به نام خدا

Amirkabir University of Technology
(Tehran Polytechnic)

پروژه

تخمین توزیع تطبیقی با میانگین گیری بیزی

عطیه غفارلوی مقدم(۴۰۱۱۳۱۰۲۹)

atiyeh.moghadam@aut.ac.ir

شناسایی الگوی آماری – دکتر رحمتی

پاییز ۱۴۰۱

فهرست

۱. مقدمه

مساله ی تخمین توزیع غیرپارامتریک روش های حل متعددی دارد. دو روش رایج برای حل این مساله روش تخمین تابع توزیع با استفاده از هسته[1] و روش ترکیبی گاوسی[2] است. در روش تخمین توزیع با هسته، از پنجره ها یا توابع هسته با پنهای باند مشخص برای تخمین توزیع در تک تک نقاط استفاده میشود. در این روش پهنای باند یک پارامتر ثابت برای تک تک داده ها میباشد.
از آنجاییکه تحقیقات نشان میدهند که تخمین صحیح توزیع بیشتر از وابستگی به تابع هسته به انتخاب صحیح پهنای باند وابسته است؛ رویکرد های دیگری برای تخمین تابع توزیع وجود دارند که پهنای باند هر نقطه را متمایز در نظر میگیرند. به این دسته از روش ها روش های تخمین تطبیقی توزیع با هسته[3] یا به اختصار AKDE میگویند. این روش ها ممکن است بسته به انتخاب پارامتر های مناسب نسبت به رویکرد غیر تطبیقی عملکرد بهتری ارائه دهند اما همچنان چالش هایی چون انتخاب پارامتر های پهنای باند مناسب، انتگرال مساوی با یک برای توزیع و حفظ محلیت[4] در توزیع مواجه هستند.
رویکرد تخمین توزیع تطبیقی با میانگین گیری بیزی یا به اختصار ADEBA یک رویکرد خود¬تنظیم گر[5] است که نیاز به مقداردهی پارامتر ها از طرف کاربر ندارد. در این روش پارامتر ها از طریق جستجو و تجمیع روی یک شبکه[6] از پارامتر ها و میانگین گیری بیزی روی آنها صورت میگیرد. این مقاله نشان میدهد که این روش مقاوم و بدون نیاز به پارامتر است و همچنین نتایج آن قابل مقایسه با روش های رایج[7] تخمین توزیع میباشد.

---

[1] Kernel density estimation
[2] Gaussian mixture models
[3] adaptive kernel density estimation(AKDE)
[4] locality
[5] self¬tuning
[6] grid
[7] state of art

۲.روش پیشنهادی

۱.۲.کلیت روش پیشنهادی

روش استاندارد تک متغیره ی تخمین توزیع با هسته به این صورت است که اگر $D = \{x_1, x_2, \dots, x_n\}$ یک مجموعه نمونه از n داده ی مستقل و به طور یکسان توزیع شده[8] باشد؛ تابع توزیع تخمین زده شده به صورت زیر است:

$$\hat{p}(x|D,h) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - x_i}{h})$$

که h ثابت پهنای باند و k تابع کرنل میباشد.

همانطور که پیش تر هم گفته شد در رویکرد ADEBA پهنای باند یک ثابت نیست بلکه بسته به نقاط تعیین میشود. در این رویکرد تابع تخمین به صورت زیر می‌باشد:

$$\hat{p}(x|D,h_i\{\alpha,\beta\}_{i=1}^{n}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i(\alpha,\beta)} k(\frac{x - x_i}{h_i(\alpha,\beta)})$$

و پهنای باند نیز برای هر نقطه به صورت متمایز و بر اساس رابطه ی زیر محاسبه می شود:

$$h(x_i,\alpha,\beta) = \frac{\alpha}{(p_0(x_i))^{\beta}}$$

پارامتر آلفا یک پارامتر برای کنترل مقیاس کلی هسته و بتا پارامتری است که میزان تاثیر توزیع پایلوت را در تعیین پهنای باند مشخص میکند. توزیع پایلوت یک توزیع اولیه است که در تعیین پهنای باند به کار میرود. در مورد تعیین این توزیع در ادامه بیشتر صحبت خواهد شد. مقدار اولیه ی این توزیع خوب است که نزدیک و شبیه به توزیع اصلی داده ها باشد ولی حتی اگر توزیع نامناسبی انتخاب شود هم رویکرد ADEBA میتواند این توزیع اولیه را بهبود دهد که در ادامه توضیح داده خواهد شد.

برای تعیین تخمین کلی از یک میانگین گیری بیزی روی همه ی پارامتر های ممکن که با مقدار احتمال پسین این پارامتر ها بر حسب داده ها وزن دهی شده است استفاده میکنیم. در این رویکرد علاوه بر استفاده از داده ها در تعیین توزیع، در محاسبه ی درست نمایی[9] پارامتر ها نیز از خود مجموعه داده کمک میگیریم.

۲.۲.استنتاج بیزی تقریبی

تخمین توزیع نهایی در این رویکرد از عبارت زیر محاسبه میشود:

$$\hat{p}(x|D) = \iint \hat{p}(x|D,h_i\{\alpha,\beta\}_{i=1}^{n}) p(\alpha,\beta|D) d\beta d\alpha$$

برای حذف انتگرال و ساده تر شدن محاسبات این انتگرال به صورت جمع تخمین زده میشود و بنابراین داریم:

$$\hat{p}(x|D) \approx \Delta\Delta\beta\alpha \sum_{\alpha}\sum_{\beta} \hat{p}(x|D,h_i\{\alpha,\beta\}_{i=1}^{n}) p(\alpha,\beta|D)$$

سپس برای محاسبه ی $p(\alpha,\beta|D)$ از روش leave¯one¯out cross validation استفاده شده است.(در آزمایشات انجام شده در مقاله ی اصلی این نکته ذکر شده است که توزیع اولیه ی آلفا و بتا تاثیر چندانی در خروجی ندارد مگر در حالتی که تابع اولیه در بعضی از نقاط صفر باشد. و همچنین با زیاد شده تعداد نمونه ها نیز اثر احتمال اولیه ی آلفا و بتا نسبت به درست نمایی کمتر و کمتر میشود. در کد پیاده سازی شده این مقدار اولیه برابر با تابع یونیفرم یک قرار گرفته است.) برای محاسبه ی $p(\alpha,\beta|D)$ داریم:

---

[8] independent and identically distributed(iid)

[9] likelihood

$$p(\alpha,\beta|D)=\prod_{i=1}^{n}\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{h_j(\alpha,\beta)}k\left(\frac{x_j-x_i}{h_j(\alpha,\beta)}\right)\frac{p(\alpha,\beta)}{p(D)}$$

که همانطور که درفرمول هم مشاهده میکنید در محاسبه ی تخمین درست نمایی در نقطه ی x_i خود این نقطه در نظر گرفته نشده است و همان LOOCV میباشد.(علت این امر این است که اگر از تخمین بیشینه درست نمایی برای تعیین پهنای باند استفاده کنیم بیشینه ی درست نمایی مقدار بهینه ی پهنای باند صفر را بر میگرداند.)

مقادیر پارامتر های آلفا و بتا هر دو باید مثبت باشند؛ منفی بود آلفا منجر به منفی شدن پهنای باند میشود و منفی بودن بتا منجر میشود که در ناحیه های چگال تر پهنای باند بزرگتر شود و در نواحی پراکنده پهنای باند کوچکتر شود که واضحا هر دوی این امور مطلوب ما نیستند.

۳.۲.تعریف توزیع پایلوت و تاثیر آن

همانطور که پیش تر هم گفته شد توزیع پایلوت خوب است شکلی شبیه توزیع اصلی داشته باشد ولی با توجه به اینکه ما شکل توزیع اصلی را نداریم، اگر توزیع نامناسبی هم انتخاب شود خود الگوریتم ADEBA میتواند آن را بهبود دهد. این بهبود به این صورت انجام میگیرد که ابتدا یک توزیع اولیه در نظر میگیریم و این الگوریتم را روی آن پیاده میکنیم و سپس مقدار تخمین جدید را به عنوان توزیع پایلوت مرحله ی بعد در نظر میگیریم. با این رویکرد در هر مرحله تابع پایلوت شباهت بیشتری به تابع توزیع اصلی دارد.

۴.۲.الگوریتم

الگوریتم اصلی این روش را در سه مرحله انجام میدهد:

پیش پردازش: در این مرحله برای کاهش اثر مقیاس و چرخش توزیع ها در ابتدار کار روی داده ها یک تبدیل PCA میزنیم که که داده ها دارای ماتریس کوواریانس همانی باشند. (تابع initialize)

پیدا کردن شبکه با پارامتر های مناسب: در این تابع به ازای هر عنصری از بتا یک مجموعه مقدار مناسب آلفا و وزن ها که در واقع همان احتمالات پسین هستند محاسبه میشود. برای پیدا کردن مقادیر آلفا یک فضای اولیه با فواصل لگاریتمی یکسان از 10^{-3} تا 10^{6} در نظر میگیرد و سپس درست نمایی هر کدام از این آلفا ها را محاسبه میکند و از میان درست نمایی ها بازه ای را که مقادیر مناسب درست نمایی را دارند به عنوان بازه ی جدید انتخاب میکند این کار را دو بار تکرار میکند و سپس روی بازه ی نهایی به دست آمده ۴۰ مقدار آلفا با فاصله های لگاریتمی یکسان تعریف میکند که برای شبکه های اصلی مورد استفاده قرار میگیرد.(تابع findIntegration و findalpha)

محاسبه ی توزیع نهایی: توزیع نهایی از میانگین بیزی پارامتر های بدست آمده در بخش قبل به دست می آید.(تابع calculatePDF)

۳.ابهامات و چالش های پیاده سازی

۱.۳.ابهامات مقاله

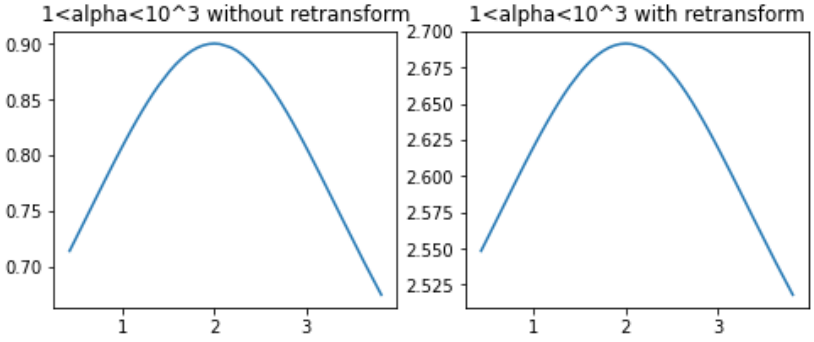
در متن مقاله و همچنین در قسمت شبه کد الگوریتم تعدادی ابهام وجود دارد که در این بخش به بعضی از آنها و فرضیه ای که در حل این ابهامات در نظر گرفته شده است اشاره میشود.

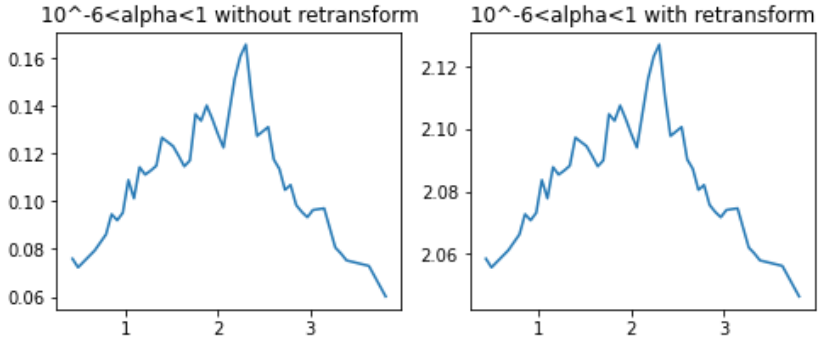
ابهام اول در مورد محاسبه ی شبه کد در الگوریتم دو میباشد. از متن مقاله میدانیم که وزن های پارامتر ها همان احتمال پسین $p(\alpha,\beta|D)$ میباشد اما در این شبه کد حاصلضرب درست نمایی در احتمال پسین قرار داده شده است و با توجه به اینکه پیش تر هم گفته شد مقدار احتمال اولیه یونیفرم در نظر گرفته شده است این عبارت معادل نرمال شده ی درست نمایی ضرب در درست نمایی است که مقدار آن بزرگ میشود. شکل زیر این قسمت از شبه کد را نشان میدهد:

1: **function** FINDINTEGRATIONGRID($D, B, \hat{p}_0$)
2: $A \leftarrow [\,]$
3: $W \leftarrow [\,]$
4: $w_\Sigma \leftarrow 0$
5: **for** $\beta \in B$ **do**
6: $A_\beta \leftarrow$ FINDALPHA(D, β)
7: $\Delta_\alpha \leftarrow$ Distance between values in A_β
8: $W_\beta \leftarrow \{\Delta_\alpha p(D|\alpha,\beta)p(\alpha,\beta|D)\}_{\alpha\in A}$ ▷ Uses pilot $\hat{p}_0$ for obtaining h
9: $w_\Sigma \leftarrow w_\Sigma + \sum_{w_i \in W_\beta} w_i$
10: $A \leftarrow [A, A_\beta]$ ▷ Add column to matrix
11: $W \leftarrow [W, W_\beta]$
12: $W \leftarrow W/w_\Sigma$
13: **return** (A, W)

برای حل این ابهام و با توجه به اینکه در متن مقاله تصریح شده بود که وزن ها همان احتمالا پسین هستند در این بخش این وزن ها به صورت احتمال پسین در نظر گرفته شدند.

ابهام دوم در خصوص روش تعیین آلفاست. در کل مقاله توضیحات بسیار کمی در رابطه با پارامتر آلفا، نحوه ی تعیین آن و علت انتخاب ثابت پارامتر های آلفا مین و آلفا مکس و حتی پارامتر ثابت آستانه وجود دارد. به نظر می آید که این نتایج و پارامتر ها ی ثابت کاملا به صورت تجربی به دست آمده اند. تعدادی آزمون و خطا در خصوص پارامتر آلفا و دستکاری سه پارامتر مرتبط با آن یعنی $\alpha_{min}, \alpha_{max}, threshold$ صورت گرفته است که نتایج آن در قسمت کد موجود است. این نتایج نشان میدهد که از آنجاییکه آلفا پارامتریست که رابطه ی مستقیمی با پهنای باند دارد، کوچکتر کردن بازه ی آن باعث کوچکتر شدن پهنای باند و در نتیجه شکسته شدن شکل توزیع نهایی میشود و همچنین زیاد کردن آن باعث هموار شدن توزیع نهایی میشود.





ابهام سوم در خصوص عدم در نظر گیری $\Delta\beta$ در محاسبات وزن است. هیچ کجا توضیح داده نشده است که چرا از در نظر گرفتن این مقدار صرف نظر شده است در حالیکه $\Delta\alpha$ با مقادیر کوچکتر در محاسبات آورده شده است.(اگر یک ثابت نتیجه ی چندانی در خروجی ندارد پس نیازی به ضرب کردن ثابت $\Delta\alpha$ هم نبود.)

ابهام چهارم ممکن است چندان به عنوان نقد به مقاله وارد نباشد و شاید مبانی ریاضی محکمی داشته باشد(هر چند که سرچ کردم چندان چیزی دستگیرم نشد)در مقاله ذکر میکند که برای کاهش اثر مقیاس و چرخش از تبدیلی استفاده میکنیم که کوواریانس داده را همانی کند. و بعد از پایان الگوریتم تبدیل معکوس را انجام میدهیم تا به فضای اولیه برگردیم. با توجه به اینکه تبدیل روی خود نمونه داده ها انجام میشود اما تبدیل معکوس روی توزیع تخمین زده شده، برای من ابهام داشت که آیا تخمین بدست آمده بعد از تبدیل معکوس نیز شکل مناسبی دارد یا خیر. نتایج نشان میدهد که شکل توزیع پس از تبدیل معکوس تغییری نمیکند و فقط مقیاس اعداد آن تغییر میکند.

۲.۳.چالش های پیاده سازی

چالش اصلی پیاده سازی زمان اجرای الگوریتم است. برای بررسی تاثیر پارامتر ها باید تعداد نمونه ها تا ۱۶۰۰ افزایش پیدا میکرد همچنین پارامتر T نیز تا مقادیر ۱۰۰ میتوانست افزایش پیدا کند اما زمان انجام برای ۱۰۰۰ نمونه و ۲۵ گام بیشتر از ۸ ساعت طول کشید. به همین دلیل تاثیر این پارامتر ها فقط در مقیاس کوچک مقایسه شد.

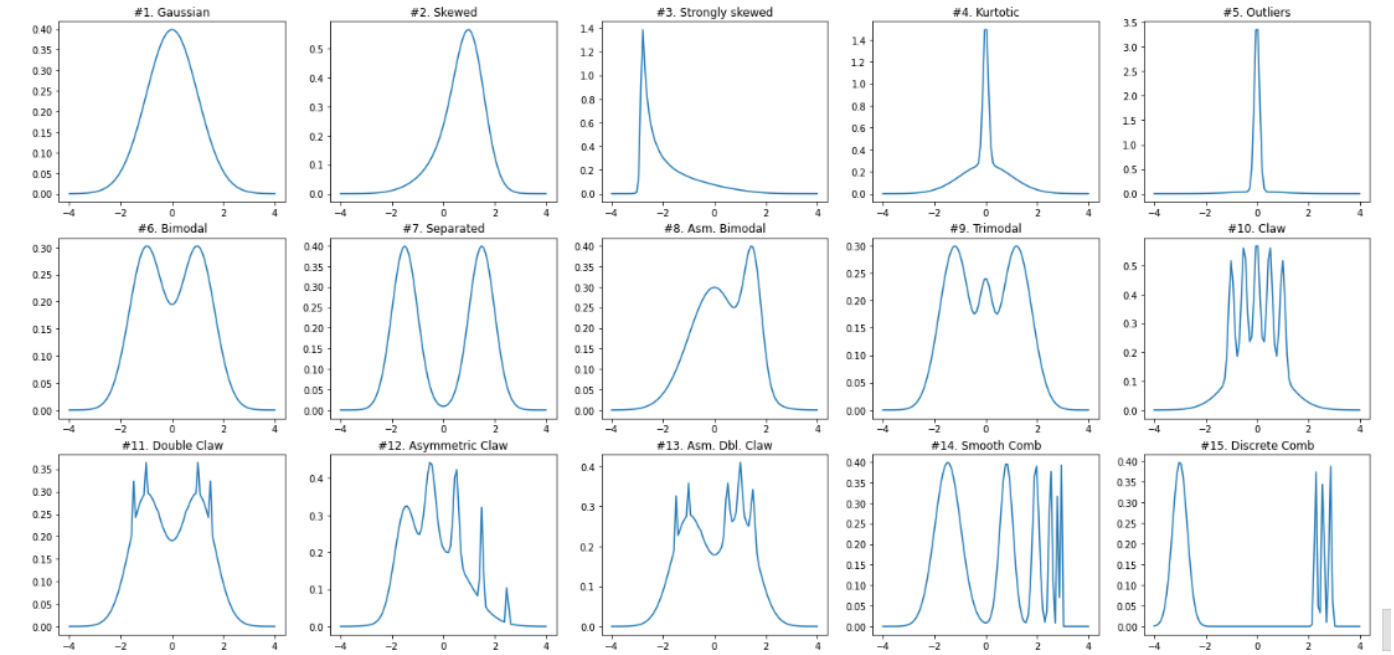
چالش دوم زیاد شدن مقدار درست نمایی در تعداد نمونه های بالا بود که باعث overflow شدن خروجی میشد. برای جلوگیری از این موضوع مقدار درست نمایی با لگاریتم درست نمایی جا به جا شد اما خروجی حاصل با محاسبه ی لگاریتم بزرگنمایی صحت خوبی نداشت به همین علت الگوریتم فعلی با همان کد اصلی درست نمایی کار میکند.(حتی این هم امتحان شد که در وزن ها مجدد مقدار ده به توان درست نمایی قرار گیرد اما نتیجه ی حاصله خوب نبود.)

۴.نتایج خروجی

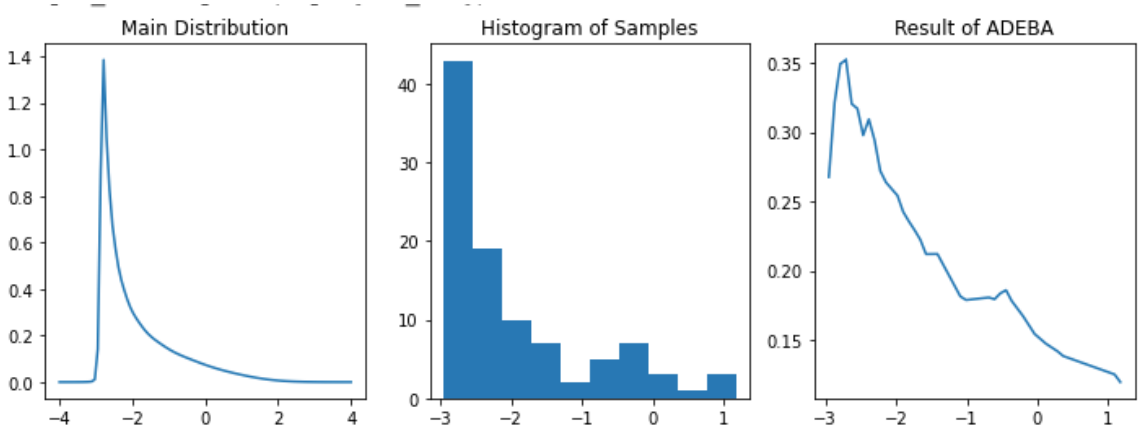
در این بخش ابتدا نتیجه ی پیاده سازی روی ۱۵ توزیع marron–wand آورده شده است و سپس به بررسی تاثیر پارامتر ها مختلف روی خروجی پرداخته شده است.

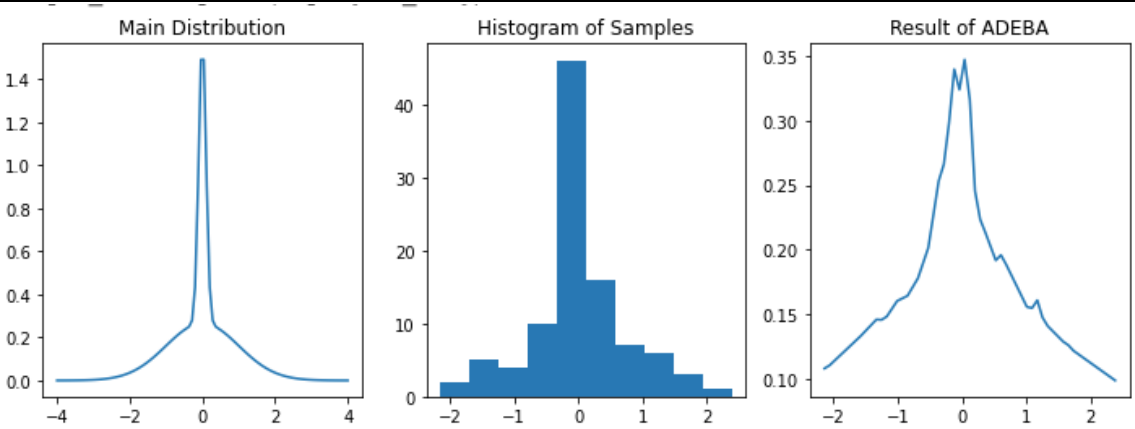
۴.۱.تست روی توزیع های مختلف

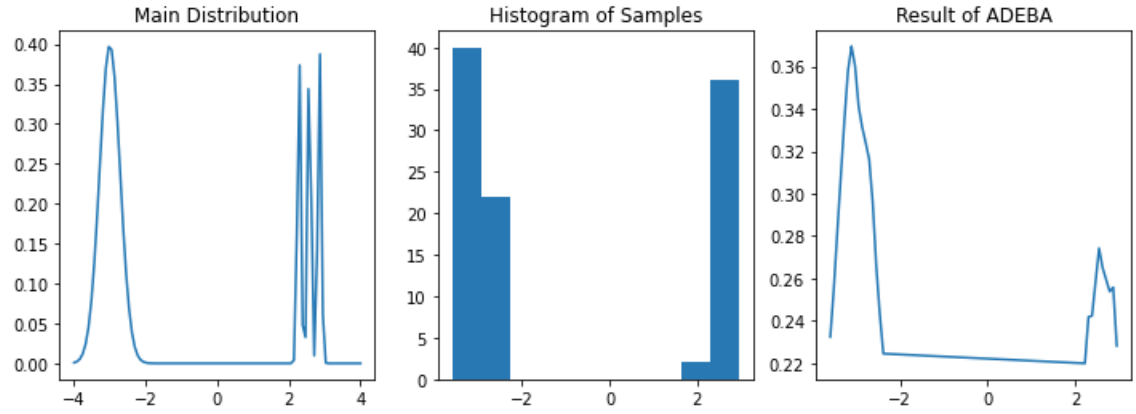
در مقاله ی اصلی ۱۵ توزیع مارون–وند به صورت زیر است که این توزیع ها در قسمت creating benchmarks تعریف شده اند.



سپس از هر کدام از توزیع ها ۱۰۰ نمونه گرفته شده و به عنوان ورودی به الگوریتم پیاده سازی داده شده و خروجی آن رسم شده است که در ادامه بعضی از این خروجی ها را مشاهده میکنید:

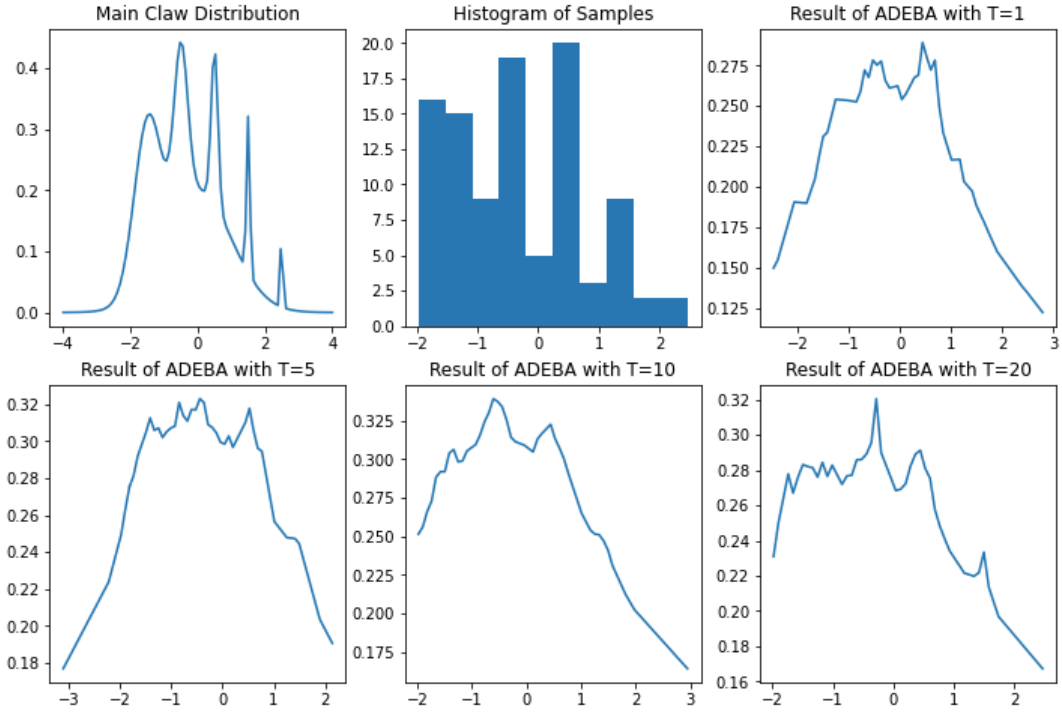






۴.۲.بررسی تاثیر پارامتر T

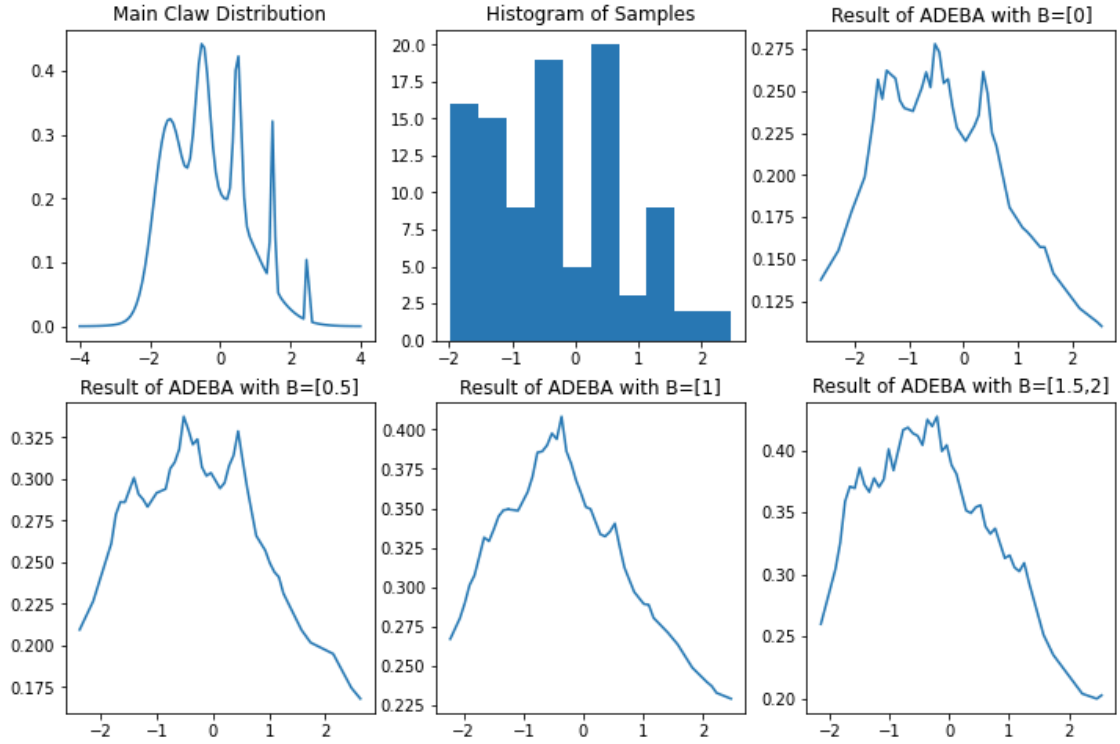
در این قسمت با ثابت نگه داشتن پارامتر های دیگر الگوریتم را روی مقادیر مختلف T تست کردیم.(به دلیل محدودیت محاسباتی امکان تست مقادیر بزرگتر وجود نداشت) و نتایج به صورت زیر است:



این مقایسه نشان میدهد که هر چقدر تعداد گام ها افزایش پیدا کند توزیع تخمین زده شده به توزیع اصلی شباهت بیشتری پیدا میکند.

۴.۳.بررسی تاثیر پارامتر β

برای بررسی تاثیر پارامتر بتا باید بررسی میشد که به ازای چه مجموعه هایی از بتا شکل توزیع به چه صورت است این بررسی روی سه مقدار مهمی که در متن مقاله هم به آنها اشاره شد و دو مقدار بزرگتر انجام شد و نتایج به صورت زیر بود:



بررسی پارامتر بتا با پارامتر های بالا به تنهایی نمیتواند نتیجه ی مشخصی در خصوص بتا بدهد چرا که بتا خودش پارامتری است که میزان تاثیر توزیع پایلوت را در خروجی مشخص میکند و خود توزیع پایلوت در هر گام بهبود بیشتری پیدا میکند. اما آنچه که از این آزمایش قابل برداشت است این است که با زیاد کردن مقدار بتا و در واقع تاثیر توزیع پایلوت تخمین حاصل شکسته شکسته تر میشود و به نظر میرسید که پهنای باند را هم کوچکتر میکند. اما این تاثیر باید در تعداد گام بالا که به یک توزیع پایلوت مناسب رسیده ایم هم بررسی شود.

۵.نتیجه گیری و کارهای آینده

در هنگام پیاده سازی این مقاله چندین ایده به نظرم رسید که به نظر میرسد میتواند در بهبود نتیجه ی حاصل موثر باشد. اول استفاده از شبه درست نمایی[10] به جای درست نمایی با توجه به اینکه درست نمایی همیشه سعی در صفر کردن پهنای باند دارد اگر به جای محاسبه ی درست نمایی در پیدا کردن آلفاهای مناسب از شبه درست نمایی استفاده کنیم نتیجه ممکن است بهبود داشته باشد و بازه ی شبکه ی بدست آمده برای آلفا بازه دقیق تر و متمرکز تری باشد. در این پروژه تلاش شد تا برای نسخه ی چند متغیره از کرنل گاوسی چند متغیره استفاده شود. میتوان به جای استفاده از کرنل گاوسی چند متغیره از ضرب کرنل ها استفاده شود. همچنین در این پروژه از بررسی انواع مختلف توابع احتمال اولیه برای آلفا و بتا چشم پوشی شد و آلفا و بتا با توزیع پیشین یونیفرم در نظر گرفته شدند. میتوان بررسی کرد که با سایر احتمال های پیشین رایج بهبود توزیع چقدر خواهد بود.و در نهایت برای تبدیل اولیه میتوان به جای PCA از ZCA و یا روش های دیگری برای همانی کردن کوواریانس داده ها استفاده کرد.

[10] pseudo–likelihood