3. (M-шаг) Максимизировать $\mathbb{E}_{\theta_0}\left[\ln p_\theta(X,Y)|X\right]$ по θ, получив $\hat{\theta}$;
4. Задать $\theta_0 = \hat{\theta}$ и повторить с пункта 2.

Пока это выглядит так, будто мы добавили изоленты к изначальному ММП. Но давайте попробуем посмотреть, чем именно является Q-функция EM-алгоритма

$$Q(\theta_0,\theta) = \mathbb{E}_{\theta_0}\left[\ln p_\theta(X,Y)|X\right],$$

интуиция в данном случае имеет под собой хороший фундамент. В общем случае мы не можем гарантировать сходимость, каждая задача в этом смысле имеет свою специфику, но некоторые общие гарантии мы можем предоставить.

Утверждение 3.1. *С ростом итерации θ_k EM-алгоритма правдоподобие модели при данных $x_1,..,x_n$ не ухудшается:*

$$p_{\theta_k}(x) \geq p_{\theta_{k-1}}(x).$$

$\triangleright$ В качестве соглашения примем, что x и y – реализации всего датасета, а в условных матожиданиях для краткости будем часто писать X вместо $X=x$. Кроме того, выкладки универсальны по модулю технических тонкостей: если работаем с недискретным Y, формулы тоже действительны с заменой суммы на интеграл. Если под логарифмом умножить и поделить на условную вероятность, то можно увидеть знакомые вещи:

$$Q(\theta_{k-1},\theta) = \mathbb{E}_{\theta_{k-1}}\left[\ln p_\theta(X,Y)|X\right] = \mathbb{E}_{\theta_0}\left[\ln\frac{p_\theta(X,Y)}{p_\theta(Y|X)} \mid X\right] + \mathbb{E}_{\theta_{k-1}}\left[\ln p_\theta(Y|X) \mid X\right].$$

Действительно, тут уже что-то похожее на знакомые энтропии. Давайте ещё немного упростим, воспользовавшись определением условной вероятности:

$$= \ln p_\theta(x) + \mathbb{E}_{\theta_{k-1}}\left[\ln p_\theta(Y|X) \mid X\right]$$

Сравним два лог-правдоподобия; используем полученную формулу для выражения каждого и сразу увидим знакомые вещи:

$$\ln p_\theta(x) - \ln p_{\theta_{k-1}}(x) = Q(\theta_{k-1},\theta) - Q(\theta_{k-1},\theta_{k-1}) + CE\left(p_{\theta_{k-1}}(\cdot|x) \mid p_\theta(\cdot|x)\right) - H\left(p_{\theta_{k-1}}(\cdot|x)\right).$$

Здесь точкой обозначен аргумент y по, которому считаются энтропии. А конкретнее

$$\ln p_\theta(x) - \ln p_{\theta_{k-1}}(x) = Q(\theta_{k-1},\theta) - Q(\theta_{k-1},\theta_{k-1}) + D_{KL}\left(p_{\theta_{k-1}}(\cdot|x) \mid p_\theta(\cdot|x)\right).$$

Так, вспомнив про неравенство Гиббса, мы получаем неравенство

$$\ln p_\theta(x) - \ln p_{\theta_{k-1}}(x) \geq Q(\theta_{k-1},\theta) - Q(\theta_{k-1},\theta_{k-1}).$$

Теперь заметим, что согласно нашей конструкции $\theta_k = \arg\max_\theta Q(\theta_{k-1},\theta)$, поэтому $Q(\theta_{k-1},\theta_k) \geq Q(\theta_{k-1},\theta_{k-1})$ и подставив $\theta=\theta_k$, получим в результате

$$\ln p_{\theta_k}(x) - \ln p_{\theta_{k-1}}(x) \geq 0.$$

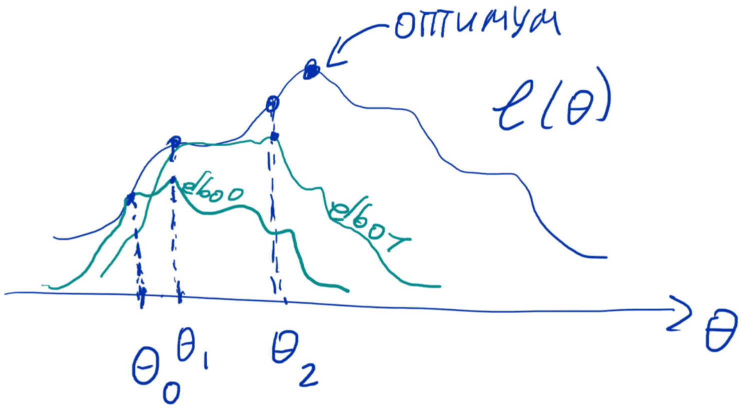
□

Можно посмотреть на Q-функцию ещё с другой стороны. Лог-правдоподобие можно получить интегрированием/суммированием совместного распределения по всему пространству:

$$\ln p_\theta(x) = \ln \int p_\theta(x, y) dy.$$

Давайте домножим и поделим на какое-то распределение $q(y)$ (главное, чтобы не было 0 там, где $p(x, y)$ не ноль) под знаком интеграла, получим отличный повод применить неравенство Йенсена:

$$\ln p_\theta(x) = \ln \int \frac{p_\theta(x, y)}{q(y)} q(y) dy \geq \int \ln \frac{p_\theta(x, y)}{q(y)} q(y) dy = ELBO(q, p_\theta).$$



Получившееся выражение можно встретить сейчас повсеместно под названием *ELBO (Evidence Lower Bound)*, сам $\ln p_\theta(x)$ называется *свидетельством (evidence)*. По сути любое технически совместимое распределение $q(y)$ (его называют *вариационным распределением, variational distribution*) даёт нижнюю оценку на лог-правдоподобие. ELBO внешне почти KL-дивергенция, используем определение условной вероятности, чтобы довершить:

$$= \int \ln \frac{p_\theta(y|x)}{q(y)} q(y) dy + \int \ln \frac{p_\theta(x)}{q(y)} q(y) dy = -D_{KL}(q | p_\theta(\cdot|x)) + \ln p_\theta(x) + H(q).$$

Если мы можем брать более-менее любое вариационное распределение, почему бы не взять сразу $q(y) = p_{\theta_0}(y|x)$, тогда

$$ELBO(q, p_\theta)(x) = \int \ln p_\theta(x, y) p_{\theta_0}(y|x) dy - \int \ln p_{\theta_0}(y|x) p_{\theta_0}(y|x) dy = Q(\theta_0, \theta) + H(p_{\theta_0}(\cdot|x)).$$

По своей сути максимазция Q-функции EM-алгоритма – это максимизация ELBO, так как правое слагаемое не зависит от θ.

Максимизируя ELBO в EM-алгоритме, мы одновременно максимизируем правдоподобие и контролируем KL-дивергенцию между вариационным распределением (в нашем случае $q(y) = p_{\theta_0}(y|x)$), которое часто выбирают не зависящим от θ и получающимся условным апостериорным распределением $p_\theta(y|x)$ с параметрами θ. Многие вероятностные модели со скрытыми переменными, которые относят к Байесовским методам (например, вариационные автокодировщики, VAE), изначально строятся как алгоритмы оптимизации ELBO. Конкретно в VAE KL-дивергенция выступает как регуляризатор обычного автоэнкодера и позволяет добиваться лучших результатов.

Заметьте, что оптимизация ELBO и в частности EM-алгоритм – это общий рецепт, мы для вывода не накладывали никаких ограничений на структуру зависимости между X и Y. Главное – суметь это всё запустить.

3.3 Где пицца лучше?

Давайте вернёмся к примеру. Модель была устроена так: сначала семплируется пиццерия (выбирается случайно из 1 с вероятностью p и 2 с вероятностью $1-p$), затем семплируется качество пиццы (из соответствующего нормального распределения). Предположим, что параметр $\theta_0 = [p^0, \mu_1^0, \mu_2^0, \sigma_1^0, \sigma_2^0]$ мы задали, тогда первая наша задача – вычислить аналитически матожидание

$$Q(\theta_0, \theta) = \mathbb{E}_{\theta_0}\left[\ln p_\theta(X, Y)|X\right].$$

Для этого воспользуемся сначала независимостью по индексу наблюдения:

$$Q(\theta_0, \theta) = \mathbb{E}_{\theta_0}\left[\sum_{i=1}^{n} \ln p_\theta(X_i, Y_i) \quad | \ X\right].$$

Здесь нам нужно посчитать условное матожидание; обозначим за x_i реализацию элемента выборки и заметим, что для этого нужно знать условные вероятности (или плотности, где нужно)

$$p_{\theta_0}(Y_i = y|X_i = x_i) = \frac{p_{\theta_0}(X_i = x_i|Y_i = y)p_{\theta_0}(y)}{p_{\theta_0}(X_i = x_i)} =$$
$$= \frac{p_{\theta_0}(X_i = x_i|Y_i = y)p_{\theta_0}(Y_i = y)}{p_{\theta_0}(X_i = x_i|Y_i = 1)p_{\theta_0}(Y_i = 1) + p_{\theta_0}(X_i = x_i|Y_i = 2)p_{\theta_0}(Y_i = 2)} =: \gamma_{Y_i}(y).$$

В силу того, что нам дан θ_0, всё это можно вычислить. К примеру, в случае $Y_i = 1, X_i = 2$

$$\gamma_{Y_i}(1) = p_{\theta_0}(Y_i = 1|X_i = 2) = \frac{N(\mu_1^0, (\sigma_1^0)^2; 2)p^0}{N(\mu_1^0, (\sigma_1^0)^2; 2)p^0 + N(\mu_2^0, (\sigma_2^0)^2; 2)(1 - p^0)},$$

Где $N(\mu, \sigma^2; x)$ – значение плотности нормального распределения с заданными параметрами в точке x.

Вернёмся в Q; по свойству линейности матожидание можно записать в другом виде:

$$\mathbb{E}_{\theta_0}\left[\sum_{i=1}^{n} \ln p_\theta(X_i, Y_i) \ \mid X\right] = \sum_{i=1}^{n} \gamma_{Y_i}(1) \ln\left(p_\theta(x_i|Y_i = 1)p_\theta(Y_i = 1)\right) + \gamma_{Y_i}(2) \ln\left(p_\theta(x_i|Y_i = 2)p_\theta(Y_i = 2)\right)$$

и более компактно получаем

$$Q(\theta_0, \theta) = \sum_{i=1}^{n} \gamma_{Y_i}(1) \ln\left(p_\theta(x_i|Y_i = 1)p\right) + \gamma_{Y_i}(2) \ln\left(p_\theta(x_i|Y_i = 2)(1 - p)\right).$$

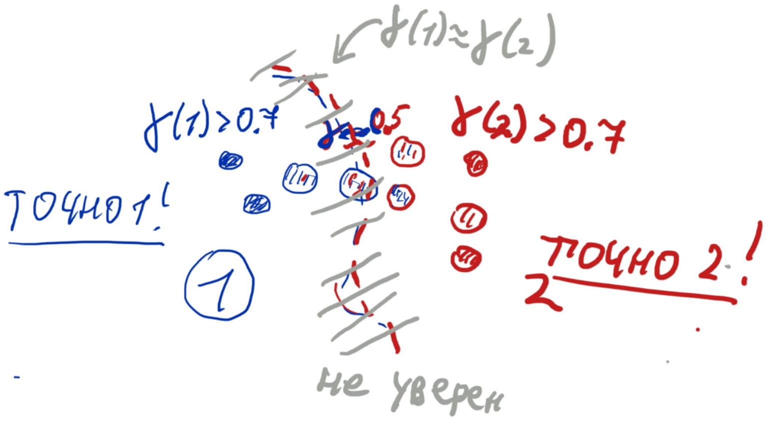
Далее для M-шага нужно максимизировать Q по θ, но это оказывается не очень сложно, так как переменные p и μ, σ разделяются на отдельные суммы и можно комфортно оптимизировать всё отдельно и условие второго порядка гарантирует оптимальность. Из-за весов γ_{Y_i} решение будет отличаться от привычного по методу максимального правдоподобия. Для краткости введём

$$\Gamma_1 = \sum_{i=1}^{n} \gamma_{Y_i}(1), \quad \Gamma_2 = \sum_{i=1}^{n} \gamma_{Y_i}(2),$$

тогда в итоге получим оценки

$$\hat{p} = \frac{\Gamma_1}{n},$$
$$\hat{\mu_j} = \frac{1}{\Gamma_j} \sum_{i=1}^{n} \gamma_{Y_i}(j) X_i,$$
$$\sigma_j^2 = \frac{1}{\Gamma_j} \sum_{i=1}^{n} \gamma_{Y_i}(j)(X_i - \hat{\mu_j})^2.$$

Коэффициенты γ_{Y_i} выступают в роли весов: чем правдоподобнее полученные наблюдения в рамках модели θ_0, тем больше вес и, следовательно, соответствующая компонента смеси будет играть большую роль при оптимизации.



Это некоторое эффективное количество наблюдения Y_i в каждой компоненте (типа на 30% это похоже на пиццерию 1). По своей конструкции они дают оценку вероятности того, что элемент выборки принадлежит конкретному компоненту смеси и оценённую смесь можно использовать для классификации новых точек, то есть назначения им своего кластера.

Всё это можно пробовать запускать: далее, согласно ЕМ-алгоритму, нужно задать $\theta_0 = \hat{\theta}$ и повторять Е- и М-шаг, пока мы не поймём, что сошлись.

3.4 Чем полезна модель смеси?

Остановимся и подумаем: что можно узнать *после* того, как мы оценили параметры смеси?

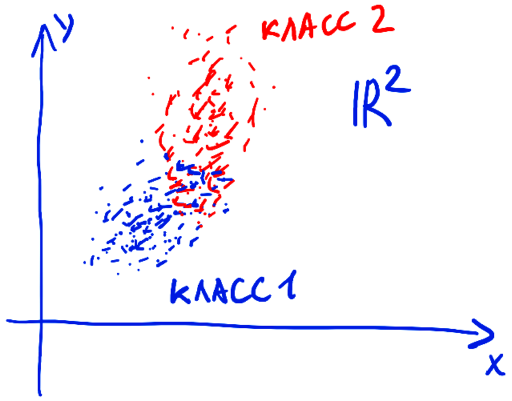
При детальном взгляде оказывается, что вес $\gamma_{Y_i}(k)$ как раз отражает вероятность того, что пицца i принадлежит компоненте смеси(пиццерии) номер k. При этом такой вес мы можем вычислить и для новых наблюдений, так как веса зависят от известных посчитанных параметров θ_0. То есть, модели смесей можно естественно использовать для кластеризации.

$X \to \gamma_0, \gamma_1$

$\gamma_0 > 0.5 \Rightarrow X$ из класса 0

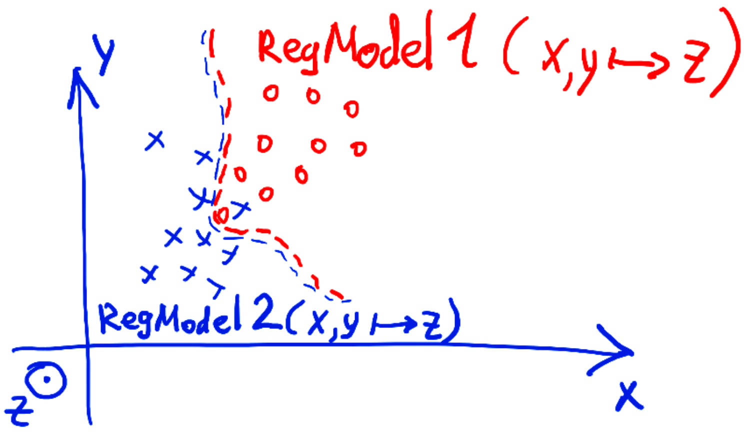
иначе: из класса 1

Мы можем свободно изменить модель одной компоненты, поставив другое более подходящее распределение с подходящей моделью параметров; например, рассмотренная нами модель носит название *гауссовской смеси* (Gaussian Mixture, например, реализован в `sklearn`). Вообще мы можем делать кластеризацию точек в $\mathbb{R}^d$, выводя оценки в многомерном случае.



Как в кластеризации, мы можем попробовать решать задачу классификации: относить пиццы к одной или к другой пиццерии в зависимости от весов γ. Но тут есть некоторая проблема: номер компоненты не обязательно совпадает с номером пиццерии. Если в инициализации параметров мы поменяем местами параметры, в результате компоненты 1 и 2 тоже поменяются местами. Если бы мы имели дополнительную информацию, например, кто-то из организаторов сказал бы: "Кажется, мы больше заказывали из ГогоПицца," – то по результату EM-алгоритма мы бы сразу соотнесли ГогоПиццу с компонентой смеси с большим количеством наблюдений, а БотманПиццу – с компонентной с меньшим числом наблюдений.

Можем ли мы пробовать решать задачу регрессии с помощью смеси? Тоже можно, если задать более конкретную параметризацию матожидания в компоненте гауссовской смеси.



Если поискать по миру, то смесь распределений, несмотря на свою простоту, имеет большое количество самых разных возможных идей приложений смеси: от биометрии до алгоритмов сегметации изображений. Важно в конкретной задаче сообразить, что можно рассмотреть как скрытую переменную и какие выбрать распределения для компонент.

Потенциал EM-алгоритма не ограничивается одной лишь смесью распределений, это на самом деле очень мощная и фундаментальная техника оценивания параметров, с которой мы только начали знакомиться.

Литература

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[10] Herbert E Rauch, F Tung, and Charlotte T Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[11] O. A. Stepanov. Kalman filtering: Past and present. an outlook from russia. (on the occasion of the 80th birthday of rudolf emil kalman). *Gyroscopy and Navigation*, 2(2):99–110, Apr 2011.

[12] R. L. Stratonovich. Conditional markov processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.

[13] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[14] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных.* Изд. дом Высшей школы экономики, 2012.

[15] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования.* М.: МЦНМО, 2013.

[16] Р.Л. Стратонович. *Условные марковские процессы и их применение к теории оптимального управления.* Московский государственный университет, 1966.

[17] А.Н. Ширяев. *Основы стохастической финансовой математики.* МЦНМО, 2016.